Social Network Analysis and Mining (2024) 14:18
https://doi.org/10.1007/s13278-023-01183-9

CASE REPORT

Offensive language and hate speech detection using deep learning in football news live streaming chat on YouTube in Thailand

Peerat Pookpanich[1] · **Thitirat Siriborvornratanakul**[1]

Received: 21 September 2023 / Accepted: 6 December 2023

Abstract
Today, hate speech is frequently seen on Thai social media platforms such as Facebook, Twitter, and even online video platforms such as YouTube. In live video broadcasts of football news, for example, some Thais expressed hate speech toward opposing football fans and players. This paper presented offensive language and hate speech detection for Thai in YouTube live streaming chat with transformer-based language models by using five BERT models, including BERT, XLM-RoBERTa, DistilBERT, WangchanBERTa, and TwHIN-BERT, which were trained with multilingual languages as well as Thai. In the data labeling process, a two-step data labeling procedure was developed. The first stage involved automated data labeling utilizing the WangchanBERTa model, and the second stage involved manual data labeling conducted by the researchers. We developed text classification models using 11 different positive and negative class ratio datasets to get the most efficient model. In terms of recall and F1 score, the results showed that XLM-RoBERTa performed the best. It yielded an average recall and F1 score of 0.9669 and 0.9530, respectively. However, neither of the five models has significantly different performance. When considering the purpose of the application, DistilBERT is most appropriate. Because of its similar performance to XLM-RoBERTa, it has smaller model sizes and works faster.

Keywords Offensive language detection · Thai natural language processing · Text classification · Deep learning

1 Introduction

YouTube is widely utilized as a prominent social media platform within the context of Thailand. In 2022, the number of users of this platform reached 42.8 million, which accounted for 61.1% of Thailand's population (Digital 2022). This platform serves as one among multiple channels for content delivery, including mainstream content broadcasted through free TV or digital TV, mainstream media broadcasts on YouTube, and sub-YouTuber content. Football, often known as soccer, is widely regarded as one of the most popular sports worldwide. In nowadays, football has emerged as a highly esteemed and popular sport on worldwide. The football event held in England's top-tier league, commonly referred to as the English Premier League, attracts a global viewership of roughly 3.2 billion individuals. A considerable proportion of football enthusiasts in Thailand are included within the aforementioned group. The growing popularity of football and the substantial user base on YouTube has resulted in a diverse range of content being offered by mainstream and niche sports channels. This content includes news reports, reviews, and analyses of competitive matches, presented in video format and often streamed live. Viewers have the opportunity to engage in real-time commentary while watching these streams, which can give way to comments expressing positive sentiments, negative feedback, and instances of hate speech. Several studies on offensive language (Gao et al. 2020; Hamdy 2021; Wei et al. 2021), abusive language (Wanasukapunt et al. 2021; Kaur et al. 2021; Gashroo and Mehrotra 2022), and hate speech (Kovács et al. 2021; Yadav et al. 2023) have addressed the issue of utilizing offensive language, abusive language, and hate speech on social media platforms, emphasizing the significance of natural language processing (NLP) in resolving this concern.

Currently, YouTube has a solution to the hate speech comment system, which detects users with inappropriate comments, sends a warning message, and promptly bans

✉ Thitirat Siriborvornratanakul
thitirat@as.nida.ac.th

Peerat Pookpanich
peerat.poo@stu.nida.ac.th

[1] Graduate School of Applied Statistics, National Institute of Development Administration, Bangkok 10240, Thailand

Published online: 03 January 2024

a user's comments for 24 h if they continue to post inappropriate comments. However, the system is still incapable of detecting Thai comments effectively, requiring channel administrators to manually ban hate speech, cyberbullying, and offensive comments. Occasionally, the live performer will need to respond to or communicate with a user who has made an inappropriate comment, causing the content broadcast to be interrupted.

According to the information presented above, no study has been found on Thai YouTube live streaming, but analogous studies have been discovered. Panchala et al. (2022); Mnassri et al. (2023); Kovács et al. (2021) have performed hate speech detection in English. In Wanasukapunt et al. (2021); Pasupa et al. (2022), there's hate speech detection in Thai. The majority of this research comes from other social media sources. We can utilize this knowledge to develop a Thai hate speech detection system for live football news streaming that has not been found in any research. The unique contribution of our study is as follows: (1) the absence of prior research on Thai hate speech identification specifically focused on YouTube live streaming chat. (2) In the previous study (Wanasukapunt et al. 2021), a single BERT model was employed for comparison with both the deep learning model and the regular machine learning model, while in this research, the performance is evaluated by utilizing five BERT models. (3) This study aims to employ a two-stage labeling approach for datasets, where the initial step involves automated data labeling, followed by a subsequent step of manual data labeling. (4) We conducted experiments by training the model using various datasets in order to determine the dataset that yields the optimal performance for the model based on the proportion of fake hate speech.

2 Related work

In recent years, researchers have made significant advancements in the identification of offensive, abusive language, and hate speech through the utilization of several methodologies. These methodologies encompass machine learning and deep learning approaches and have been applied to research conducted in both English and Thai languages.

