ViHateT5: Enhancing Hate Speech Detection in Vietnamese With a Unified Text-to-Text Transformer Model

Anonymous ACL submission

Abstract

Recent advancements in hate speech detection (HSD) in Vietnamese have made significant progress, primarily attributed to the emergence of transformer-based pre-trained language models, particularly those built on the BERT architecture. However, the necessity for specialized fine-tuned models has resulted in the complexity and fragmentation of developing a multitasking HSD system. Moreover, most current methodologies focus on fine-tuning general pre-trained models, primarily trained on formal textual datasets like Wikipedia, which may not accurately capture human behavior on online platforms. In this research, we introduce ViHateT5, a T5-based model pre-trained on our proposed large-scale domain-specific dataset named VOZ-HSD. By harnessing the power of a text-to-text architecture, ViHateT5 can tackle multiple tasks using a unified model and achieve state-of-the-art performance across all standard HSD benchmarks in Vietnamese. Our experiments also underscore the significance of label distribution in pre-training data on model efficacy. We provide our experimental materials for research purposes, including the VOZ-HSD dataset[1], pre-trained checkpoint[2], the unified HSD-multitask ViHateT5 model[3], and related source code on GitHub[4].

Warning: This paper contains examples from actual content on social media platforms that could be considered toxic and offensive.

1 Introduction

Hate speech refers to harmful expression targeting individuals or groups based on their inherent characteristics, potentially inciting violence or discrimination (Brown, 2017). Its detrimental impact on mental well-being includes differnet levels of anxiety, depression, or stress among affected individuals (Ghafoori et al., 2019). Due to the rise of social media on the internet, where people can easily leave their toxic content that may negatively affect anyone who reads it, the consequences that hate speech brings to use become worse and worse. To address these issues, automatic systems have been explored to detect harmful content online and mitigate its dissemination (Gitari et al., 2015; MacAvaney et al., 2019; Aïmeur et al., 2023).

Different languages have unique forms of harmful expressions, necessitating specific text processing methodologies for developing HSD systems. In the context of English, one of the most prevalent languages, there exist several effective strategies for addressing HSD-related tasks, such as employing machine learning models (Abro et al., 2020) or deep learning models for identifying harmful content (Badjatiya et al., 2017; Zimmerman et al., 2018). Furthermore, transfer learning approaches have garnered considerable research interest, showcasing remarkable performance in hate speech detection tasks (Ali et al., 2022; Mozafari et al., 2020). In the case of low-resource languages, numerous studies have been conducted, yielding promising results in tackling this issue (Bigoulaeva et al., 2021; Nkemelu et al., 2022; Arango Monnar et al., 2022).

Vietnamese, considered a low-resource language, has seen limited research in natural language processing concerning large-scale datasets and pre-trained models. Recent efforts in hate speech detection tasks based on Vietnamese language characteristics have yielded significant achievements (Vu et al., 2020; Luu et al., 2021; Hoang et al., 2023). However, current state-of-the-art models only fine-tune general transformer-based models, which may have been pre-trained on formal textual data sources (Nguyen et al., 2020). Moreover, even pre-trained on social media text, models like ViSoBERT (Nguyen et al., 2023) still necessitate

[1]Link provided upon acceptance
[2]Link provided upon acceptance
[3]Link provided upon acceptance
[4]Link provided upon acceptance

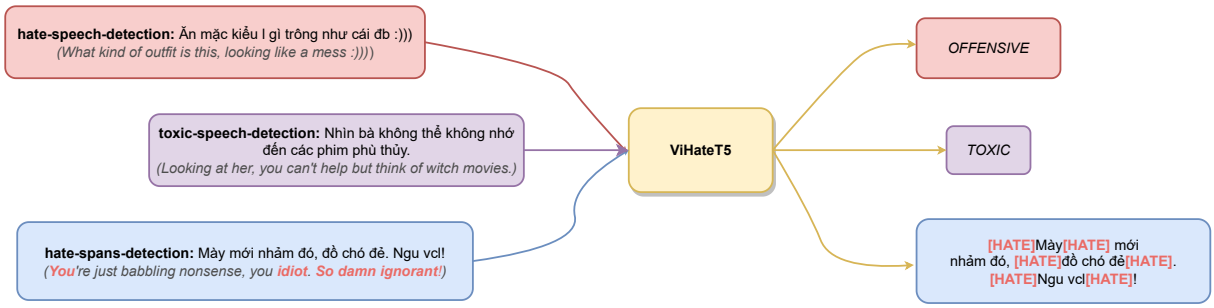

Figure 1: An overview of the proposed ViHateT5 model incorporating various prefix tasks tailored for hate speech detection in Vietnamese.

separate fine-tuning for specific tasks, resulting in system fragmentation.

In this paper, we present a new approach to address the existing limitations of HSD systems in Vietnamese. The contributions of this research are outlined as follows:

- A novel domain-specific model named ViHateT5 is presented in this study to address HSD-related problems in the Vietnamese language. Unlike prior endeavors, which merely fine-tune general transformer models, our innovative model was explicitly trained on a specific dataset derived from social media texts called VOZ-HSD with 10M+ comments with labels.d
- A unified T5-based model by fine-tuning the pre-trained ViHateT5 model advances the state-of-the-art performance on all available HSD benchmark datasets in Vietnamese.
- We illustrate our empirical strategy and data preparation to establish a comprehensive model for tackling HSD problems. Moreover, we highlight the significance of data pre-training on our pre-trained model, showing its substantial impact on model performance.

