Ex-ThaiHate: A Generative Multi-task Framework for Sentiment and Emotion Aware Hate Speech Detection with Explanation in Thai

Krishanu Maity[1], Shaubhik Bhattacharya[1], Salisa Phosit[2],
Sawarod Kongsamlit[2], Sriparna Saha[1], and Kitsuchart Pasupa[2](✉)

[1] Department of Computer Science and Engineering, Indian Institute of Technology Patna, Patna 801103, India
{krishanu_2021cs19,shaubhik_2111cs19,sriparna}@iitp.ac.in

[2] School of Information Technology, King Mongkut's Institute of Technology Ladkrabang, Bangkok 10520, Thailand
{63070242,63070245,kitsuchart}@it.kmitl.ac.th

Abstract. Social media platforms have both positive and negative impacts on users in diverse societies. One of the adverse effects of social media platforms is the usage of hate and offensive language, which not only fosters prejudice but also harms the vulnerable. Additionally, a person's sentiment and emotional state heavily influence the intended content of any social media post. Despite extensive research being conducted to detect online hate speech in English, there is a lack of similar studies on low-resource languages such as Thai. The recent enactment of laws like the "right to explanations" in the General Data Protection Regulation has stimulated the development of interpretable models rather than solely focusing on performance. Motivated by this, we created the first benchmark hate speech corpus, called *Ex-ThaiHate*, in the Thai language. Each post is annotated with four labels, namely hate, sentiment, emotion, and rationales (explainability), which specify the phrases that are responsible for annotating the post as hate. In order to investigate the effect of sentiment and emotional information on detecting hate speech posts, we propose a unified generative framework called *GenX*, which redefines this multi-task problem as a text-to-text generation task to simultaneously solve four tasks: hate-speech identification, rationale detection, sentiment, and emotion detection. Our extensive experiments demonstrate that *GenX* significantly outperforms all baselines and state-of-the-art models, thereby highlighting its effectiveness in detecting hate speech and identifying the rationales in low-resource languages. The code and dataset are available at https://github.com/dsmlr/Ex-ThaiHate.
Disclaimer: The article contains offensive text and profanity. This is due to the nature of the work and does not reflect any opinion or stance of the authors.

Keywords: Hate Speech · Sentiment · Emotion · Explainability · Thai · Multi-task

G. De Francisci Morales et al. (Eds.): ECML PKDD 2023, LNAI 14174, pp. 139–156, 2023.
https://doi.org/10.1007/978-3-031-43427-3_9

1 Introduction

Social media platforms have become an integral part of people's lives, providing opportunities to connect, express, and share ideas with individuals worldwide. While these platforms have numerous positive effects, they are often plagued by the prevalence of hate speech and offensive language. Hate speech refers to any form of communication that aims to attack the dignity of a group based on characteristics such as race, gender, ethnicity, sexual orientation, nationality, religion, or other features [23]. According to the Pew Research Center, approximately 40% of social media users have encountered online harassment or bullying [6]. Between July and September 2021, Facebook detected and took action against 22.3 million instances of hate speech content [22]. These hate posts, which may seem harmless on social media, have real-world consequences, including violence and riots [6]. Therefore, it is crucial to prioritize the detection and control of hate speech.

Over the past decade, significant research has been conducted to develop models and datasets for automatic hate speech detection in the English language, utilizing traditional machine learning techniques [8,9,30] as well as deep learning techniques [1,2,37]. However, limited studies have been conducted for other languages, such as Italian [35], Indonesian [14], and Thai [26], primarily due to inadequate resources or conflicting interests. Given the variation in the perception of hate speech across different languages and cultures, it is crucial to develop automatic hate speech detection techniques for low-resource languages to improve classification and understanding of the corresponding contexts. According to a recent report by Reuters, Thailand has witnessed a rapid surge in hate speech incidents during the COVID-19 outbreak [34]. Specifically, the infection of many Myanmar workers at a fish market in Samut Sakhon led to the spread of hate speech against them on social media platforms, including YouTube, Facebook, and Twitter. Consequently, migrant and immigrant workers from Myanmar became extremely fearful for their safety. To address this issue, we have developed an advanced model for detecting online hate speech in the Thai language. Our goal is to automatically identify and flag hateful messages using these hate speech detection systems.

However, researchers have primarily focused on enhancing the performance of hate speech detection by utilizing various models but have largely overlooked the importance of explainability in these models. The emergence of explainable artificial intelligence (AI) [13] has made it necessary to provide explanations or interpretations for the decisions made by machine learning algorithms. This is crucial for building trust and confidence when deploying AI models in practical scenarios. Furthermore, legislation such as the General Data Protection Regulation (GDPR) [10] in Europe has introduced a "right to explanation" law, highlighting the need to develop interpretable models. As a result, there is a pressing demand to prioritize the development of interpretable models rather than solely focusing on model complexity for enhanced performance.

Multi-task learning is a training technique that utilizes data from related tasks to efficiently learn the relationship between them [5]. Numerous studies have demonstrated that incorporating an auxiliary task can enhance the performance of the primary task. For instance, in the context of cyberbullying detection [20], complaint identification [33], and tweet act classification [31], the inclusion of auxiliary tasks has proven beneficial. Considering that a person's sentiments and emotions can significantly impact the meaning of social media posts, it is crucial to incorporate sentiment and emotional analysis in hate speech detection.

Motivated by these considerations, we have developed the first explainable hate speech dataset, called "*Ex-ThaiHate*," in the Thai language. This dataset addresses four tasks simultaneously: hate speech detection (HSD), sentiment analysis (SA), emotion recognition (ER), and rationale detection (RD) — which focuses on providing explainability. To construct *Ex-ThaiHate*, we re-annotated the existing Thai Hate Speech dataset [26] by adding the sentiment and emotion labels and marking rationales. Rationales are text fragments from a source text that justify classification decisions. In cases where a post is a non-hate speech, we do not indicate any rationales. Our study specifically emphasizes the application of rationales to enhance model interpretability, aiming to achieve more human-like decision-making and improve the model's trustworthiness, transparency, and reliability. Previous studies, such as e-SNLI [4] and commonsense explanations [29], have also utilized rationales to enhance their models.