In the study performed on the English language, Gao et al. (2020) utilized three transformer-based pre-trained models, including BERT (Devlin et al. 2018), RoBERTa (Liu et al. 2019), and XLNet (Yang et al. 2019), to detect offensive language. Among the models evaluated, RoBerta had the highest level of effectiveness, as indicated by its F1-score of 0.795%. Hamdy (2021) has also conducted research on Offensive Language Detection in English utilizing the 2019 and 2020 OLD dataset, a text from Twitter, to create a classification using four techniques: BERT Embeddings, *n*-gram (range 1–3), TFIDF, and Doc2Vec. BERT embeddings with an accuracy of 83.7% and an F1-score of 0.791% have the highest precision and F1-score. Wei et al. (2021) conducted another study on offensive language and hate speech detection in English text using deep learning and transfer learning and the same Twitter data. In a comparison of the efficacy of the BI-LSTM, DistilBERT, and GPT-2 models, the BI-LSTM model was determined to be the most effective.

For the research on the Thai language, Wanasukapunt et al. (2021) conducted a study on the classification of offensive Thai language on social media using deep learning techniques. They compared the performance of traditional machine learning algorithms such as the Discriminative Multinomial Naive Bayes (DMNB), Random Forest (RF), the Maximum Entropy, the Support Vector Machine (SVM), the Bernoulli Naive Bayes (BNB) classifier, the Decision Table/Naive Bayes Hybrid (DTNB), the Repeated Incremental Pruning to Produce Error Reduction (RIPPER), the k-Nearest Neighbor (kNN) and the C4.5 Decision Tree with deep learning models including the Bidirectional Long-Short Term Model (Bi-LSTM) and DistilBERT. The evaluation was conducted using both binomial and multinomial approaches. The outcome indicates that the deep learning module produces more accurate results (accuracy 0.8965, precision 0.9128, recall 0.9006, F1-score 0.9067). Pasupa et al. (2022) has also conducted a study of hate speech detection in Thai social media utilizing the WangchanBERTa model, one of the RoBERTa architectures trained with a set of Thai data. It was fine-tuned to optimize the model, and its performance was evaluated using the F1-score and a hybrid loss function comprised of the Ordinal regression loss function and Pearson correlation coefficients. The F1-score results were 78.38–0.88%, which is significantly greater than the conventional loss function, and the relative improvement in average mean square error was 0.24–78.5%.

The results shown in references Gao et al. (2020); Hamdy (2021); Wei et al. (2021) demonstrate that the utilization of RoBerta, BERT, and BI-LSTM yields favorable outcomes in the context of English hate speech detection. In the Thai study, it was seen that the outcomes obtained from the application of DistilBERT were superior to those obtained from BI-LSTM, as reported in reference Wanasukapunt et al. (2021). The study conducted by WangchanBERTa Pasupa et al. (2022) employed a Thai-specific model and implemented text processing rules suitable for the Thai language. The outcomes obtained from this investigation were deemed satisfactory.

Based on the aforementioned study, we intend to construct a hate speech detection system tailored to the research context. To achieve this, we will employ various models including RoBerta, BERT, DistilBERT as a multilingual version, and WangchanBERTa, a Thai-trained model. The primary objectives are to evaluate the performance of each

model and to compare the outcomes between an unimodal model trained specifically in Thai and a multilingual model.

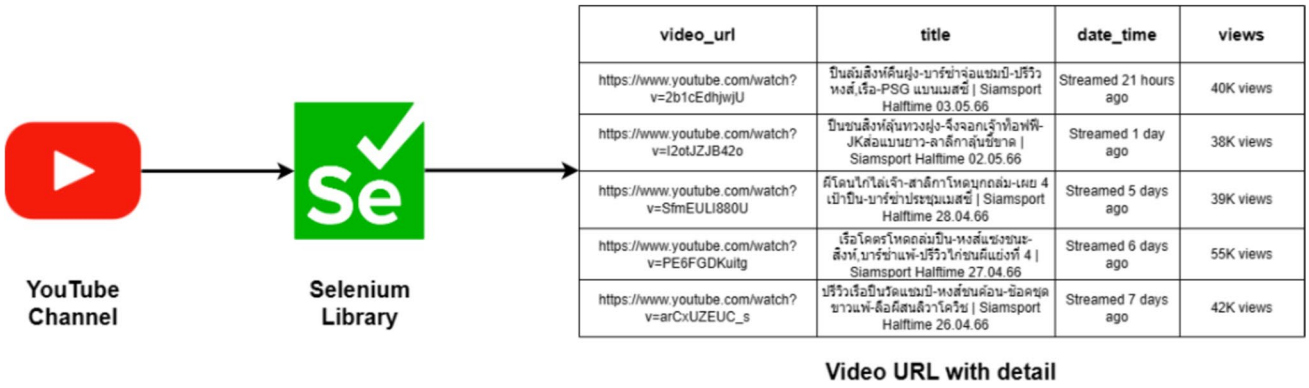
3 Dataset

3.1 Data collection & data pre-processing

The objective of this study is to conduct an analysis of the Thai messages that are posted by viewers in the live chat of Thai broadcasting videos during live streaming sessions. The primary area of interest will be around news articles, football critiques, and match analyses available on the YouTube platform in Thailand spanning the period from 2021 to 2023. The subject matter of attention relates to the five most significant football matches in Europe, including Premier League (England), Laliga (Spain), Bundesliga (Germany), Series A (Italy), and Ligue 1 (France).

The five football leagues enjoy significant popularity both within Thailand and internationally. Their weekly competitions attract substantial viewership for live news broadcasts and match review videos. Consequently, a significant amount and diverse range of analytical data is generated. The duration of a live broadcast exceeds 30 min in order to ensure a sufficient quantity of data is available for analysis throughout each live session.