This paper is organized into distinct sections. Section 2 examines relevant prior research on hate speech detection tasks in Vietnamese. Subsequently, Section 3 introduces ViHateT5, our text-to-text model, covering its automatically generated pre-training dataset, pre-training methodologies, and fine-tuning for downstream tasks. Section 4 presents experimental results obtained by comparing various baseline methods with our proposed ViHateT5 model across a range of hate speech detection-related tasks and address discussions. Section 5 concludes the paper with a summary of our findings. Section 6 addresses the current limitations of our proposed method, while Section 7 provides ethical statements related to our research.

2 Related Work

Since the emergence of the transformer architecture (Devlin et al., 2019), numerous challenges in NLP have been successfully addressed, including tasks related to hate speech detection. Additionally, domain-specific models like HateBERT (Caselli et al., 2021), fBERT (Sarkar et al., 2021), or ToxicBERT[5] have been introduced. However, there remains a deficiency in hate-speech-focused pre-trained models for low-resource languages like Vietnamese, hindering the effective resolution of HSD tasks.

Besides, there exist diverse endeavors pertaining to HSD tasks, which involve the contribution of large-scale, high-quality datasets, thus facilitating precise research in this domain (Luu et al., 2021; Nguyen et al., 2021; Hoang et al., 2023). Furthermore, competitions such as the VLSP-2019 shared-task (Vu et al., 2020), dubbed Hate Speech Detection on Social Networks, are geared towards the development of automated systems aimed at identifying harmful content on internet-based social media, yielding remarkable outcomes. Additionally, transformer-based models have demonstrated remarkable efficacy across various NLP tasks. The advent of monolingual pre-trained models in Vietnamese, which have been observed to surpass their multilingual counterparts (Nguyen et al., 2022; To et al., 2021), has paved the way for the creation of precise systems for hate speech detection.

On the other hand, despite the remarkable performance achieved by current BERT-based approaches, the necessity persists to develop separate

[5] https://huggingface.co/unitary/toxic-bert

systems tailored to individual tasks. Addressing this issue of fragmentation, a unified text-to-text-based model such as FT5 for English or mFT5 (Ranasinghe and Zampieri, 2023) has demonstrated its effectiveness in amalgamating multiple tasks ranging from syllable-level to sentence-level-based HSD challenges. Hence, in this study, we introduce ViHateT5, a pre-trained text-to-text model specifically designed for the hate speech domain, aiming to streamline complex separate systems while ensuring optimal performance in addressing HSD issues in Vietnamese.

3 ViHateT5

This section reveals the methodologies for the creation of pre-training data, the pre-training techniques utilized, and the fine-tuning procedures undertaken to assemble the unified ViHateT5 model.

3.1 Automated Pre-training Data Creation

That Vietnamese is a low-resource language results in a shortage of extensive datasets for training targeted language models, particularly in specific NLP tasks. In this research, we present a significant Vietnamese hate speech classification dataset alongside an automated data annotation system. The entire process, which includes several modules, is illustrated in Figure 2.

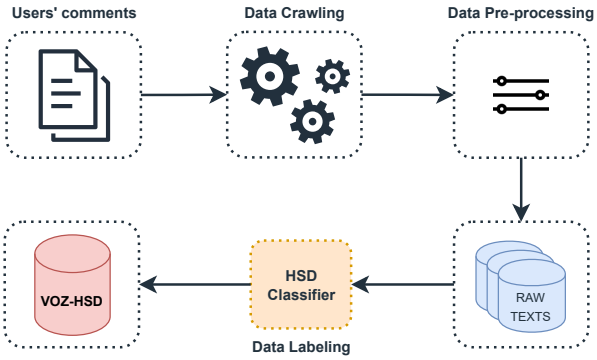

Figure 2: The process of creating VOZ-HSD by the automated data labeling approach.

Data Crawling: Initially, data was crawled from VOZ Forums[6], recognized as one of the most popular forums among young Vietnamese individuals. In comparison to other mostly used social media platforms like Facebook or TikTok, which have been utilized as pre-training data sources for other transfer learning models (Nguyen et al., 2023), data sourced from VOZ presents a potentially richer resource due to its characteristic of unrestricted freedom of speech. Consequently, it represents a valuable asset for research into hate speech.

The primary source of data collection was the main chat parent-thread[7], where users typically share their personal thoughts, often incorporating toxic content and emotional expressions. The crawling process involved the utilization of the BeautifulSoup4[8] tool.

Data Pre-processing: Given that the raw data comprises social media content, it includes noise and undesirable elements like user identities, URLs, or references to other comments. Therefore, the pre-processing of texts is exceedingly crucial before inputting them into models. In our research, we adopt the data pre-processing approach outlined by Nguyen et al. (2023), which involves tasks such as eliminating mentioned links, @username, retaining emojis and emoticons, and further excluding quotes, considered a distinctive element in a forum-based social media platform. The process results in approximately 1.7GB of uncompressed textual data.

AI Data Annotator: The advancements observed in AI data labeling systems (Desmond et al., 2021) have motivated our research to explore automated data annotation, with the objective of generating extensive datasets for hate speech classification. Since we experiment with pre-trained models using different data ratios that require raw texts to be labeled (as discussed in Section 4.6), we initially convert the ViHSD dataset (Luu et al., 2021), a recognized benchmark for hate speech detection in Vietnamese, into two labels: `CLEAN` $\Rightarrow$ `NONE`, and (`OFFENSIVE`, `HATE`) $\Rightarrow$ `HATE`, and employ it as a training dataset for training our classifier.

Following this, we fine-tune several pre-trained models designed for Vietnamese to identify the best-performing one. Results, shown in Table 6 in Appendix A, reveal that the ViSoBERT-based fine-tuned model achieves the highest Macro F1-score. Thus, we select this model as the HSD Classifier.