A typical multi-task model consists of a shared encoder that incorporates representations from data of different tasks, along with task-specific layers or heads attached to that encoder. However, this approach has several drawbacks. One such drawback is negative transfer, where multiple tasks, instead of optimizing the learning process, start to hinder the training process [7]. Additionally, there are concerns related to model capacity, wherein if the size of the shared encoder becomes too large, there will be no effective transfer of information across different tasks [38]. Furthermore, the optimization scheme for assigning weights to different tasks during training poses challenges [38].

To address the challenges mentioned earlier in multi-task learning, we have proposed the idea of employing a generative model to simultaneously solve multiple classification tasks in a text-to-text generation manner. In this work, we introduce a unified generative framework called "*GenX*," which is capable of solving all four tasks concurrently. The input to the *GenX* model is a social media post written in Thai, and the output target sequence is the concatenation of corresponding hate, sentiment, emotion labels, and rationales, separated by a special character. Through extensive experiments, we demonstrate that *GenX* consistently outperforms other baselines and state-of-the-art (SOTA) models across various evaluation metrics. The following is a summary of our contributions:

1. We investigate two new tasks: (i) explainable HSD in Thai and (ii) formulating the multi-task problem as a text-to-text generation problem.
2. We have developed *Ex-ThaiHate*, a new benchmark dataset for explainable HSD in the Thai language. This dataset includes sentiment and emotion

labels. To the best of our knowledge, this is the first study to focus on explainable HSD in Thai.
3. We propose a unified generative framework called "*GenX*" with reinforcement learning (RL) -based training to simultaneously solve four tasks: HSD, SA, ER, and RD.
4. Experimental results demonstrate that incorporating rationales, sentiment, and emotion information significantly enhances the performance of the main task, i.e., HSD.

2 Related Works

HSD heavily relies on linguistic subtleties, and researchers have recently devoted significant attention to automatically identifying hate speech in social media. In this section, we will review recent works on both stand-alone and multi-task learning-based methods for HSD.

Several studies have been conducted to develop and enhance algorithms for the detection of cyberbullying and hate speech in the English language. Reynolds et al. [30] utilized data from formspring.me to create a cyberbullying dataset and achieved an accuracy of 78.5% using the C4.5 decision tree method. In 2020, Balakrishnan et al. [3] developed a cyberbullying detection algorithm that employed multiple machine learning techniques while considering the psychological characteristics of Twitter users. Another notable system, CyberBERT, was proposed by Paul et al. [27], which utilized BERT-based models and demonstrated SOTA performance on benchmark hate speech datasets from Formspring (12k posts), Twitter (16k posts), and Wikipedia. Furthermore, Badjatiya et al. [2] conducted extensive experiments with various deep learning architectures to learn semantic word embeddings. Their results on a hate speech dataset consisting of 16K annotated tweets showed that deep learning methods outperformed traditional char/word n-gram algorithms by an 18% F1 score.

In 2021, Wanasukapunt et al. [36] developed both binomial models—Support Vector Machine (SVM), Random Forest (RF)—and multinomial models—Long short-term memory (LSTM), DistilBERT)—to detect abusive speech from social media specifically in the Thai language. Their study revealed that deep learning models outperformed machine learning models, and the best F1 score of 90.67% was achieved using DistilBERT. In a separate study, Pasupa et al. [26] constructed a benchmark Thai hate speech dataset by collecting posts from platforms such as Facebook, Twitter, and YouTube. They fine-tuned the WangchanBERTa model using the ordinal regression loss function, resulting in a SOTA performance for HSD in the Thai language. Recently, Maity et al. [18] introduced a two-channel deep learning model called FastThaiCaps. This model combines BERT embedding with a capsule network, as well as FastText embedding with BiLSTM and attention. Notably, extensive experiments demonstrated that their proposed model surpassed the performance of the baseline models.

In [40], the authors developed a multi-task framework that incorporates sentiment knowledge for HSD. Saha et al. [31] proposed a multi-modal tweet act classification framework. Their approach involves an ensemble adversarial learning strategy, where the inclusion of sentiment and emotion information improves the performance of the main task. Maity et al. [19] created a Hindi-English code-mixed dataset specifically for cyberbullying detection. They developed an attention-based deep multi-task framework based on BERT and VecMap embeddings.

Zaidan et al. [39] introduced the concept of rationales, which involves annotators underlining a section of text to support their tagging decision. The authors found that using these rationales improved the performance of sentiment classification. In a similar vein, Mathew et al. [21] introduced the HateXplain benchmark dataset for HSD. They discovered that models trained using human rationales were more effective at reducing inadvertent bias against targeted communities. Karim et al. [15] developed an explainable HSD approach (DeepHate-Explainer) in Bengali based on different variants of transformer architectures (BERT-base, mMERT, XLM-RoBERTa). They provided explainability by highlighting the most important words for which the sentence is labeled as hate speech.

After conducting an in-depth literature review, it can be concluded that the majority of research on HSD focuses on the English language. It has been observed that incorporating sentiment and emotional information greatly improves the performance of the primary task. However, there is a notable absence of studies investigating sentiment and emotion-aided HSD in the Thai language.

