ViHOS: Hate Speech Spans Detection for Vietnamese

Phu Gia Hoang, Canh Duc Luu, Khanh Quoc Tran,
Kiet Van Nguyen, Ngan Luu-Thuy Nguyen
University of Information Technology, Ho Chi Minh City, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam
{19520215, 19521272}@gm.uit.edu.vn
{khanhtq, kietnv, ngannlt}@uit.edu.vn

Abstract

The rise in hateful and offensive language directed at other users is one of the adverse side effects of the increased use of social networking platforms. This could make it difficult for human moderators to review tagged comments filtered by classification systems. To help address this issue, we present the ViHOS (**Vi**etnamese **H**ate and **O**ffensive **S**pans) dataset, the first human-annotated corpus containing 26k spans on 11k comments. We also provide definitions of hateful and offensive spans in Vietnamese comments as well as detailed annotation guidelines. Besides, we conduct experiments with various state-of-the-art models. Specifically, XLM-R$_{Large}$ achieved the best F1-scores in Single span detection and All spans detection, while PhoBERT$_{Large}$ obtained the highest in Multiple spans detection. Finally, our error analysis demonstrates the difficulties in detecting specific types of spans in our data for future research. Our dataset is released on GitHub[1].

Disclaimer: This paper contains real comments that could be considered profane, offensive, or abusive.

1 Introduction

Social networking sites have been widely used all over the world. Here, users can easily share their thoughts, connect with others, or earn money by selling items, creating content, and so on. Since these sites are universally accepted, many extreme users misuse comment functions to abuse other individuals or parties with hate and offensive language. Consequently, it has been proved that these types of speech could harm other users' health (Mohan et al., 2017; Anjum et al., 2018). Sometimes these behaviors can be considered cyberbullying, cyber threats, or online harassment.

However, current studies are mainly about classifying comments as a whole with binary labels

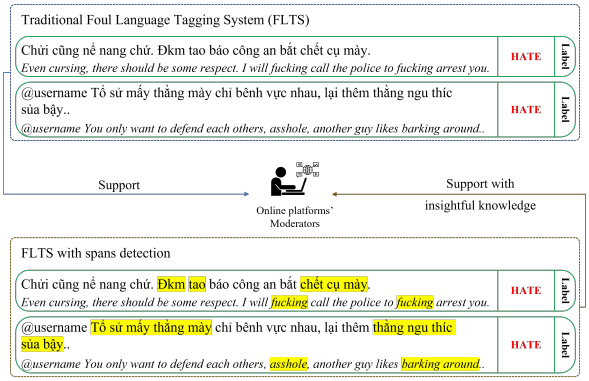


Figure 1: An example of the aid of spans detection for traditional foul language tagging system can provide additional insightful knowledge about tagged comments for human moderators.

(Zampieri et al., 2019; Nguyen et al., 2021b) or multiple labeling schemes of abusive behaviors (Davidson et al., 2017; Founta et al., 2018; Mathur et al., 2018). These efforts are made to aid human moderators, who need to review a massive number of online tagged comments that violate their community standards. However, a system that can highlight the spans that make a comment hateful or offensive can be more advantageous to human moderators who frequently deal with long and tedious comments and prefer explanations over a system-generated unexplained tag per comment. Furthermore, in some cases, using highlighted spans and moderators' context knowledge, they can take some actions to stop cyberbullying or online harassment. Nevertheless, there is only a study on toxic spans, SemEval-2021 Task 5: Toxic Spans Detection (Pavlopoulos et al., 2021). On the other hand, in a study of Mathew et al. (2020), hate and offensive spans worked as a rationale to support models in classifying the whole comments.

In Vietnamese, the resources about hate and offensive language are limited, namely ViHSD (Luu et al., 2021), HSD-VLSP (Vu et al., 2020), and UIT-ViCTSD (Nguyen et al., 2021b). Indeed, there

[1] https://github.com/phusroyal/ViHOS

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 652–669
May 2-6, 2023

is no study about hate and offensive spans in Vietnamese. This motivated us to (i) develop a new task of extracting hate and offensive spans from Vietnamese social media texts that conceivably impact research and downstream applications and (ii) provide the Natural Language Processing (NLP) research community with a new dataset for recognizing hate and offensive spans in Vietnamese social media texts.

Our two main contributions are summarized:

1. We created the first human-annotated dataset for Vietnamese Hate and Offensive Spans (ViHOS) comprising 26,467 human-annotated spans on 11,056 comments. Our dataset is annotated with a clear definition of hate and offensive spans, along with detailed and specialized guidelines for a less-studied language like Vietnamese. Compared to the toxic spans dataset at SemEval-2021 Task 5 (Pavlopoulos et al., 2021), which is built to detect toxic spans from toxic comments, or the HateXplain dataset (Mathew et al., 2020), which has spans working as a rationale for classifying the whole sentence, ours includes not only a large number of texts with annotated hate and offensive spans but also clean texts without any spans. This effort is made to serve a new task of detecting hate and offensive spans from Vietnamese online social media comments.