Automated Data Labeling: Utilizing the selected HSD Classifier, we proceed to label all textual data within the raw dataset automatically. The resultant dataset comprises approximately 10 million user comments annotated with hate speech labels. According to the statistics, there are over

[6]https://voz.vn/

[7]https://voz.vn/#khu-vui-choi-giai-tri.16

[8]https://pypi.org/project/beautifulsoup4/

Figure 3: The word cloud of VOZ-HSD dataset.

500K+ comments labeled as containing harmful content, constituting a significant portion of the total dataset. Notably, the dataset still maintains a substantial number of comments labeled as hate speech, particularly when compared to label distributions observed in the previous study by Luu et al. (2021) focusing on the Vietnamese hate speech detection task. We designate the final dataset as VOZ-HSD, indicating its purpose for hate speech detection and its origin from VOZ.

The word cloud depicted in Figure 3 showcases the prevalent terms found in harmful comments within VOZ-HSD. It is apparent that the dataset predominantly contains offensive language in Vietnamese, encompassing profanity, explicit content, colloquialisms, and informal expressions. As a result, this dataset holds promise for improving the ability of language models to identify and address hate speech effectively.

3.2 Model Pre-training

Inspired by the success of domain-specific pre-training and the efficacy of text-to-text models in addressing HSD tasks (Ranasinghe and Zampieri, 2023), we embark on the pre-training of the ViHateT5 model, leveraging the T5 architecture. The constructed VOZ-HSD dataset is employed as the pre-training dataset, comprising samples extracted from real-life comments.

3.3 Model Fine-tuning

To evaluate the efficacy of the trained model, we proceed to fine-tune the pre-trained ViHateT5 on various hate-speech-based datasets currently available, focusing on three tasks in Vietnamese.

Hate Speech Detection (ViHSD): Initially devised to identify harmful content in user comments across social media platforms in Vietnam, the Vietnamese Hate Speech Detection (ViHSD) dataset (Luu et al., 2021) has been extensively employed for text classification tasks. It involves categorizing texts into three labels: `HATE`, `OFFENSIVE`, and `CLEAN`. The ViHSD dataset comprises over 33K comments collected from comment sections of Facebook pages and YouTube videos.

Toxic Speech Detection (ViCTSD): The ViCTSD dataset (Nguyen et al., 2021) was initially formulated to identify constructiveness and toxicity in user comments. However, this study focuses solely on detecting toxicity within the comments. Originating from online news websites, where users are typically middle-aged individuals who often express themselves in formal styles, the level of offensiveness may not be as overt as in other datasets. Consequently, conducting toxicity detection on this dataset presents a challenge for language models.

Hate Spans Detection (ViHOS): The ViHOS dataset (Hoang et al., 2023) represents the first human-annotated corpus for identifying hateful and offensive spans within Vietnamese texts, providing a syllable-level task for HSD in Vietnamese. With over 11K comments and 26K annotated spans, this dataset diverges from its predecessors by focusing on the syllable level of hate speech, thereby augmenting the complexity of the task. Current BERT-based fine-tuning approaches typically adopt IOB sequence taggings to pre-process data, treating this task as a token classification task.

4 Experiments and Results

4.1 Data

This section outlines the experiments conducted, including details regarding the training data utilized, model configurations, baselines, evaluation metrics, and the ensuing results. We present the performance of our proposed ViHateT5 model across multiple HSD tasks in Vietnamese, providing insights into its efficacy relative to other previous state-of-the-art approaches. Note that no specific data pre-processing is applied to any models to ensure fairness.

Pre-training Data: We use raw texts from the VOZ-HSD dataset to pre-train the ViHateT5 mod-

Table 1: Statistics of datasets used in the experiments. Note that all samples in datasets are comments written in Vietnamese.

Dataset	Samples			Labels	Source(s)
Pre-training Data					
VOZ-HSD	10.8M			NONE, HATE	Voz Forum
Finetuning Data					
ViHSD (Binary)	24,048	2,672	6,680	NONE, HATE	Facebook, Youtube
ViHSD (Luu et al., 2021)	24,048	2,672	6,680	CLEAN, OFFENSIVE, HATE	Facebook, Youtube
ViCTSD (Nguyen et al., 2021)	7,000	2,000	1,000	NONE, TOXIC	VnExpress
ViHOS (Hoang et al., 2023)	8,844	1,106	1,106	Hate Speech Spans	Facebook, Youtube

els. As labeled by the HSD Classifier, raw texts in the VOZ-HSD dataset with generated labels can be helpful for data analysis and experiments with different proportions of hate-labeled samples.

Downstream Task Data: Next, we select several benchmark datasets for hate speech detection in Vietnamese to assess the performance of our proposed ViHateT5 model compared to others. These datasets encompass both sentence-level tasks, such as hate speech detection and toxic speech detection, and syllable-level tasks, such as hate spans detection, using the ViHSD, ViCTSD, and ViHOS datasets, respectively.

Specific Data Pre-processing for T5-based Models: Given the utilization of a text-to-text architecture, T5-based models require specific data pre-processing prior to fine-tuning downstream tasks. Figure 1 illustrates the multitasking input-output of our proposed ViHateT5 model, along with other T5-based models. Initially, we append task-specific prefixes, namely `'hate-speech-detection'`, `'toxic-speech-detection'`, and `'hate-spans-detection'` for texts sourced from the ViHSD, ViCTSD, and ViHOS datasets, respectively. For the syllable-level task of ViHOS, we incorporate tags `[HATE]` before and after the spans to encompass multiple spans based on the given index spans, thereby producing target texts for model training. Table 11 in Appendix D provides several samples of processed texts for both BERT-based and T5-based models employed in this study.

4.2 Model Setup

We follow the original pre-training strategy outlined for the T5 model (Raffel et al., 2023) to pre-train our ViHateT5. Both training and validation are conducted with a batch size of 128. Continual pre-training is executed over 20 epochs, employing the Adam optimizer with a lower learning rate set at 5e-3. Additionally, a weight decay of 0.001 is applied, with the initial 2,000 steps designated for warm-up during training.