3 *Ex-ThaiHate* Dataset Development

To start the process, we conducted a literature review to identify existing Thai hate speech datasets. Our search yielded two relevant Thai datasets [26,36]. After careful consideration, we decided to use the Thai Hate Speech dataset by Pasupa et al. [26] for further annotation with sentiment and emotion labels. This dataset was collected from three widely used social media platforms: Facebook, Twitter, and YouTube. The data collection period spanned from 18/12/2020 to 23/12/2020, following the news of a COVID-19 infection case involving a merchandiser at a market in Samut Sakhon, Thailand, who was subsequently admitted to a hospital.

3.1 Data Annotation

The annotation process was carried out by a team consisting of three Ph.D. scholars specializing in cyberbullying, hate speech, and offensive content, and three undergraduate students who were proficient in the Thai language. To recruit undergraduate students, we sent out a voluntary hiring notice through the school's email list, and they were compensated with gift vouchers for their participation. Initially, the Thai hate speech dataset [26] had been annotated with a binary hate speech class (Hate/non-hate). In order to train the annotators for the annotation of sentiment and emotion classes, we needed gold-standard samples with these annotations. Our expert annotators randomly

Table 1. Samples from annotated *ExThaiHate* dataset. The underlined tokens provide the rationale behind the hate speech.

<table>
<tr><th>Post</th><th>Sentiment Class</th><th>Emotion Class</th><th>Hate Speech Class</th></tr>
<tr><td><b>T1</b>: สงสารพม่าเขามากเลย สู้นะพม่าทุกคน
<b>Token</b>: สงสารพม่า|เขา|มากเลย|สู้|นะ|พม่า|ทุกคน
<b>Translation</b>: I feel pity for the Burmese so much. Keep fighting, all Burmese.</td><td>Negative</td><td>Sadness</td><td>Non-hate Speech</td></tr>
<tr><td><b>T2</b>:จะรับมาให้โรคติด ให้ตายกันหมดประเทศหรือไง แล้วให้เหลือแต่พม่าครองเมือง
<b>Token</b>: จะรับ|มาให้|โรค|ติด|ให้ตาย|กันหมด|ประเทศ|หรือไง|แล้ว|ให้|เหลือแต่|พม่า|ครองเมือง
<b>Translation</b>: Accepting Burmese into the country to allow the disease to infect us and people in the country to die? Then leave only Burmese to rule the city.</td><td>Negative</td><td>Anger</td><td>Hate Speech</td></tr>
<tr><td><b>T3</b>: สมัยนี้พม่ายังตีไทยอีกนะ
<b>Token</b>: สมัยนี้|พม่า|ยัง|ตี|ไทย|อีกนะ
<b>Translation</b>: Burma still invades Thailand these days.</td><td>Negative</td><td>Disagreeable</td><td>Hate Speech</td></tr>
</table>

selected 600 samples from the dataset and highlighted specific words (rationales) for providing textual explanations. They also assigned suitable sentiment labels (Positive/Neutral/Negative) and emotion labels based on Plutchik's eight emotion categories (Sadness, Joy, Surprise, Fear, Disgust, Anger, Anticipation, and Trust). For the rationale annotation, we followed the same strategy as mentioned in [21], where each word in a tweet was marked with either 0 or 1, with 1 indicating the presence of a rationale. During the emotion class annotation, we observed that out of the eight emotion categories, only four (Anger, Trust, Sadness, and Anticipation) were utilized, and a significant portion of the samples fell into the "Other" category. Upon reviewing the "Other" category samples, we found that many of them were of a disagreeable nature. Based on this observation, we introduced the additional emotion class of "disagreeable" in our Thai hate speech dataset. Throughout the annotation process, expert annotators had discussions to resolve any differences in their annotations and ensure consistency. This resulted in the creation of 600 gold standard samples with annotations for sentiment, emotion, hate speech, and rationales. These 600 annotated examples were divided into three sets, each containing 200 samples, to facilitate a three-phase training approach. After each phase of training, expert annotators met with novice annotators to correct any incorrect annotations and update the annotation guidelines. Upon completing the third round of training, the top three annotators were selected to annotate the entire dataset.

We initiated our main annotation process with a small batch of 100 samples and later raised it to 500 as the annotators became well-experienced with the tasks. We tried to maintain the annotators' agreement by correcting some errors they made in the previous batch. On completion of each set of annotations, final sentiment and emotion labels were decided by the majority voting method. If the selections of three annotators vary, we enlist the help of an expert annotator to break the tie. We also directed annotators to annotate the posts without regard for any particular demography, religion, or other factors. We use the Fleiss' Kappa [11] score to calculate the inter-annotator agreement (IAA) to affirm the annotation quality. IAA obtained scores of 0.79, 0.72, and 0.74 for sentiment, emotion, and rationales labels, respectively, signifying the dataset being of acceptable quality.

Table 1 presents a selection of samples obtained from the *Ex-ThaiHate* dataset. The dataset comprises a total of 7,597 posts, with 2,685 posts labeled as hate and 4,912 posts marked as non-hate. Class-wise statistics of the *Ex-ThaiHate* dataset can be found in Table 2.

Table 2. Dataset Statistics of different classes of *Ex-ThaiHate* dataset

Total Samples	Hate Speech		Sentiment			Emotion					
	Hate	Non-Hate	Positive	Neutral	Negative	Anger	Trust	Sadness	Disagreeable	Anticipation	Others
7597	2685	4912	2655	2257	2685	2133	2020	251	482	160	2551

4 Methodology

This section presents our proposed *GenX* model, shown in Fig. 1, for sentiment- and emotion-aware HSD with explainability in the Thai language.