The data collection process involves gathering live conversation text from live streaming videos on 11 Thai YouTube channels that meet the specified criteria for live video selection. This process (Fig. 1) begins with the extraction of URLs using the Selenium library (version 4.8.3) to obtain all video URLs within each channel. Next, in accordance with the research's purpose, choose a URL that is related to the subject matter. The demonstrated process in Fig. 2 employs the chat_downloader library (version 0.2.4) to collect and merge the messages acquired



video_url	title	date_time	views
https://www.youtube.com/watch?v=2b1cEdhjwjU	ปืนลั่นสิงห์คืนฝูง-บาร์ซ่าจ่อแชมป์-ปรีวิวหงส์,เรือ-PSG แบนเมสซี่ | Siamsport Halftime 03.05.66	Streamed 21 hours ago	40K views
https://www.youtube.com/watch?v=I2otJZJB42o	ปืนขนสิงห์ลุ้นหวงฝูง-จิ้งจอกเจ้าท้องฟ้า-JKส่อแบนยาว-ลาลีกาลุ้นชิชาด | Siamsport Halftime 02.05.66	Streamed 1 day ago	38K views
https://www.youtube.com/watch?v=SfmEULI880U	ผีโดนไก่ไล่เจ๊า-สาลิกาโหดบุกถล่ม-เผย 4 เบ้าปืน-บาร์ซ่าประชุมเมสซี่ | Siamsport Halftime 28.04.66	Streamed 5 days ago	39K views
https://www.youtube.com/watch?v=PE6FGDKuitg	เรือโดดโหดถล่มปืน-หงส์แช่งชนะ-สิงห์,บาร์ซ่าแพ้-ปรีวิวไก่ขนผีแข่งที่ 4 | Siamsport Halftime 27.04.66	Streamed 6 days ago	55K views
https://www.youtube.com/watch?v=arCxUZEUC_s	ปรีวิวเรือปืนวัดแชมป์-หงส์ขนค้อน-ข้อศอกขาวแพ้-ล็อผีสนสิวาโควิช | Siamsport Halftime 26.04.66	Streamed 7 days ago	42K views

Fig. 1 Video URL collection process

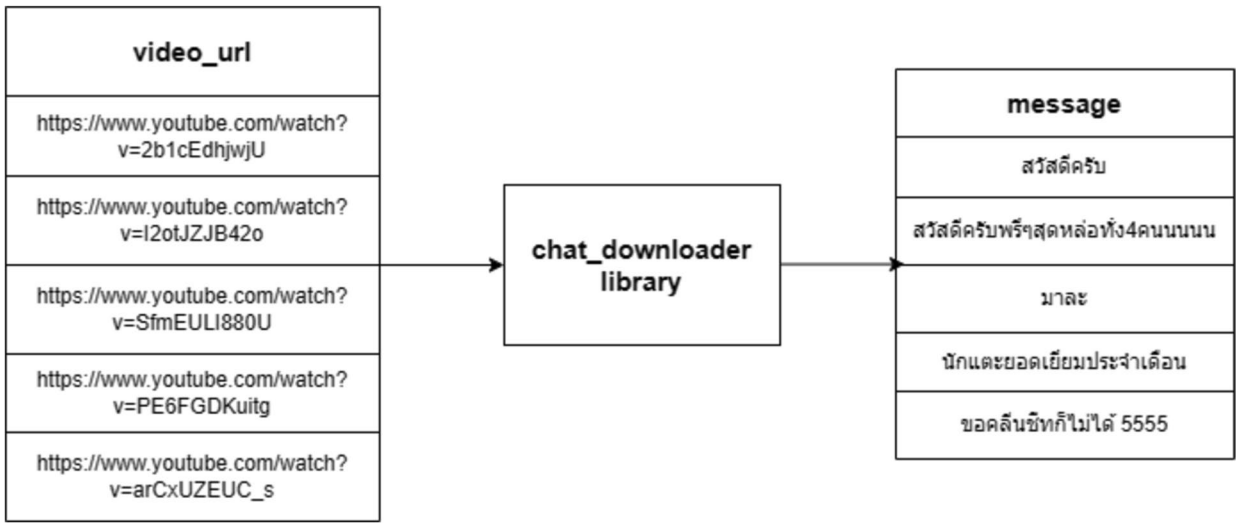


Fig. 2 Message collection process

from individual URLs. The total number of messages acquired is 2,028,434.

In the data pre-processing section, the data were subjected to cleansing procedures. These procedures involved the removal of special characters, stop words, pure numeric records, and emoji. The purpose of these procedures was to ensure that each text exclusively consists of understandable phrases and sentences.

3.2 Data labeling

The dataset has a grand total of 2,028,434 unlabeled messages. The process of assigning linguistic major students (Wanasukapunt et al. 2021) or specialists (Pasupa et al. 2022) to categorize messages can be resource-intensive in terms of time and budget (Zhang et al. 2021). While GPT-3 (Dou et al. 2021) and ChatGPT (Gilardi et al. 2023) have demonstrated superior performance compared to human crowd-workers in English text-annotation tasks, their performance in Thai text-annotation tasks is yet to be established. In this study, a two-step data labeling procedure was developed. The first stage involved automated data labeling utilizing the WangchanBERTa model, and the second stage involved manual data labeling conducted by the researchers. The specific details of this process are as follows.