In the fine-tuning phase, we maintain uniform settings for all BERT-based baseline models across specific tasks. Similarly, the same model settings are applied to T5-based models. For detailed information regarding the model settings for fine-tuning downstream tasks, please refer to Appendix B.2.

It is worth noting that all experiments are carried out with a limited resource setup utilizing a single NVIDIA A6000 GPU.

4.3 Baseline

We establish various baselines based on BERT-based architecture to compare with the performance of our proposed ViHateT5 model. The selected BERT-based pre-trained language models, encompassing both multilingual and monolingual variants, are readily available and extensively utilized for Vietnamese. Details of these pre-trained models, along with our proposed ViHateT5, are provided in Table 2. This information encompasses their architectures, total parameters, maximum sequence length of the model, pre-training data domain, vocabulary size, and data size.

4.4 Evaluation

The downstream tasks in this study are evaluated using metrics consistent with those employed in previous publications (Nguyen et al., 2022, 2023), which include accuracy score (Acc), weighted F1-score (WF1), and macro F1-score (MF1). For each task, MF1 serves as the primary evaluation metric, as per the original research. Besides, we calculate the Average MF1, derived from the MF1 scores across three benchmark datasets, to depict the overall performance of each model on HSD tasks.

Additionally, hate spans detection is a syllable-level task, necessitating the processing of output

Table 2: Details on baseline pre-trained models and our ViHateT5 used in the experiments, including model architecture, number of total parameters, max sequence length, pre-training data domain, vocab size, and the total of data size. Note that the data size for pre-training multilingual models reflects the total, not just Vietnamese texts.

Model	#archs	#params	#max_len	Data Domain	#vocab	Size
BERT (multilingual, cased) (Devlin et al., 2019)	base	177M	512	BookCorpus+EnWiki	120K	20GB
BERT (multilingual, uncased) (Devlin et al., 2019)	base	167M	512	BookCorpus+EnWiki	106K	20GB
DistilBERT (multilingual) (Sanh et al., 2019)	base	135M	512	BookCorpus+EnWiki	120K	20GB
XLM-RoBERTa (Conneau and Lample, 2019)	base	270M	512	CommonCrawl	250K	2.5TB
PhoBERT (Nguyen et al., 2020)	base	135M	256	ViWiki+ViNews	64K	20GB
PhoBERT_v2 (Nguyen et al., 2020)	base	135M	256	ViWiki+ViNews+OscarCorpus	64K	140GB
viBERT (Tran et al., 2020)	base	115M	256	Vietnamese News	38K	10GB
ViSoBERT (Nguyen et al., 2023)	base	98M	256	Vietnamese Social Media	15K	1GB
ViHateT5 (Ours)	base	223M	256	VOZ-HSD	32K	1.7GB

from T5-based models before computing evaluation metrics. To accomplish this, we follow Process 1 to obtain index spans consistent with the original dataset structure.

Process 1: Index spans retrieval from T5-based models' output

Data: `[HATE]`vcl`[HATE]` thật. Chịu luôn `[HATE]`đm m`[HATE]`!!!
(Original text $\mathbb{T}$: "vcl thật. Chịu luôn đm m!!!")
Result: [0, 1, 2, 20, 21, 22, 23]

1 Construct list $\mathbb{H}$ containing sub-strings covered by two `[HATE]` tokens;

2 From $\mathbb{H}$, find the corresponding index spans $\mathbb{I}$ of each sub-string in the original text $\mathbb{T}$;

3 **return** $\mathbb{I}$;

Next, we construct the binary form of indices by Process 2. Also, note that this second process is also applied to the ground truth data in order to compute the evaluation metrics consistently.

Process 2: Converting index spans for evaluation computation

Data: [0, 1, 2, 20, 21, 22, 23] (along with the original text $\mathbb{T}$)
Result: [1, 1, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0]

1 Calculate the length $\mathbb{L}$ of the original text $\mathbb{T}$;

2 Initialize the list $\mathbb{F}$ with '0' elements corresponding to the length $\mathbb{L}$;

3 Replace elements in list $\mathbb{F}$ whose indices are in list index spans $\mathbb{I}$ to '1';

4 **return** $\mathbb{F}$;

4.5 Experimental Results

Table 3 shows the performance of ViHateT5 compared to other approaches across various HSD tasks. Through experiments conducted under identical settings, ViHateT5 consistently outperforms other models, establishing itself as SOTA for most HSD-related tasks in Vietnamese.

In the realm of sentence-level tasks, specifically hate speech detection on the ViHSD dataset and toxic speech detection on the ViCTSD dataset, our proposed ViHateT5 model demonstrates outstanding performance, surpassing previous models with MF1 scores of 68.67% and 71.63%, respectively. Meanwhile, for the remaining baseline models, ViSoBERT achieves its highest performance on the hate speech detection task with an MF1 score of 67.71%, whereas XLM-RoBERTa attains the highest MF1 score of 71.53% for toxic speech detection.

In the domain of syllable-level tasks, such as hate spans detection, the ViHateT5 model showcases its effective ability to identify harmful segments by leveraging its text-to-text architecture, achieving the highest MF1 score of 86.37%. Additionally, ViSoBERT ranks second on the leaderboard with an MF1 score of 86.04%. Both models, being pre-trained specifically on social media domain data, yield consistent results on social media benchmark datasets, with a relatively small gap between them[9].

4.6 Discussion

In this section, we delve into the comparison between the unified HSD-fine-tuned ViHateT5 model and other T5-based models fine-tuned on HSD-

[9]The performance of ViSoBERT reported by Nguyen et al. (2023) on similar tasks may slightly differ from our experiments due to variations in model settings during reproduction owing to resource constraints.