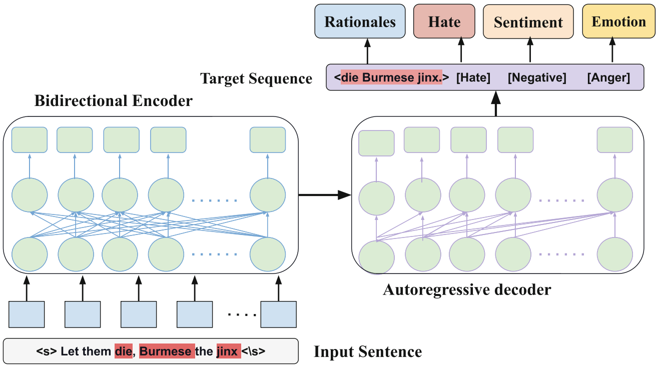

Fig. 1. *GenX* architecture

4.1 Redefining Explainable HSD Task as Text-to-Text Generation Task

Traditional multi-tasking methods leverage separate task-specific heads for different tasks making them difficult to add a new task to the model without having that task's specific head. Here, we propose a text-to-text generation paradigm for solving explainable HSD and other auxiliary tasks in a single unified manner. To transform this problem into a text generation problem, we first construct a natural language target sequence, Y_i, for input sentence, X_i, for training purposes by concatenating all the labels of all four tasks. For the rationale detection task, we only consider those $\{r\}$s which belong to the offensive label set in R_{Labels} represented by R_{Off}. In case of an empty offensive label set, we will use a NONE token to represent 0 offensive tokens in the text. Finally, the target sequence Y_i is represented as:

$$Y_i = \{< R_{Off} >< b >< s >< e >\} \tag{1}$$

where R_{Off}, b, s, and e represent the corresponding rationales, hate, sentiment, and emotion labels of an input post, X_i.

We have added special characters <> after each task's prediction, as shown in (1) so that we can extract task-specific predictions during testing or inference. Now, both the input sentence and the target are in the form of natural language to leverage large pre-trained sequence-to-sequence models for solving this task of text-to-text generation. The problem can be reformulated as given an input sequence X, the task is to generate an output sequence, Y', containing all the predictions defined in (1) using a generative model defined as $Y' = G(X)$, where G is a generative model. The advantage of this approach is that now we can add any new task just by concatenating that task's labels to the target sequence Y or solve any subtask with ease.

4.2 Sequence-to-Sequence Learning (Seq2Seq)

This problem of text-to-text generation can easily be solved with the help of a sequence-to-sequence model, which consists of two modules: 1) Encoder and 2) Decoder. We employed the pre-trained BART [16] and T5 [28] models as the sequence-to-sequence models. BART and T5 are encoder-decoder-based transformer models, mainly pre-trained for text generation tasks such as summarization and translation. As we are working on the Thai language so, multilingual BART (mBART) and T5 (mT5) have been used for the experiment. We delineate the training and inference process for sequence-to-sequence learning as follows.

Training Process. We are given a pair of input sentences and target sequence (X, Y), the first step is to feed $X = \{x_0, x_1, \ldots, x_i, \ldots, x_n\}$ to the encoder module to obtain the hidden representation of input as

$$H_{EN} = G_{Encoder}(\{x_0, x_1, \ldots, x_i, \ldots, x_n\}), \tag{2}$$

where $G_{Encoder}$ represents encoder computations.

After obtaining the hidden representation, H_{EN}, we will feed H_{EN} and all the output tokens till time step $t-1$ represented as $Y_{<t}$ to the decoder module to obtain the hidden state at time step t as

$$H^t_{DEC} = G_{Decoder}(H_{EN}, Y_{<t}), \tag{3}$$

where $G_{Decoder}$ denotes the decoder computations.

The conditional probability for the predicted output token at t^{th} time step, given the input and previous $t-1$ predicted tokens, is calculated by applying the Softmax function over the hidden state, H^t_{DEC}, as follows:

$$P(Y'_t|X, Y_{<t}) = F_{Softmax}(H^t_{DEC} W_{Gen}), \tag{4}$$

where $F_{Softmax}$ represents Softmax computation and W_{Gen} denotes weights of our model.

Training Objective. We initialize the weights W_{Gen} for our model with the pre-trained weights of the pre-trained sequence-to-sequence generative models (T5 or BART). We then fine-tune the model with negative log-likelihood, i.e., the maximum likelihood estimation (MLE) objective function in a supervised manner to optimize the weights, W_{Gen} as

$$\max_{W_{Gen}} \prod_{t=0}^{T} P(Y'_t|X, Y_{<t}). \tag{5}$$

In the context of transformers, MLE typically involves finding the best weights for the model's layers that maximize the probability of observing a given sequence of tokens in a training dataset. The loss function takes into account the information from earlier time steps in the decoder by considering the cumulative error in the model's predictions over all time steps. Further, we have incorporated RL-based Training to enhance the performance of the *GenX* model.

RL-Based Training. On top of the MLE objective function, we also employ a reward-based training objective function. Inspired from [32], we use a BLEU [25] based reward function. We define BLEU based Reward R_{BLEU} as:

$$R_{BLEU} = (BLEU(Y'_i, Y_i) - BLEU(Y^g_i, Y_i)), \tag{6}$$

where Y'_i denotes the output sequence sampled from the conditional probability distribution at each decoding time stamp and Y^g_i denotes the output sequence obtained by greedily maximizing the conditional probability distribution at each time step.

To maximize the expected reward, R_{BLEU} of Y'_i, we use the policy gradient technique, which is defined as

$$\nabla_\theta J(\theta) = R_{BLEU} \cdot \nabla_\theta \log P(Y'_i|X_i; \theta). \tag{7}$$

Inference. During the training process, we have access to both the input sentence, X, and the target sequence, Y. Thus, we train the model using the teacher forcing approach, i.e., using the target sequence as the input instead of tokens predicted at prior time steps during the decoding process. However, the inference must be done in an autoregressive manner as we do not have access to target sequences to guide the decoding process replacing $Y_{<t}$ with $Y'_{<t}$ in (3)–(5) where $Y'_{<t}$ represents tokens predicted till time step $t-1$. So we use the beam search algorithm to obtain the predicted sequence, Y', as it considers multiple alternative options based on the hyperparameter beamwidth (B) which is optimal than a simple greedy search technique which only selects the single best token at each time step. In beam search, the decoder generates a set of candidate output sequences in parallel, each with a different starting token. At each time step, the decoder calculates the probability distribution over the vocabulary for each candidate sequence and generates a set of new candidate sequences by extending each existing candidate sequence with the top K most likely next tokens, where K is the beam size. The candidate sequences are ranked based on their accumulated probabilities, and the K sequences with the highest probabilities are kept for the next time step.