3.2.1 Automated data labeling

For the purpose of automating data labeling, the model selected was wangchanberta-base-att-spm-uncased. This model has been trained using the most extensive Thai dataset available, encompassing diverse information sources and ensuring high data quality, and it has applied text processing rules that are specific to Thai (Lowphansirikul et al. 2021). We automated the data labeling process using the wangchanberta-base-att-spm-uncased, which was fine-tuned with the Wisesight sentiment dataset. This dataset comprises text data from Thai social media, enabling our chosen model to perform sentiment classification in the Thai language. Utilizing this model on our unlabeled live chat messages allows us to automatically assign sentiment labels, as depicted in Fig. 3.

In Fig. 3, the sentiments of the message are categorized as neutral, negative, positive, and question. From a total of 2,028,434 messages, we can determine the number of labeled messages in each category: neutral 1,751,285 (86.34%), negative 210,774 (10.39%), positive 65,083 (3.21%), and question 1,292 (0.06%), as shown in Fig. 4. Since the ratio of questions is extremely low, the study only includes neutral, negative, and positive messages.

3.2.2 Manual data labeling

When we have labeled all three types of messages, we have an amount of 210,774 labeled negative messages containing offensive language (https: www.lawinsider.com, dictionary, offensive-language 2023) or hate speech (https: dictionary.cambridge.org, dictionary, English, hate-speech 2023) by filtering messages using commonly used negative terms (blasphemy, cruelty, disrespect or sexual insult etc.) in Thai social media such as กาก, เหี้ย, ควย, กระจอก, ปัญญาอ่อน, ตุ๊ด (You are so mean. sucky, asshole, dick, beggarly, retarded, gay etc.) that were collected from various sources by researcher.

The demonstrated process in Fig. 5 yielded a total of 13,981 instances of hate speech messages, which accounts for 6.63% of the overall 210,774 negative messages. In the subsequent phase, the labeling procedure will be divided into two distinct processes in order to enhance the precision of the labeling outcome.

- For positive and neutral labeled messages, we use hate speech keywords to filter out messages containing these keywords in order to minimize information errors.
- For hate speech labeled messages, because the text contains hate speech keywords, there may not always be