Table 3: Comparative performance results of diverse models, encompassing fine-tuned models from multilingual pre-trained language models, monolingual models, and our proposed ViHateT5 model. Evaluation metrics include Accuracy (Acc), Weighted F1-score (WF1), and Macro F1-score (MF1) across various Hate Speech Detection (HSD)-related tasks.

Model	Average MF1	Hate Speech Detection			Toxic Speech Detection			Hate Spans Detection		
		Acc	WF1	MF1	Acc	WF1	MF1	Acc	WF1	MF1
BERT (multilingual, cased)	0.6930	0.8736	0.8680	0.6444	0.8983	0.8855	0.6710	0.8601	0.8464	0.7637
BERT (multilingual, uncased)	0.6827	0.8666	0.8606	0.6292	0.8993	0.8877	0.6796	0.8520	0.8172	0.7393
DistilBERT (multilingual)	0.6933	0.8630	0.8606	0.6334	0.8962	0.8873	0.6850	0.8585	0.8428	0.7615
XLM-RoBERTa	0.7265	0.8729	0.8697	0.6508	0.9015	0.9007	0.7153	0.8834	0.8754	0.8133
PhoBERT	0.6963	0.8675	0.8652	0.6476	0.9078	0.9027	0.7131	0.8465	0.8112	0.7281
PhoBERT_v2	0.7050	0.8742	0.8733	0.6660	0.9023	0.8978	0.7139	0.8492	0.8151	0.7351
viBERT	0.6780	0.8633	0.8579	0.6285	0.8881	0.8817	0.6765	0.8463	0.8128	0.7291
ViSoBERT	0.7507	0.8817	0.8786	0.6771	0.9035	0.9016	0.7145	0.9016	0.9007	0.8604
ViHateT5 (Ours)	**0.7556**	**0.8876**	**0.8914**	**0.6867**	**0.9080**	**0.9178**	**0.7163**	**0.9100**	**0.9020**	**0.8637**

Table 4: ViHateT5 versus other T5-based models in terms of Vietnamese HSD-related task performance with Macro F1-score.

	#archs	ViHSD	ViCTSD	ViHOS
mT5	base	0.6676	0.6993	0.8660
ViT5	base	0.6695	0.6482	0.8690
ViHateT5	base	**0.6867**	**0.7163**	**0.8637**

related tasks in Vietnamese. Additionally, we explore the performance of our proposed pre-trained ViHateT5 model across different pre-training data settings. Furthermore, we assess the model's ability to tackle syllable-level tasks.

The effectiveness of the T5 text-to-text architecture in addressing HSD tasks in Vietnamese has been demonstrated by ViHateT5. This study evaluates other T5-based models supporting Vietnamese for HSD tasks. We experiment with mT5-base, mT5-large (Xue et al., 2021) for multilingual models, and ViT5-base, ViT5-large (Phan et al., 2022) for monolingual models. The fine-tuning phases of these models are conducted under the same settings as ViHateT5, listed in Table 9 in Appendix B.2.2. Due to resource limitations, the batch size for large versions is reduced. Table 4 compares the performance of ViHateT5 with other T5-based models across three benchmark HSD datasets in Vietnamese.

The results attained highlight the superior performance of our proposed ViHateT5 model across various HSD-related tasks in comparison to other T5-based models supporting Vietnamese. The primary reason for this disparity lies in the nature of HSD benchmark datasets, which predominantly consist of spoken textual data, such as users' comments on the internet. These data exemplify social media characteristics, comprising informal written style texts accompanied by abbreviations, emojis, or teencode.

In contrast, while mT5 and ViT5 were pre-trained on formal content sources such as news or wiki pages, ViHateT5 was pre-trained on a domain-specific social media pre-training dataset. This domain-specific pre-training dataset ensures that ViHateT5 is more adept at understanding and processing informal language used in social media contexts, thereby yielding superior performance on HSD tasks.

The effectiveness of pre-training a transformer model on a domain-specific dataset was further validated by the ViSoBERT model, which demonstrated superior performance across various social media benchmark datasets (Nguyen et al., 2023).

How Pre-training Data Affects ViHateT5: We also assess how varying the data ratio in pre-training data affects our proposed models. This evaluation involves pre-training under different data conditions: utilizing full-data samples as in this study, employing a balanced-label pre-trained model with equal samples for both labels, and utilizing a hate-only pre-trained model where only hate labels are retained for pre-training.

Based on the generated labels, we conducted experiments to pre-train ViHateT5 under different data ratio conditions. The first condition used the entire dataset, while the second balanced the labels by reducing the number of `CLEAN` samples. The final condition exclusively pre-trained on `HATE` labeled samples. Table 5 presents the performance of these models after fine-tuning them on downstream tasks. It is worth noting that due to the relatively small size of the training samples in the 100% ratio condition, which is not sufficient for pre-training from scratch, we opted to use the continual pre-

Table 5: The performance of ViHateT5, measured by Macro F1-score, under various data pre-training conditions. The "Ratio" column indicates the percentage of hate data in the total dataset.

Ratio	Samples	Epochs	ViHSD	ViCTSD	ViHOS
100%	584,495	10	0.6548	0.6134	0.8542
		20	0.6577	0.6258	0.8601
50%	1,168,990	10	0.6600	0.6022	0.8577
		20	0.6620	0.6642	0.8588
5.54%	10,747,733	10	0.6286	**0.7358**	0.8591
		20	**0.6800**	0.7027	**0.8644**

training approach for all these experiments, utilizing weights from the ViT5-base[10].