5 Experimental Results and Analysis

This section describes the outcomes of various baseline models and our proposed model, tested on the *Ex-ThaiHate* dataset. The experiments are intended to address the following research questions: **RQ1** How is the performance of our *GenX* model for HSD over the SOTA machine learning models? **RQ2** How does multi-tasking help in enhancing the performance of HSD with the help of additional rationale, sentiment and emotion information? **RQ3** What is the effect of the BLEU-based reward function in RL-based training? **RQ4** To handle noisy social media Thai data, which embedding is better, BERT or FastText?

5.1 Experimental Settings and Baselines Setup

We split our dataset into 80% train, 10% validation, and 10% test sets. We experimented with mBART and mT5 and attained optimal performance with mBART. During training, we trained for a total of 20 epochs and used the Adam optimizer with a weight decay of 1e−3 (to avoid overfitting).

Classification Baselines. (i) Standard machine learning baselines as mentioned in [36], i.e., Naïve Bayes, SVM, and RF have been used for our experiments. We used the pooled result of dimension 768 returned by WangchanBERTa as input for machine learning-based baselines. On the other hand, for FastText embedding, we first tokenized the phrase using PyThaiNLP[1], then we extracted the embedding of each token from the pre-trained Thai FastText model, and we averaged it out to represent the full sentence by a 300-dimensional vector. (ii) We passed the pooled output from BERT through a Fully Connected (FC) layer that consisted of 100 neurons. Then, we utilized a Softmax output layer to generate the final prediction probabilities. (iii) We pass input text to BiLSTM followed by the attention layer [17]. Attended features of the text are passed through a dense layer to predict the labels.

[1] https://pythainlp.github.io/docs/2.2/.

Rationales Detection Baselines. (i) To comprehensively evaluate our proposed *GenX* model for the RD task, we established a baseline by selecting a Bidirectional Long Short-Term Memory-Conditional Random Field (BiLSTM-CRF) model [24], as this task involves sequence learning. The BiLSTM-CRF model has three components: a word embedding layer, a Bidirectional Long Short-Term Memory network (BiLSTM), and a Conditional Random Field (CRF). We used the sequence output of mBERT and WangchanBERTa (wBERT) as word embeddings. The BiLSTM network captures complete contextual information, while the CRF model predicts the label sequence.

There are four multi-task variants based on how many tasks we want to solve simultaneously, e.g., HSD+RD, HSD+RD+SA, HSD+RD+ER, HSD+RD+SA+ ER, etc. It should be noted that the *GenX* model can be used for both single and multi-task settings. The only difference in a single-task setting is that the target sequence contains token/tokens specific to the task being addressed. In contrast, in a multi-task setting, the target sequence is formed by concatenating all labels (tokens), with each token corresponding to a specific task.

5.2 Findings from Experiments

Table 3 presents the performance of machine learning baselines, different variants of single-task and multi-task frameworks in terms of accuracy (Acc), and weighted F1 score. Table 4 presents the results of the RD task. For the quantitative assessment of the RD task, we used the Jaccard Similarity (JS), Hamming Distance (HD), and Ratcliff-Obershelp Similarity (ROS) metrics as mentioned in [12]. The following are the findings from our experimental results presented in Tables 3 and 4:

- **RQ1:** Our proposed *GenX* model, in both single-task and multi-task settings, surpasses all machine learning-based baselines by a considerable margin. The MT(RD+HSD+SA+ER)+RL with mBART outperformed the best ML baseline (BERT-SVM) by 6.6%, 9.1%, and 15.0% for the HSD, SA, and ER tasks, respectively. Furthermore, *GenX* outperforms the deep learning-based baseline BiLSTM-Attn by a significant margin.
- **RQ2:** The MT(RD+HSD+SA+ER)+RL model with mBART shows better performance than ST-GenX, with accuracy improvements of 3.2%, 2.1%, and 2.0% for HSD, SA, and ER tasks, respectively. These findings suggest that incorporating sentiment and emotion knowledge significantly enhances the performance of the HSD task.
- Comparing the proposed *GenX* model, based on text-to-text generation, with the BiLSTM+CRF model (Classical Named Entity Recognition model), we observe that *GenX* outperforms BiLSTM+CRF for the RD task (see Table 4). This result demonstrates the effectiveness of utilizing a text-to-text generation model to solve two distinct categories of tasks, classification tasks (HSD, SA, ER), and sequence labeling tasks (RD), simultaneously with a single model.