2. To evaluate the efficacy of our dataset, strong baselines are empirically investigated on ViHOS, including BiLSTM-CRF (Lample et al., 2016), XLM-R (Conneau et al., 2019), and PhoBERT (Nguyen and Nguyen, 2020). We conducted various experiment schemas, including comparing the full dataset having additional clean comments with the dataset that does not have; Single span detection, Multiple spans detection, and All spans detection. We obtain that: (i) Additional clean comments help the baselines have better performance than the dataset without them for 10$\pm$2% (ii) After fine-tuning the deep learning model and pre-trained language models, results show that the pre-trained language models outperform the deep learning models.

2 Related work

To the best of our knowledge, much of the research in the field of hate speech detection has been conducted in English due to the abundance of corpora and the robust pre-trained models. Many benchmark datasets for hate and offensive speech in other languages have also been published in recent years, including Arabic (Mubarak et al., 2020), Dutch (Tulkens et al., 2016), and French (Chiril et al., 2019). Novel models are introduced to improve the efficiency of hate and offensive speech detection. Initial approaches were based on typical machine learning and deep neural networks with word embeddings. Transformer-based models such as BERT (Devlin et al., 2018), BERTology (Rogers et al., 2020), and BERT-based transfer learning (Ruder et al., 2019) have recently been used to detect hate and offensiveness that achieved competitive results in major SemEval shared tasks such as SemEval-2020 Task 12 (Zampieri et al., 2020), and SemEval-2021 Task 5 (Pavlopoulos et al., 2021). However, research in Vietnamese is still limited in terms of the dataset and experimental methods. Only a few outstanding research exist, such as ViHSD (Luu et al., 2021), HSD-VLSP (Vu et al., 2020), and UIT-ViCTSD (Nguyen et al., 2021b).

For the topic of detecting foul spans, there are only a few case studies in English that are closely related, namely the SemEval-2021 Task 5: Toxic Spans Detection dataset (Pavlopoulos et al., 2021) and the HateXplain dataset (Mathew et al., 2020). The toxic spans, defined in the SemEval-2021 Task 5 dataset, are the sequences of words that make a text toxic. There are a total of 10,629 posts in this dataset, which stems from the Civil Comments dataset (Borkan et al., 2019). Another dataset with hate and offensive spans at the word level is HateXplain. The HateXplain contains 20,148 Gab and Twitter posts. Each post is manually classified into one of three labels: hateful, offensive, and normal.

In this study, we focus on Vietnamese to close the gap and **develop the first Vietnamese hate and offensive spans detecting benchmark**.

3 Dataset Creation

3.1 Dataset Source

ViHOS consists of 11,056 comments derived from the ViHSD dataset (Luu et al., 2021). The Vietnamese Hate Speech Detection dataset (ViHSD) is one of the few large and credible social media text datasets in a low-resource language like Vietnamese. ViHSD contains 27,624, 3,514, and 2,262 of CLEAN, HATE, and OFFENSIVE comments, respectively. Comments in ViHSD are public and collected from social media platforms.

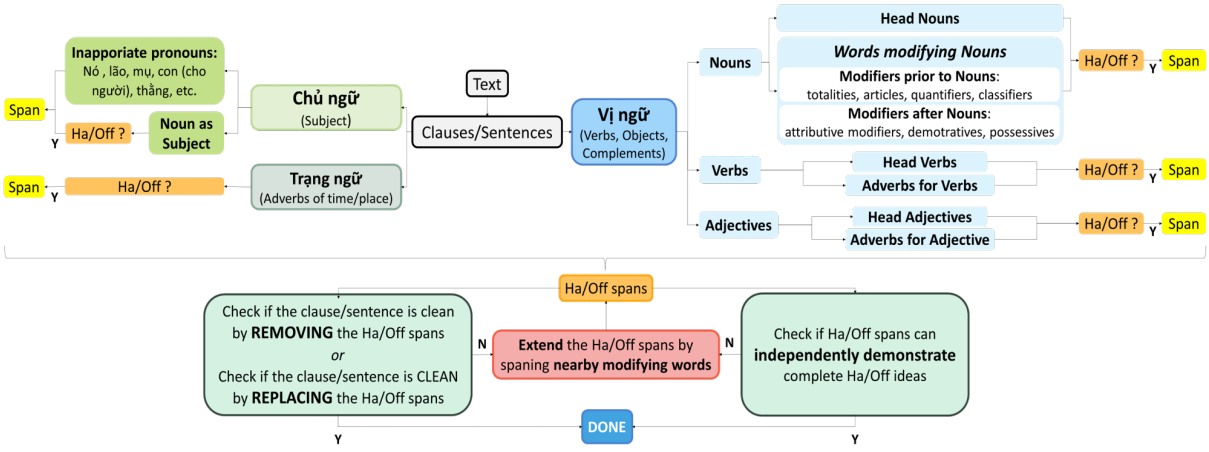


Figure 2: Detailed annotation guidelines for annotating comments for annotators. In which, *Ha/Off?* stands for a requirement for annotators to check whether the associated components are hate (Ha) or offensive (Off) or not.