The analysis reveals that pre-training with balanced or hate-labeled datasets does not improve model performance and can even lower MF1. However, different pre-training conditions affect ViHateT5 performance across various HSD tasks, suggesting additional pre-training on another T5 architecture model could be beneficial despite limited data. Also, increasing the number of pre-training epochs improves performance. Further research could explore resource-intensive setups to enhance ViHateT5 performance.

ViHateT5 in Syllable-level Hate Speech Detection: ViHateT5 has demonstrated its effectiveness in tackling syllable-level challenges, particularly in detecting hate speech spans within the ViHOS dataset. Leveraging an innovative architecture and training methodology derived from the T5 text-to-text transformer architecture, ViHateT5 surpasses baseline methods relying on BERT-based models, which primarily encounter limitations due to their token-level processing approach. Operating at the syllable level empowers ViHateT5 to pinpoint harmful spans within textual contexts accurately. Furthermore, its text-to-text framework presents ViHateT5 with opportunities to extend its capabilities to other tasks, such as hate speech detection question-answering or summarization, through adjustments to the prefix for fine-tuning.

5 Conclusions

Advancements in hate speech detection tasks in Vietnamese have recently gained notable progress thanks to the use of transformer models. However, these efforts remain fragmented due to the reliance on separate fine-tuned models for distinct tasks. Hence, our research aims to introduce a unified text-to-text transformer model, ViHateT5, with the potential to address prevailing issues in hate speech detection in Vietnamese and attain state-of-the-art performance. Moreover, ViHateT5's pre-training on domain-specific datasets enables it to grasp the nuances of social media content in Vietnamese deeply. The open-source nature of both the dataset and the model facilitates researchers and developers in leveraging our work, fostering further advancements in Vietnamese NLP and online safety.

6 Limitations

Training a language model through pre-training demands a substantial volume of data and computational power. In this investigation, we built an initial pre-training dataset, VOZ-HSD, suitable for experimentation with the ViT5-base, based on the base version of T5 architecture. However, it might not be adequate for larger versions. Previous research (Phan et al., 2022) outlines the effectiveness of these large-setting models, demonstrating their performance relative to smaller versions like the T5-base, which is employed in this experiment.

7 Ethical Statements

The proposed ViHateT5 model is specifically designed to handle various hate speech detection tasks in the Vietnamese language. Trained on a substantial auto-labeled dataset VOZ-HSD, as discussed in Section 3.1, the collected data undergoes meticulous preprocessing to eliminate all user identities, safeguarding user privacy.

With the rise of social media platforms and the corresponding increase in harmful content, there are unintended repercussions that require content moderation to protect users in online conversations. The proposed ViHateT5 model aims to make a meaningful contribution by delivering accurate performance across various hate speech detection tasks in Vietnamese. This initiative seeks to enhance content moderation on social media, promoting transparency and fostering a healthier online environment.

Acknowledgement

We express our gratitude to the anonymous ACL reviewers for their valuable and constructive feedback. Their contributions have significantly enriched the quality and thoroughness of our work.

[10] https://huggingface.co/VietAI/vit5-base

References

Sindhu Abro, Sarang Shaikh, Zahid Hussain Khand, Ali Zafar, Sajid Khan, and Ghulam Mujtaba. 2020. Automatic hate speech detection using machine learning: A comparative study. *International Journal of Advanced Computer Science and Applications*, 11(8).

Esma Aïmeur, Sabrine Amri, and Gilles Brassard. 2023. Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, 13(1):30.

Raza Ali, Umar Farooq, Umair Arshad, Waseem Shahzad, and Mirza Omer Beg. 2022. Hate speech detection on twitter using transfer learning. *Computer Speech & Language*, 74:101365.

Ayme Arango Monnar, Jorge Perez, Barbara Poblete, Magdalena Saldaña, and Valentina Proust. 2022. Resources for multilingual hate speech detection. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 122–130, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 759–760.

Irina Bigoulaeva, Viktor Hangya, and Alexander Fraser. 2021. Cross-lingual transfer learning for hate speech detection. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 15–25, Kyiv. Association for Computational Linguistics.

Alexander Brown. 2017. What is hate speech? part 1: The myth of hate. *Law and Philosophy*, 36:419–468.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Michael Desmond, Evelyn Duesterwald, Kristina Brimijoin, Michelle Brachman, and Qian Pan. 2021. Semi-automated data labeling. In *NeurIPS 2020 Competition and Demonstration Track*, pages 156–169. PMLR.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

B Ghafoori, Y Caspi, C Salgado, M Allwood, J Kreither, JL Tejada, T Hunt, L Waelde, O Slobodin, M Failey, et al. 2019. Global perspectives on the trauma of hate-based violence: An international society for traumatic stress studies briefing paper. *International Society for Traumatic Stress Studies*.

Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

Phu Gia Hoang, Canh Duc Luu, Khanh Quoc Tran, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2023. ViHOS: Hate speech spans detection for Vietnamese. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 652–669, Dubrovnik, Croatia. Association for Computational Linguistics.

Son T Luu, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2021. A large-scale dataset for hate speech detection on vietnamese social media texts. In *Advances and Trends in Artificial Intelligence. Artificial Intelligence Practices: 34th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2021, Kuala Lumpur, Malaysia, July 26–29, 2021, Proceedings, Part I 34*, pages 415–426. Springer.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152.

Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2020. A bert-based transfer learning approach for hate speech detection in online social media. In *Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019 8*, pages 928–940. Springer.

Dat Quoc Nguyen et al. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042.

Luan Nguyen, Kiet Nguyen, and Ngan Nguyen. 2022. SMTCE: A social media text classification evaluation benchmark and BERTology models for Vietnamese. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*, pages 282–291, Manila, Philippines. Association for Computational Linguistics.