Table 3. Results of different baselines, SOTA, and proposed frameworks for Hate speech detection (HSD), sentiment analysis (SA), and emotion recognition (ER) tasks; wBERT: WangchanBERTa, mBERT: Multilingual BERT; MT: Multi-Task; ST: Single Task

Embedding	Model	Hate		Sentiment		Emotion	
		Acc	F1	Acc	F1	Acc	F1
Baselines							
BERT	Naïve Bayes	75.32	76.21	64.73	65.55	57.96	56.78
	SVM	83.22	83.26	70.98	71.23	60.53	60.18
	Random Forest	77.03	74.08	64.53	64.32	59.67	57.72
FastText	Naïve Bayes	72.56	72.45	58.35	58.71	52.43	49.83
	SVM	81.71	81.65	70.98	71.32	66.11	64.37
	Random Forest	80.92	79.53	67.30	67.87	62.56	60.13
SOTA							
BERT	Fine-tune	85.87	85.83	–	–	–	–
Deep Learning Baselines							
mBERT	FC	76.15	78.32	63.27	63.18	62.15	62.76
	BiLSTM-Attn	82.36	82.64	64.16	65.28	63.47	64.33
wBERT	FC	77.58	75.61	62.16	60.12	62.45	62.21
	BiLSTM-Attn	83.45	84.78	66.38	67.39	65.63	64.89
Proposed Model – *GenX*							
mBART	**ST**	**85.48**	**85.34**	**78.47**	**78.64**	**75.23**	**75.54**
	MT(HSD+SA)	87.67	87.43	79.34	79.58	–	–
	MT(HSD+ER)	86.84	86.66	–	–	75.92	75.88
	MT(HSD+SA+ER)	88.53	87.94	79.57	79.63	76.54	76.48
	MT(RD+HSD+SA)	86.54	86.50	79.63	79.46	–	–
	MT(RD+HSD+SA)+RL	87.46	87.42	80.48	80.31	–	–
	MT(RD+HSD+ER)	86.63	86.45	–	–	76.04	75.98
	MT(RD+HSD+ER)+RL	87.55	87.46	–	–	76.85	78.45
	MT(RD+HSD+SA+ER)	87.74	86.84	79.74	79.86	76.42	76.53
	MT(RD+HSD+SA+ER)+RL	**88.67**	**88.21**	**80.57**	**80.46**	**77.24**	**79.37**
T5	ST	84.93	85.26	77.49	77.63	72.68	72.14
	MT(HSD+SA)	86.24	85.78	77.14	76.89	–	–
	MT(HSD+ER)	86.43	86.11	–	–	72.44	72.37
	MT(HSD+SA+ER)	86.75	85.69	77.34	77.47	72.83	72.11
	MT(RD+HSD+SA)	86.07	85.94	78.16	77.83	–	–
	MT(RD+HSD+ER)	86.41	86.34	–	–	73.32	74.11
	MT(RD+HSD+SA+ER)	86.48	86.44	78.43	78.64	73.59	74.26
Improvements over ST		3.19	2.87	2.10	1.82	2.01	3.83
Improvements over SOTA		2.80	2.38	–	-	–	–

- **RQ3:** We observe that RL-based training improves performance by an average of 1.0% for all tasks. We report the results with RL only for those task combinations where RD is included, as without the RD task, the target string

Table 4. Results of different baselines and proposed frameworks for Rationales Detection (RD) task; JS: Jaccard Similarity, HD: Hamming Distance, and ROS: Ratcliff-Obershelp Similarity

Embedding	Model	Rationales		
		JS	HD	ROS
Baselines				
mBERT	BiLSTM+CRF	59.24	50.95	65.28
wBERT	BiLSTM+CRF	60.13	51.93	65.86
Proposed Model - GenX				
mBART	ST	62.19	53.48	69.56
	ST+RL	63.31	55.47	71.25
	MT(HSD+RD)	65.67	57.37	73.25
	MT(HSD+RD)+RL	66.45	58.07	74.01
	MT(HSD+RD+SA)	65.78	57.42	73.35
	MT(HSD+RD+SA)+RL	66.53	58.18	74.13
	MT(HSD+RD+ER)	65.81	57.46	73.24
	MT(HSD+RD+ER)+RL	66.50	58.08	74.03
	MT(HSD+RD+SA+ER)	65.87	57.64	73.36
	MT(HSD+RD+SA+ER)+RL	**66.60**	**58.22**	**74.16**

has a very minimal length, i.e., 2 or 3. To prevent the model from generating sentences with out-of-sentence vocabulary, we use BLEU similarity measures. Training the model with this reward function encourages the generation of sequences with high overlap with the target sequence, leading to improved results in the RD tasks.

- **RQ4:** Comparing the individual performance between BERT and FastText embedding, we find that BERT consistently outperforms FastText for all tasks, except for Random Forest. Another noteworthy finding is that wBERT outperforms mBERT, indicating wBERT's greater efficiency in handling Thai data than mBERT. Additionally, between the two generative models, BART achieved better results, which is why we only reported the RL variants and RD task results with mBART settings.
- The proposed mBART-GenX model outperforms the SOTA with an improved F1 score of 2.4% for the HSD task. This result demonstrates the efficacy of our proposed model.

We have conducted a statistical t-test on the results of ten different runs of our proposed model and other baselines and obtained a p-value less than 0.05.

5.3 Error Analysis

We conducted an analysis of prediction errors for hate speech by randomly selecting the results of the multi-task model from one out of ten trials. We have identified two primary concerns related to the sentiment and emotion predictions of multi-task models as follows.

1. The model was confused in predicting negative sentiment 22.4% (37/165) of the statements with the following observation: (i) Predicted negative as neutral 46.0% (17/37) of the statements. Most were found to be caused by ambiguous or metaphorical words that can be used in ironic or sarcastic contexts, for example, "พม่าพี่น้องชาวสมุทรสาครทำพิษซะแล้ว" (Our Burmese siblings in Samut Sakhon have already caused trouble). The word "ทำพิษ" (trouble) is a metaphor. This makes it difficult for the model to determine the true sentiment, thus predicting neutral instead of negative.
2. The model incorrectly predicted 30.5% (57/187) emotion classes of the statements. Most of them predicted anger emotion incorrectly in 86.0% (49/57) of the statements. We observed that the model predicted anger as disagreeable in 22.5% (11/49) of the statements. For example, "ตอนนี้คือจิตตกแล้ว ตกงานมาก็จะปีแล้ว จะอดตายแล้ว ยังมามีรอบสองรอบนี้คือตายทั้งขึ้นทั้งร่อง เอาพม่ากลับบ้านเกิดเลยค่ะ มาทางไหนก็กลับไปก่อนเลย ป้องกันใส่แมสจะตายทุกวันสุดท้ายเสียเปล่ามาก" (Now I am very depressed. It has been a year since I lost my job, and I am about to starve to death. Bring Burmese people back to their country. Go back the way you came. I protect myself by wearing a mask every day. In the end, it is very wasteful.), the author expressed feelings of injustice to the Thai people, which is often accompanied by anger emotion. It should be noted that the message is very long and complex.