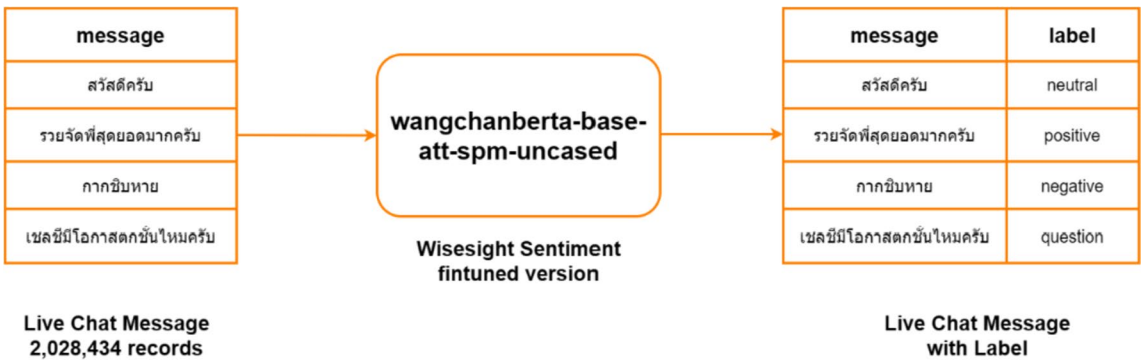


Fig. 3 Automated data labeling process

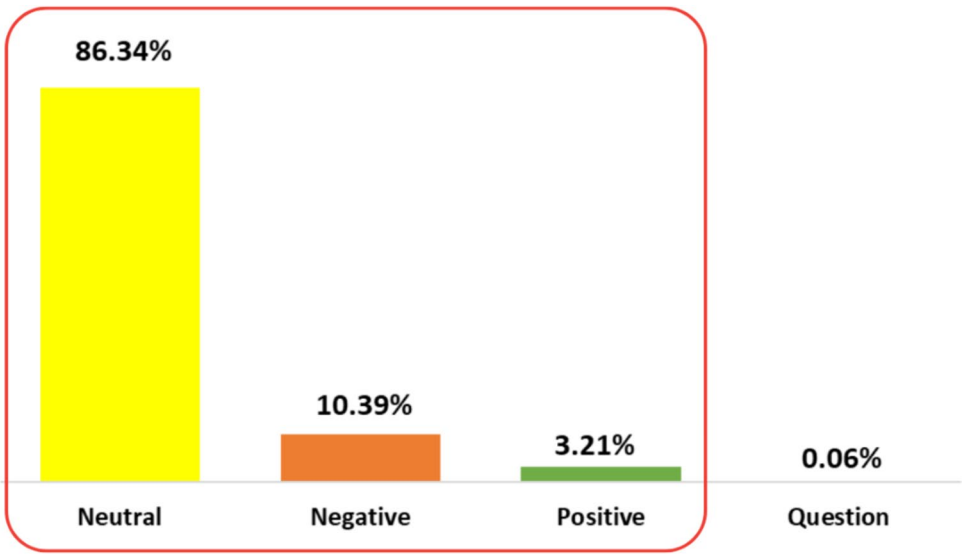


Fig. 4 Automated data labeling result

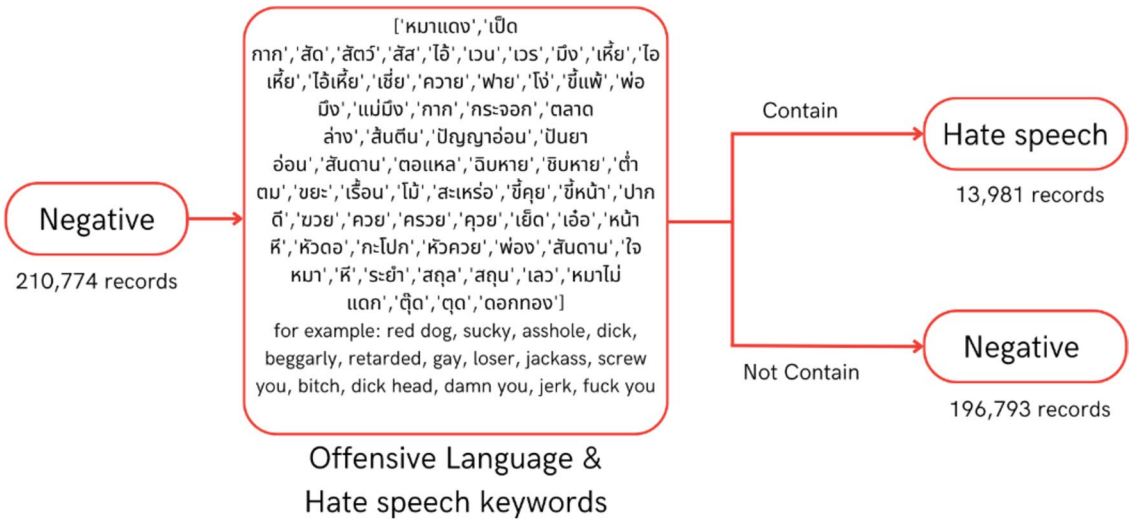


Fig. 5 Filtering offensive language & hate speech in negative message by using offensive & hate speech keywords

intentional abuse or an inappropriate comment. The researchers labeled the text by reading it to see what it really meant.

Based on the hate speech label messages received from the previous process in Fig. 6, 13,981 messages were divided into 10,085 hate speech messages (72.13%) and 3,896 non-hate speech messages (27.87%). Finally, the dataset we have includes five kinds of labels:

- **Positive** is a positive or encouraging message.
- **Neutral** is a nonspecific text.
- **Negative** is a text with a common negative meaning, but not one containing obscene or offensive language.
- **Hate speech** is the use of words with offensive, rude, contemptuous, or abusive meanings.
- **Fake hate speech** is a text that contains the hate speech keyword in the message but is not abusive or impolite.

In the context of this study, binary classification is employed, wherein the hate speech category is assigned a label of 1, while the categories of positive, neutral, negative, and fake hate speech are assigned a label of 0.

3.3 Split data

The entire dataset was first split into two distinct datasets: the Training and validation dataset, and the test dataset.

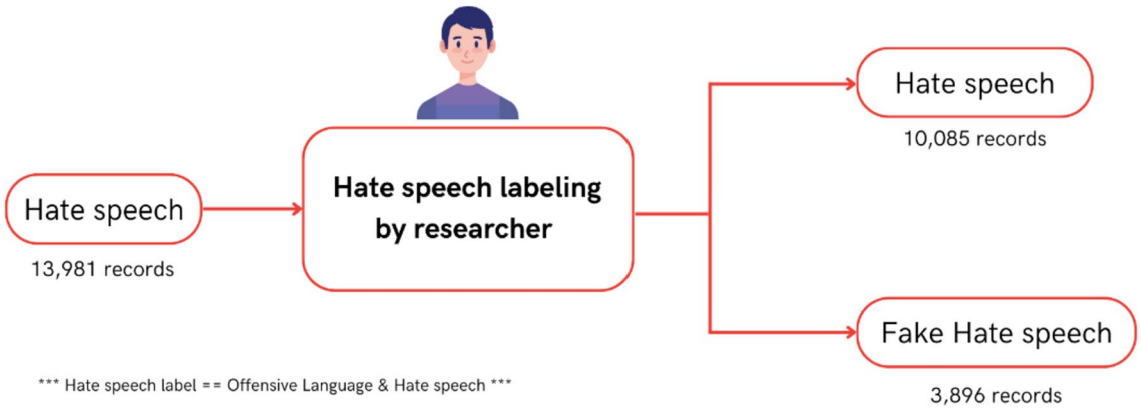


Fig. 6 Classify the true hate speech messages by researcher

Given that we have all four subclasses of non-hate speech labels, denoted as label = 0, we have conducted multiple variations on the proportions of each dataset. The specific information pertaining to each dataset is shown in Fig. 7.

3.3.1 Training & validation dataset

Initially, a dataset comprising 9000 instances of hate speech messages (label = 1) is established. Subsequently, in order to identify the datasets that yield the most optimal model performance, the proportion of non-hate speech labels (label = 0) is varied across all four categories. Table 1 presents detailed information for each dataset.

From Table 1, we divide the dataset into two groups: (1) balance class dataset (equal numbers of class 1 and 0) and (2) imbalance class dataset (non-equal numbers of class 1 and 0). In fact, the model must accommodate all message sentiments. In order to evaluate the efficacy of each model and prevent overfitting, we incorporated fake hate speech into class 0 at 10%, 20%, and 30% of the total amount of hate speech. For dataset types 1–8, we desire to fix the number of data points. Therefore, when we increase the amount of fake hate speech, we decrease the proportion of other sentiments by the same amount. For dataset types 9–11, there is no reduction in any sentiment because there is no fixed number of data points. Finally, we configured the ratio of training data to validation data to be 70:30 and made this a standard test for every model.