Luan Thanh Nguyen, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2021. Constructive and toxic speech detection for open-domain social media comments in vietnamese. In *Advances and Trends in Artificial*

Intelligence. Artificial Intelligence Practices: 34th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2021, Kuala Lumpur, Malaysia, July 26–29, 2021, Proceedings, Part I 34, pages 572–583. Springer.

Nam Nguyen, Thang Phan, Duc-Vu Nguyen, and Kiet Nguyen. 2023. ViSoBERT: A pre-trained language model for Vietnamese social media text processing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5191–5207, Singapore. Association for Computational Linguistics.

Daniel Nkemelu, Harshil Shah, Michael Best, and Irfan Essa. 2022. Tackling hate speech in low-resource languages with context experts. In *Proceedings of the 2022 International Conference on Information and Communication Technologies and Development*, pages 1–11.

Long Phan, Hieu Tran, Hieu Nguyen, and Trieu H. Trinh. 2022. Vit5: Pretrained text-to-text transformer for vietnamese language generation.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.

Tharindu Ranasinghe and Marcos Zampieri. 2023. A text-to-text model for multilingual offensive language identification. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 375–384.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Diptanu Sarkar, Marcos Zampieri, Tharindu Ranasinghe, and Alexander Ororbia. 2021. fBERT: A neural transformer for identifying offensive content. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1792–1798, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Huy Quoc To, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen, and Anh Gia-Tuan Nguyen. 2021. Monolingual vs multilingual bertology for vietnamese extractive multi-document summarization. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 692–699.

Thi Oanh Tran, Phuong Le Hong, et al. 2020. Improving sequence tagging for vietnamese text using transformer-based neural models. In *Proceedings of the 34th Pacific Asia conference on language, information and computation*, pages 13–20.

Xuan-Son Vu, Thanh Vu, Mai-Vu Tran, Thanh Le-Cong, and Huyen Nguyen. 2020. Hsd shared task in vlsp campaign 2019: Hate speech detection for social good. *arXiv preprint arXiv:2007.06493*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer.

Steven Zimmerman, Udo Kruschwitz, and Chris Fox. 2018. Improving hate speech detection with deep learning ensembles. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

A Fine-tuning Hate Speech Classifier

To develop hate speech classifiers, we fine-tune existing pre-trained language models designed for the Vietnamese language. All experiments utilize a common set of pre-training language models. The training process is conducted over 3 epochs, employing a batch size of 16 for both training and evaluation phases. The maximum sequence length is defined as 128, and the learning rate is set to 1e-5. The remaining parameters adhere to recommendations from prior research. Table 6 provides an overview of the performance of various models in detecting hate speech in Vietnamese, utilizing two labels: `HATE` and `NONE`. Note that selecting the best model for our proposed system is based on its classification performance using the Macro F1-score (MF1). The chosen hate speech classifier, ViSoBERT-HSD, is publicly available at HuggingFace[11].

Table 6: Classification performances of various classifiers fine-tuned from different pre-trained models on the task of hate speech classification in order to find the best one for automated data annotation.

	Accuracy	Weighted F1-score	Macro F1-score
Multilingual Pre-trained Models			
BERT (Multilingual, base, cased) (Devlin et al., 2019)	0.8615	0.8483	0.7089
BERT (Multilingual, base, uncased) (Devlin et al., 2019)	0.8524	0.8335	0.6742
DistilBERT (Multilingual) (Sanh et al., 2019)	0.8344	0.7992	0.5895
XLM-RoBERTa (base) (Conneau and Lample, 2019)	0.8477	0.8070	0.5965
XLM-RoBERTa (large) (Conneau and Lample, 2019)	**0.8877**	0.8836	0.7861
Monolingual Pre-trained Models			
PhoBERT (base) (Nguyen et al., 2020)	0.8603	0.8479	0.7095
PhoBERT (large) (Nguyen et al., 2020)	0.8678	0.8464	0.6936
PhoBERT_v2 (base) (Nguyen et al., 2020)	0.8754	0.8723	0.7676
viBERT (Tran et al., 2020)	0.8612	0.8463	0.7028
ViSoBERT (Nguyen et al., 2023)	0.8477	**0.9033**	**0.8227**

B Experimental Settings

B.1 Model Pre-training

We initially pre-train ViHateT5 from scratch on the VOZ-HSD dataset and its variants with different pre-training data settings with parameters illustrated in Table 7. Note that validation split means the ratio for the validation set taken from the original dataset.

Table 7: Model settings for pre-training ViHateT5 variants. Note that all pre-trained models were trained on a low-resource setting with a single GPU NVIDIA A6000.

Name	Initial Weights	#archs	Pre-training Data	Valid Split	Epochs	l_r	batch_size	max_seq_len
ViHateT5	From scratch	base	VOZ-HSD	0.02	20	5e-3	128	256
ViHateT5	ViT5-base	base	VOZ-HSD	0.02	[10, 20]	5e-3	128	256
ViHateT5	ViT5-base	base	Balanced-label VOZ-HSD	0.05	[10, 20]	5e-3	128	256
ViHateT5	ViT5-base	base	Hate-label VOZ-HSD	0.1	[10, 20]	5e-3	128	256

B.2 Model Fine-tuning

B.2.1 BERT-based Models

To establish BERT-based models as baselines for fine-tuning each dataset, we implement the experimental configurations outlined below, as depicted in Table 8. These settings adhere closely to those recommended in the original publications.

[11]The link will be provided upon acceptance.

Table 8: Fine-tuning parameters for BERT-based models on each HSD-related task.