6 Conclusion and Future Works

The present study addresses the issue of HSD in the Thai language, with a focus on the aspect of explainability. The current work contributes in two main ways: (a) the development of the first-ever explainable HSD dataset in the Thai language, which includes annotations of rationale/phrases used for explainability, as well as hate, sentiment, and emotion labels; (b) the proposal of a unified generative framework, called *GenX*, with RL-based training, to simultaneously solve four tasks: HSD, SA, ER, and RD. This work demonstrates how a multi-task problem can be formulated as a text-to-text generation task, leveraging the knowledge of large pre-trained sequence-to-sequence models in low-resource language settings. Experimental results showcase the superiority of the proposed model over baselines and its outperformance of the SOTA, achieving an improved accuracy score of 2.8% for the hate speech task.

In future works, efforts will be made to extend explainable HSD to a multi-modal setting by considering both image and text modalities.

Acknowledgments. This work was supported by the Ministry of External Affairs (MEA) and the Department of Science & Technology (DST), India, under the ASEAN-India Collaborative R&D Scheme.

References

1. Agrawal, S., Awekar, A.: Deep learning for detecting cyberbullying across multiple social media platforms. In: Pasi, G., Piwowarski, B., Azzopardi, L., Hanbury, A. (eds.) ECIR 2018. LNCS, vol. 10772, pp. 141–153. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-76941-7_11
2. Badjatiya, P., Gupta, S., Gupta, M., Varma, V.: Deep learning for hate speech detection in tweets. In: Barrett, R., Cummings, R., Agichtein, E., Gabrilovich, E. (eds.) Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, April 3–7, 2017, pp. 759–760. ACM (2017). https://doi.org/10.1145/3041021.3054223
3. Balakrishnan, V., Khan, S., Arabnia, H.R.: Improving cyberbullying detection using twitter users' psychological features and machine learning. Comput. Secur. **90**, 101710 (2020). https://doi.org/10.1016/j.cose.2019.101710
4. Camburu, O., Rocktäschel, T., Lukasiewicz, T., Blunsom, P.: e-snli: natural language inference with natural language explanations. In: Bengio, S., Wallach, H.M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3–8, 2018, Montréal, Canada, pp. 9560–9572 (2018)
5. Caruana, R.: Multitask learning. Mach. Learn. **28**(1), 41–75 (1997). https://doi.org/10.1023/A:1007379606734
6. Chan, T.K.H., Cheung, C.M.K., Wong, R.Y.M.: Cyberbullying on social networking sites: the crime opportunity and affordance perspectives. J. Manag. Inf. Syst. **36**(2), 574–609 (2019). https://doi.org/10.1080/07421222.2019.1599500
7. Crawshaw, M.: Multi-task learning with deep neural networks: a survey. CoRR abs/2009.09796 (2020)
8. Dadvar, M., Trieschnigg, D., de Jong, F.: Experts and machines against bullies: a hybrid approach to detect cyberbullies. In: Sokolova, M., van Beek, P. (eds.) AI 2014. LNCS (LNAI), vol. 8436, pp. 275–281. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-06483-3_25
9. Dinakar, K., Reichart, R., Lieberman, H.: Modeling the detection of textual cyberbullying. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 5, pp. 11–17 (2021). https://doi.org/10.1609/icwsm.v5i3.14209
10. European Parliament and of the Council: Protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec. EC General Data Protection Regulation 679 (2016)
11. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychol. Bull. **76**(5), 378–382 (1971)
12. Ghosh, S., Roy, S., Ekbal, A., Bhattacharyya, P.: CARES: CAuse recognition for emotion in suicide notes. In: Hagen, M., et al. (eds.) ECIR 2022. LNCS, vol. 13186, pp. 128–136. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-99739-7_15
13. Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., Yang, G.: XAI - explainable artificial intelligence. Sci. Robot. **4**(37) (2019). https://doi.org/10.1126/scirobotics.aay7120
14. Ibrohim, M.O., Budi, I.: Multi-label hate speech and abusive language detection in Indonesian Twitter. In: Proceedings of the Third Workshop on Abusive Language Online, Florence, Italy, pp. 46–57. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/W19-3506