3.3.2 Test dataset

The number of data points was set to 4000. Each sentiment begins with a value of 1000. Class 1 has a constant value of 1000, and class 0 has a ratio adjustment based on the same method as training & validation dataset, as detailed in Table 2.

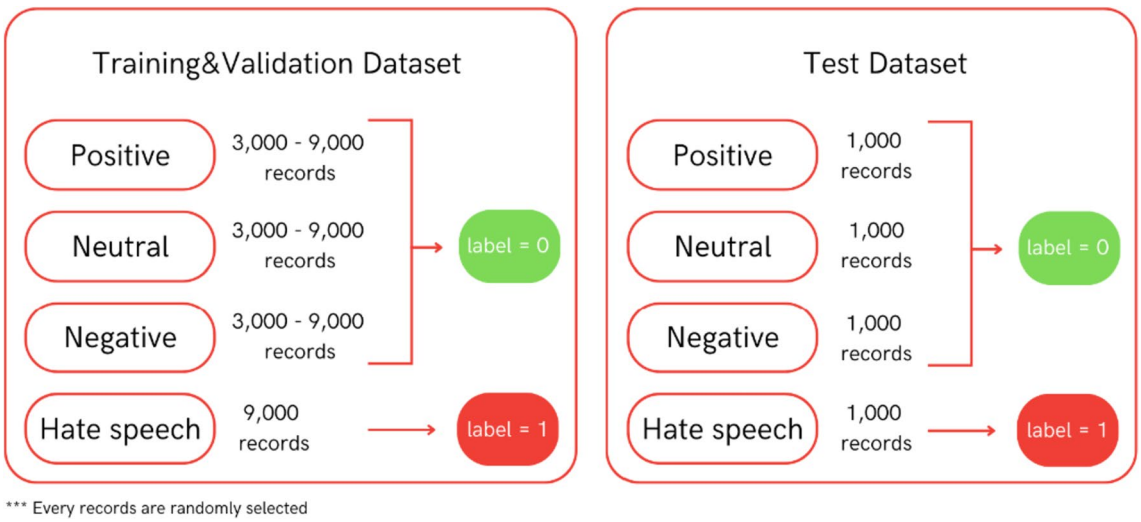


Fig. 7 Detail of initial datasets

Table 1 Training & validation dataset detail

Training & validation dataset								
Dataset group	Dataset type	%Fake hate	Positive	Neutral	Negative	Fake hate speech	Hate speech	Total
Balance class	1	0.00%	3000	3000	3000	0	9000	18,000
	2	10.00%	2700	2700	2700	900	9000	18,000
	3	20.00%	2400	2400	2400	1800	9000	18,000
	4	30.00%	2100	2100	2100	2700	9000	18,000
Imbalance class	5	0.00%	9000	9000	9000	0	9000	36,000
	6	10.00%	8700	8700	8700	900	9000	36,000
	7	20.00%	8400	8400	8400	1800	9000	36,000
	8	30.00%	8100	8100	8100	2700	9000	36,000
	9	10.00%	9000	9000	9000	900	9000	36,900
	10	20.00%	9000	9000	9000	1800	9000	37,800
	11	30.00%	9000	9000	9000	2700	9000	38,700

Table 2 Test dataset detail

Test dataset							
Dataset type	%Fake hate	Positive	Neutral	Negative	Fake hate speech	Hate speech	Total
1	10.00%	967	967	967	100	1000	4000
2	20.00%	933	933	933	200	1000	4000
3	30.00%	900	900	900	300	1000	4000
4	40.00%	867	867	867	400	1000	4000
5	50.00%	833	833	833	500	1000	4000

4 Proposed method

This research presented a text classification with transformer-based language models by using five BERT architecture models (Devlin et al. 2018), which were trained with multilingual languages, including Thai, to compare the results. The five models to be used in the experiment are all available on the website Hugging Face. Including bert-base-multilingual-cased (Devlin et al. 2018), xlm-roberta-base (Conneau et al. 2019), distilbert-base-multilingual-cased (Sanh et al. 2019), wangchanberta-base-att-spm-uncased (Lowphansirikul et al. 2021) and twhin-bert-base (Zhang et al. 2023).

The criteria for selecting the model for that experiment are based on reliability and performance. Each model has an international publication paper, and each model has a different architecture (except bert-base-multilingual-cased and twhin-bert-base), even though it is based on the same BERT.

xlm-roberta-base, bert-base-multilingual-cased, and distilbert-base-multilingual-cased are transformer-based models that are the most downloaded on Hugging Face (last accessed on August 14, 2023), which represent reliability and popularity from around the world. Wangchanberta-base-att-spm-uncased is the best-performing model in Thai. And Twhin-bert-base is a model that was just released in September 2022 and was trained with social media data on Twitter, which is similar to the data we brought into our research.

In this experiment, the researchers fine-tune a pre-train model with the simpletransformers library (version 0.64.0), which is a library built on the transformers library (versions 4.30.2) of Hugging Face, by processing with a V100 GPU 1 unit on Google Colab and setting the hyperparameters of all models to the same values: random_seed = 42; learning rate = 0.001; optimizer = AdamW; batch_size = 64; epochs = 10; max_seq_length = 128. We only adjust the classifier layer; there is no configuration of the model's architecture.