Dataset	batch_size	max_seq_len	l_r	weight_decay	epochs
ViHSD	16	256	2e-5	0.01	4
ViCTSD	16	256	2e-5	0.01	4
ViHOS	16	256	2e-5	0.01	10

B.2.2 T5-based Models

The model configurations for fine-tuning T5-based models, including our ViHateT5 utilized in this paper, are showcased in Table 9. The difference in the value of batch size occurs because of the limitation of GPU resources, leading to reduce the batch size for the training phase.

Table 9: Fine-tuning parameters for T5-based models on the tasks of hate speech detection in Vietnamese. Note that ViHateT5-based indicates fine-tuned models from any variants of the pre-trained ViHateT5.

	#archs	batch_size	max_seq_len	l_r	epochs
mT5	base	16	256	3e-4	4
ViT5	base	32	256	3e-4	4
ViHateT5-based	base	32	256	3e-4	4

C What is inside the VOZ-HSD dataset?

Table 10 illustrates the distribution of topic text data within the VOZ-HSD dataset. It is evident that we have gathered a wide range of conversation topics, indicating that the dataset is not skewed towards any particular domain and closely reflects real-life textual content. Previous studies by Nguyen et al. (2023) have further demonstrated that even with a limited dataset size of only 1GB in an uncompressed format for pre-training a transformer on social media texts, the model can still exhibit strong performance across multiple tasks, achieving state-of-the-art results.

Table 10: The distribution of comments in terms of conversation topics in the VOZ-HSD datasets.

No.	Parent Thread	N.o. Threads	N.o. Comments	Size (Uncompressed)
1	Random conversation	142,387	6,104,792	945MB
2	News	76,107	2,030,315	304MB
3	Sports	10,121	1,154,658	144MB
4	Cars	12,348	552,717	96MB
5	Movies - Music - Books	6,467	329,601	52MB
6	Bikes	5,093	258,728	41MB
7	Fashion	1,845	137,548	19MB
8	Food - Travel	3,492	136,649	19MB
9	Other Hobbies	690	42,737	6MB
	Total	**258,550**	**10,747,745**	**1,66GB**

D Actual examples in benchmarks dataset and their pre-processed representations for BERT-based baseline models and our proposed ViHateT5

Table 11: Examples from HSD benchmark datasets in experiments featuring diverse input and output data formats aligned with BERT-based and T5-based models.

Examples		BERT-based Models		T5-based Models	
Input	**Output**	**Source**	**Target**	**Source**	**Target**
ViHSD		Original text	CLEAN (0) OFFENSIVE (1) HATE (2)	Text with specific prefix	CLEAN OFFENSIVE HATE
Từ lý thuyết đến thực hành là cả 1 câu chuyện dài =)) (*Translated: From theory to practice is a whole long story* =)))	CLEAN	Từ lý thuyết đến thực hành là cả 1 câu chuyện dài =))	0	hate-speech-detection: Từ lý thuyết đến thực hành là cả 1 câu chuyện dài =))	CLEAN
Giống nhau như 2 giọt nước. Mà mỗi cái là 1 giọt nước mắt với 1 giọt nước sh!t thôi ạ (*Translated: Similar as two drops of water. But each one is a teardrop with a drop of shit too*)	OFFENSIVE	Giống nhau như 2 giọt nước. Mà mỗi cái là 1 giọt nước mắt với 1 giọt nước sh!t thôi ạ	1	hate-speech-detection: Giống nhau như 2 giọt nước. Mà mỗi cái là 1 giọt nước mắt với 1 giọt nước sh!t thôi ạ	OFFENSIVE
Im mẹ đi thằng mặt lon (*Translated: Shut up you big-faced idiot*)	HATE	Im mẹ đi thằng mặt lon	2	hate-speech-detection: Im mẹ đi thằng mặt lon	HATE
ViCTSD		Original text	NONE (0) TOXIC (1)	Text with specific prefix	NONE TOXIC
Một thời để nhớ, bao kỷ niệm gắn liền với những ca khúc của anh. (*Translated: A time to remember, with so many memories attached to his songs.*)	NONE	Một thời để nhớ, bao kỷ niệm gắn liền với những ca khúc của anh.	0	toxic-speech-detecion: Một thời để nhớ, bao kỷ niệm gắn liền với những ca khúc của anh.	NONE
nghe xong máu điên trong người nổi lên. muốn đánh cho thằng cha một trận quá..... (*Translated: After listening, rage surged through my veins. I feel like giving that guy a beating.....*)	TOXIC	nghe xong máu điên trong người nổi lên. muốn đánh cho thằng cha một trận quá.....	1	toxic-speech-detecion: nghe xong máu điên trong người nổi lên. muốn đánh cho thằng cha một trận quá.....	TOXIC
ViHOS		Original text	IOB Tags: O, B-T, I-T	Text with specific prefix	Text with [HATE] tokens
Hãnh diện về ng thầy có tâm nhất của năm. (*Translated: Proud of the most dedicated teacher of the year.*)	[]	Hãnh diện về ng thầy có tâm nhất của năm.	[]	hate-spans-detection: Hãnh diện về ng thầy có tâm nhất của năm.	Hãnh diện về ng thầy có tâm nhất của năm.
Chương trình ln gì vậy ? :D (*Translated: What is this pussy program ? :D*)	[13, 14]	Chương trình ln gì vậy ? :D :)))	O O B-T O O O	hate-spans-detection: Chương trình ln gì vậy ? :D :)))	Chương trình [HATE] ln [HATE] gì vậy ? :D :)))
t deo hieu no cuoi cl me gi nua (*Translated: I don't fucking understand what the fuck he is laughing at*)	[2, 3, 4 19, 20, 21, 22, 23]	t deo hieu no cuoi cl gi nua	O B-T O O O B-T I-T O O	hate-spans-detection: t deo hieu no cuoi cl me gi nua	t [HATE] deo [HATE] hieu no cuoi [HATE] cl me [HATE] gi nua