15. Karim, M.R., et al.: Deephateexplainer: explainable hate speech detection in under-resourced Bengali language. In: 8th IEEE International Conference on Data Science and Advanced Analytics, DSAA 2021, Porto, Portugal, October 6–9, 2021, pp. 1–10. IEEE (2021). https://doi.org/10.1109/DSAA53316.2021.9564230
16. Lewis, M., et al.: BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J.R. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, 5–10 July 2020, pp. 7871–7880. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.acl-main.703
17. Liu, B., Lane, I.R.: Attention-based recurrent neural network models for joint intent detection and slot filling. In: Morgan, N. (ed.) Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, 8–12 September 2016, pp. 685–689. ISCA (2016). https://doi.org/10.21437/Interspeech.2016–1352
18. Maity, K., Bhattacharya, S., Saha, S., Janoai, S., Pasupa, K.: Fastthaicaps: a transformer based capsule network for hate speech detection in Thai language. In: Tanveer, M., Agarwal, S., Ozawa, S., Ekbal, A., Jatowt, A. (eds.) ICONIP 2022, Part II. LNCS, vol. 13624, pp. 425–437. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-30108-7_36
19. Maity, K., Kumar, A., Saha, S.: A multitask multimodal framework for sentiment and emotion-aided cyberbullying detection. IEEE Internet Comput. **26**(4), 68–78 (2022). https://doi.org/10.1109/MIC.2022.3158583
20. Maity, K., Saha, S.: A multi-task model for sentiment aided cyberbullying detection in code-mixed Indian languages. In: Mantoro, T., Lee, M., Ayu, M.A., Wong, K.W., Hidayanto, A.N. (eds.) ICONIP 2021. LNCS, vol. 13111, pp. 440–451. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-92273-3_36
21. Mathew, B., Saha, P., Yimam, S.M., Biemann, C., Goyal, P., Mukherjee, A.: Hatexplain: a benchmark dataset for explainable hate speech detection. CoRR abs/2012.10289 (2020)
22. Meta: Community standards enforcement – hate speech. Meta Transparency Centre (2022), https://transparency.fb.com/data/community-standards-enforcement/hate-speech. Accessed 1 Apr 2023
23. Nockleby, J.T.: Hate speech in context: the case of verbal threats. Buffalo Law Rev. **42**, 653–713 (1994)
24. Panchendrarajan, R., Amaresan, A.: Bidirectional LSTM-CRF for named entity recognition. In: Politzer-Ahles, S., Hsu, Y., Huang, C., Yao, Y. (eds.) Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation, PACLIC 2018, Hong Kong, 1–3 December 2018. Association for Computational Linguistics (2018)
25. Papineni, K., Roukos, S., Ward, T., Zhu, W.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6–12, 2002, Philadelphia, PA, USA, pp. 311–318. ACL (2002). https://doi.org/10.3115/1073083.1073135
26. Pasupa, K., Karnbanjob, W., Aksornsiri, M.: Hate speech detection in Thai social media with ordinal-imbalanced text classification. In: 19th International Joint Conference on Computer Science and Software Engineering, JCSSE 2022, Bangkok, Thailand, June 22–25, 2022, pp. 1–6. IEEE (2022). https://doi.org/10.1109/JCSSE54890.2022.9836312
27. Paul, S., Saha, S.: Cyberbert: BERT for cyberbullying identification. Multimedia Syst. **28**(6), 1897–1904 (2022). https://doi.org/10.1007/s00530-020-00710-4

28. Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**, 140:1-140:67 (2020)
29. Rajani, N.F., McCann, B., Xiong, C., Socher, R.: Explain yourself! leveraging language models for commonsense reasoning. CoRR abs/1906.02361 (2019)
30. Reynolds, K., Kontostathis, A., Edwards, L.: Using machine learning to detect cyberbullying. In: 2011 10th International Conference on Machine Learning and Applications and Workshops, vol. 2, pp. 241–244 (2011). https://doi.org/10.1109/ICMLA.2011.152
31. Saha, T., Upadhyaya, A., Saha, S., Bhattacharyya, P.: A multitask multimodal ensemble model for sentiment- and emotion-aided tweet act classification. IEEE Trans. Comput. Soc. Syst. **9**(2), 508–517 (2022). https://doi.org/10.1109/TCSS.2021.3088714
32. Sancheti, A., Krishna, K., Srinivasan, B.V., Natarajan, A.: Reinforced rewards framework for text style transfer. In: Jose, J.M., et al. (eds.) ECIR 2020, Part I. LNCS, vol. 12035, pp. 545–560. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-45439-5_36
33. Singh, A., Saha, S., Hasanuzzaman, M., Dey, K.: Multitask learning for complaint identification and sentiment analysis. Cogn. Comput. **14**(1), 212–227 (2022). https://doi.org/10.1007/s12559-021-09844-7
34. Thepgumpanat, P., Naing, S., Tostevin, M.: Anti-myanmar hate speech flares in thailand over virus. Reuters (2020). https://www.reuters.com/article/us-health-coronavirus-thailand-myanmar-idUSKBN28Y0KS. Accessed 1 Apr 2023
35. Vigna, F.D., Cimino, A., Dell'Orletta, F., Petrocchi, M., Tesconi, M.: Hate me, hate me not: Hate speech detection on facebook. In: Armando, A., Baldoni, R., Focardi, R. (eds.) Proceedings of the First Italian Conference on Cybersecurity (ITASEC17), Venice, Italy, 17–20 January 2017. CEUR Workshop Proceedings, vol. 1816, pp. 86–95. CEUR-WS.org (2017)
36. Wanasukapunt, R., Phimoltares, S.: Classification of abusive Thai language content in social media using deep learning. In: 18th International Joint Conference on Computer Science and Software Engineering, JCSSE 2021, Lampang, Thailand, 30 June–2 July 2021, pp. 1–6. IEEE (2021). https://doi.org/10.1109/JCSSE53117.2021.9493829
37. Waseem, Z., Hovy, D.: Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In: Proceedings of the Student Research Workshop, SRW@HLT-NAACL 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, 12–17 June 2016, pp. 88–93. The Association for Computational Linguistics (2016). https://doi.org/10.18653/v1/n16-2013
38. Wu, S.: Emmental: a framework for building multimodal multi-task learning systems (2019)
39. Zaidan, O., Eisner, J., Piatko, C.D.: Using "annotator rationales" to improve machine learning for text categorization. In: Sidner, C.L., Schultz, T., Stone, M., Zhai, C. (eds.) Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, April 22–27, 2007, Rochester, New York, USA, pp. 260–267. The Association for Computational Linguistics (2007), https://aclanthology.org/N07-1033/

40. Zhou, X., et al.: Hate speech detection based on sentiment knowledge sharing. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, 1–6 August 2021, pp. 7158–7166. Association for Computational Linguistics (2021). https://doi.org/10.18653/v1/2021.acl-long.556