To determine the most efficient model for classifying offensive language and hate speech messages, the entire experiment will consist of two major phases.

(1) **Experiment on dataset type** is experimenting to determine the type of dataset that provides the best-performing model.
(2) **Experiment on models** is experimenting to determine the best-performing model by using the dataset that resulted in the previous section

4.1 Experiment on dataset type

DistilBERT (distilbert-base-multilingual-cased) (Sanh et al. 2019) will serve as the baseline mode for the dataset evaluation because it has the shortest training time. We will train a model using all eleven types of datasets. One model will be evaluated with all five types of test datasets. Then, the validation results and the average of the test results will be used to evaluate the model's performance.

From Table 3, we can see that the most efficient dataset type for the model is the "Type 2" dataset. Consider the recall value first, as we believe the model predicts false negatives to be more negative than false positives, and then we consider the F1-score. Although the "Type 1" dataset has the highest validation recall and F1 score, it returns to a significantly lower F1 score due to a decrease in precision value during the test section. The model trained with the Type-1 dataset has never previously learned a fake hate speech text. When evaluated with a test dataset containing all sentiment text, the model performs insufficiently.

4.2 Experiment on models

In this section, we will use the Type-2 dataset to train and compare the performance of the five previously mentioned models. One model will be trained and evaluated in three rounds by sequentially assigning the random seeds 42, 52, and 62. The results will be displayed according to Table 4.

The performance achieved by five distinct models is presented in Table 4. The results indicate that the XLM-Roberta-base model had better outcomes compared to the other models in terms of recall and F1 score. The obtained results demonstrated an average recall of 0.9669 and an F1-score of 0.9530.

5 Experimental results & discussion

The results of the research indicate that the xlm-roberta-base model has the highest efficacy in identifying hate speech, as evidenced by its recall value. Nevertheless, the outcomes of the comprehensive study examination revealed that there were no statistically significant differences among the five models.

The comparative results in Table 5 show that the appropriateness of utilizing XLM-Roberta-Base may not align optimally with the objective of developing a real-time text detection system. The training and prediction processes of the model require a substantial investment of time due to their size. The utilization of distilbert-base-multilingual-cased is advised due to its efficient training time (~40 s per epoch) and prediction time (~5 s for 4000 data) on a single V100 GPU unit.

Table 3 The result of the dataset type that provides the best-performing baseline model

Dataset type	Model type	Validation result			Test result		
		Precision	Recall	F1	Avg. precision	Avg. recall	Avg. F1
1	Distilbert	0.9904	0.9963	0.9934	0.7673	0.9990	0.8655
2	Distilbert	0.9296	0.9630	0.9460	0.8448	0.9540	0.8950
3	Distilbert	0.9007	0.9441	0.9219	0.8693	0.9310	0.8985
4	Distilbert	0.8896	0.9226	0.9058	0.8756	0.9140	0.8940
5	Distilbert	0.9918	0.9859	0.9889	0.7764	0.9920	0.8685
6	Distilbert	0.9390	0.9519	0.9454	0.8577	0.9480	0.8995
7	Distilbert	0.9177	0.9289	0.9232	0.8891	0.9260	0.9065
8	Distilbert	0.8968	0.9048	0.9008	0.9089	0.8880	0.8979
9	Distilbert	0.9346	0.9530	0.9437	0.8570	0.9440	0.8973
10	Distilbert	0.9111	0.9296	0.9203	0.8877	0.9140	0.9000
11	Distilbert	0.8899	0.9096	0.8996	0.9008	0.8970	0.8984

Table 4 Performance comparison of models

Dataset type	Model name	Validation result		
		Precision	Recall	F1
2	Distilbert-base-multilingual-cased	0.9307 ± 0.0002	0.9644 ± 0.0033	0.9473 ± 0.0017
2	XLM-Roberta-base	0.9394 ± 0.0031	0.9669 ± 0.0030	0.9530 ± 0.0004
2	bert-base-multilingual-cased	0.9323 ± 0.0029	0.9632 ± 0.0013	0.9475 ± 0.0019
2	Wangchanberta-base-att-spm-uncased	0.9410 ± 0.0044	0.9609 ± 0.0022	0.9508 ± 0.0031
2	twhin-bert-base	0.9349 ± 0.0065	0.9636 ± 0.0067	0.9490 ± 0.0013

Table 5 Comparative results of model training and testing time

Dataset type	Model name	Model run time	
		Training (sec/epoch)	Testing (sec/4000 data points)
2	Distilbert-base-multilingual-cased	39.3333 ± 0.5774	5.3333 ± 0.5774
2	XLM-Roberta-base	78.0000 ± 1.000	12.6667 ± 0.5774
2	bert-base-multilingual-cased	75.3333 ± 0.5774	12.0000 ± 1.7321
2	Wangchanberta-base-att-spm-uncased	74.3333 ± 0.5774	12.0000 ± 1.7321
2	Twitter/twhin-bert-base	80.6667 ± 0.5774	14.3333 ± 0.5774

In this section of the data labeling process, a single examiner does manual data labeling. The potential consequences include the occurrence of misclassification errors and a decrease in the overall predictive accuracy of the model. The effectiveness of automated data labeling is contingent upon the precision of the wangchanberta-base-att-spm-uncased (Lowphansirikul et al. 2021) model. Various methods can be employed to classify and assess the dataset for the purpose of comparing outcomes and acquiring the dataset of the highest quality.

Due to the purpose of the experiment, which was to develop a model for real-time detection, and the constraints on computational resources, the model may not be as efficient as possible due to the fine-tuned model, which has a maximum sequence length of 128 tokens. If a sufficient quantity of resources were available for training a model with a higher token capacity, it may potentially yield improved outcomes and facilitate the identification of longer sentences.

6 Conclusion

For offensive language and hate speech detection in Thai, we developed five Transformer-based language models, including xlm-roberta-base, bert-bas-multilingual-cased, distilbert-base-multilingual-cased, wangchanberta -base-att-spm-uncased, and twhin-bert-based on a type 2 dataset, which is a balance class dataset that contained 10% of the fake hate speech sentiment. In terms of recall and F1-score, the results showed that xlm-roberta-base performed the best. However, neither of the five models has significantly different performance. When considering the purpose of the application, distilbert-base-multilingual-cased is most appropriate. Because of its performance close to xlm-roberta-base, it has smaller model sizes and works faster.

In the future, we will have a larger dataset and more updated messages in the dataset, as the language used on social media is evolving rapidly and there is a new movement of meaning every year, necessitating a model update. In the case of data labeling, we will employ more than one person to label the dataset in order to improve its reliability. And other approaches may exist to compare the outcomes of additional experiments.

Author contributions All authors contributed equally to this manuscript.

Funding No funding.

Data availability The data that support the findings of this study are not openly available due to reasons of sensitivity and are available from the corresponding author upon reasonable request. Data are located in local computer of researcher.

Declarations

Conflict of interest The authors have no conflicts of interest to declare that are relevant to the content of this article.

References

Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, Grave E, Ott M, Zettlemoyer L, Stoyanov V (2019) Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116

Devlin J, Chang MW, Lee K, Toutanova K (2018) BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805

Digital 2022: THAILAND: https://datareportal.com/reports/digital-2022-thailand, last accessed 2023/01/15

Dou Y, Forbes M, Koncel-Kedziorski R, Smith NA, Choi Y (2021) Is GPT-3 text indistinguishable from human text? SCARECROW: A framework for scrutinizing machine text. arXiv preprint arXiv:2107.01294

Gao Z, Yada S, Wakamiya S, Aramaki E (2020) Offensive language detection on video live streaming chat. In: Proceedings of the 28th international conference on computational linguistics, pp 1936–1940

Gashroo OB, Mehrotra M (2022) Analysis and classification of abusive textual content detection in online social media. In intelligent communication technologies and virtual mobile networks. In: Proceedings of ICICV 2022, Springer, Singapore, pp 173–190

Gilardi F, Alizadeh M, Kubli M (2023) ChatGPT outperforms crowd-workers for text-annotation tasks. arXiv preprint arXiv:2303.15056

Hamdy E (2021) Neural Models for Offensive Language Detection. arXiv preprint arXiv:2106.14609

https://dictionary.cambridge.org/dictionary/english/hate-speech, last accessed 2023/08/14

https://www.lawinsider.com/dictionary/offensive-language, last accessed 2023/08/14
Kaur S, Singh S, Kaushal S (2021) Abusive content detection in online user-generated data: a survey. Procedia Comput Sci 189:274–281
Kovács G, Alonso P, Saini R (2021) Challenges of hate speech detection in social media: data scarcity, and leveraging external resources. SN Comput Sci 2:1–15
Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) RoBERTa: a robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692
Lowphansirikul L, Polpanumas C, Jantrakulchai N, Nutanong S (2021) WangchanBERTa: Pretraining transformer-based Thai language models. arXiv preprint arXiv:2101.09635
Mnassri K, Rajapaksha P, Farahbakhsh R, Crespi N (2023) Hate speech and offensive language detection using an emotion-aware shared encoder. arXiv preprint arXiv:2302.08777
Panchala GH, Sasank VVS, Adidela DRH, Yellamma P, Ashesh K, Prasad C (2022) Hate speech & offensive language detection using ML &NLP. In: 2022 4th international conference on smart systems and inventive technology (ICSSIT), pp 1262–1268, IEEE
Pasupa K, Karnbanjob W, Aksornsiri M (2022) Hate speech detection in Thai social media with ordinal-imbalanced text classification. In: 2022 19th international joint conference on computer science and software engineering (JCSSE), pp 1–6, IEEE
Sanh V, Debut L, Chaumond J, Wolf T (2019) DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108
Wanasukapunt R, Phimoltares S (2021) Classification of abusive Thai language content in social media using deep learning. In: 2021 18th international joint conference on computer science and software engineering (JCSSE), pp 1–6, IEEE
Wei B, Li J, Gupta A, Umair H, Vovor A, Durzynski N (2021) Offensive language and hate speech detection with deep learning and transfer learning. arXiv preprint arXiv:2108.03305
Yadav AK, Kumar M, Kumar A, Shivani K, Yadav D (2023) Hate speech recognition in multilingual text: hinglish documents. Int J Inf Technol 15(3):1319–1331
Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR, Le QV (2019) XLNet: generalized autoregressive pretraining for language understanding. Advances in neural information processing systems, vol 32
Zhang S, Jafari O, Nagarkar P (2021) A survey on machine learning techniques for auto labeling of video, audio, and text data. arXiv preprint arXiv:2109.03784
Zhang X, Malkov Y, Florez O, Park S, McWilliams B, Han J, El-Kishky A (2023) TwHIN-BERT: a socially-enriched pre-trained language model for multilingual tweet representations at twitter. In: Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining, pp 5597–5607