Thus, metaphors, idioms, proverbs, and other tricky characteristics of online comments abound.

All of the HATE, OFFENSIVE comments from ViHSD after removing duplicates (5,528 comments left) are used to annotate the hate and offensive spans. Otherwise, 5,528 CLEAN comments, which also come from ViHSD and do not violate any hate or offensive definition defined in Section 3.2, are manually annotated for our dataset. We append the 5,528 CLEAN comments because: (*) We aim to detect the hate and offensive spans directly in online comments; (**) With an equal number of span and non-span (clean) comments helps models not be biased towards any type.

3.2 Annotation Guidelines

Our goal is to create a dataset that contains comprehensive hate and offensive thoughts, meanings, or opinions within the comments rather than just a lexicon of hate and offensive terms. We define the hate or offensive spans as follows to help annotators understand our goals:

- Harassing, cursing, insulting, disrespecting others.
- Sexual or verbal abuse towards one or a group of individuals based on their sensitive characteristics such as region, religion, politics, body, gender, etc.
- Insinuations, metaphors, metonymy used for hate, offensive or controversial purposes on sensitive issues such as region, gender, religion, politics, human rights, etc.
- Disuniting any factions or parties based on their politics, religion, ideologies, genders, etc.
- Causing verbal disrespect by using inappropriate pronouns.
- If replaced or removed, the sentence will no longer be hateful or offensive.

However, the hate or offensiveness in Vietnamese comments might cover one or even many components of a sentence. For instance: "thằng ad thở ra cái tư duy như trẻ lớp mầm" (**Eng:** *the admin speaks as his mind just like a kindergarten boy.)* This comment consists of three nouns/nouns phrases: "thằng ad"$_{\textit{the admin}}$ (it is offensive when calling a guy as a "thằng") as subject; "cái tư duy"$_{\textit{mind}}$ (the appearance of the word "cái" causes this noun phase become offensive) and "trẻ lớp mầm"$_{\textit{kindergarten boy}}$ as objects; one verb: "thở" (it usually means *breathe*, but in this context, we could consider it as *speak* but in a hate manner). As defined above, we must annotate a part of the subject, "thằng," and the whole phrase of the verb with its objects, "thở ra cái tư duy như trẻ lớp mầm" (**Eng:** *speaks as his mind just like a kindergarten boy*) in order to capture the whole hate or offensive ideas.

Therefore, we provided detailed guidelines (Figure 2) to assist annotators in determining when to annotate one or multiple components in a hateful and offensive sentence. As we observed, most of the comments in our dataset are colloquial. These comments are written freely and lack many grammatical rules. As a result, we frequently witness comments that lack subject(s), verb(s), conjunc-

	Phase 1				Phase 2	Phase 3
	1^{st}	2^{nd}	3^{rd}	4^{th}		
Kappa score	0.4161	0.4568	0.4936	0.6402	0.7239	0.7215
F1-score	0.7085	0.7186	0.7534	0.8219	0.8219	0.8585

Table 1: Inner-annotator agreement scores in three phases of annotation. In which, 1^{st}, 2^{nd}, 3^{rd}, 4^{th} are corresponding with four rounds of training annotators in Phase 1.

tions, punctuation marks, and so on. To deal with this, the comments were split into clauses/sentences (units). The smallest subdivided unit must have at least a cluster of "Chủ ngữ"-"Vị ngữ" ("Subject"-"Verb, objects, complements"), which is considered as a simple sentence or clause, or a cluster of "Chủ ngữ"-"Vị ngữ" nesting in another "Chủ ngữ"-"Vị ngữ" as a complex sentence or clause. From this, annotators can recognize components in clauses or sentences before annotating them.

Furthermore, in Appendix A and Appendix B, we also provided notices for annotators in Table 9 and ways to help them deal with nine online foul linguistic phenomena in Table 8 while annotating the data.

3.3 Dataset Construction Process

For dataset construction, we conducted three phases in which Phases 2 and 3 were inspired by Truong et al. (2021) and used metrics as bellow to calculate Inter-Annotator Agreement (IAA) among annotators. LightTag (Perry, 2021) is the tool we used for annotating data.

Assessment of Inter-Annotator Agreement Cohen's Kappa is widely used to measure inter-annotator agreement (IAA) in most tasks and is accepted as the standard measure (McHugh, 2012). However, numerous studies indicated that Kappa is not the most proper measure for the NLP sequence tagging task like NER (Hripcsak and Rothschild, 2005; Grouin et al., 2011). The reason is that the definite number of negative cases required to calculate the Kappa does not exist for named entity spans. Spans in our task are the sequences of characters rather than sequences of tokens since hate and offensive spans could be icon(s), word(s), or distinct character set(s) (see Table 8 for more details). Therefore, the pre-existing fixed number of characters to consider in the process of annotating is not existent.

A solution to deal with this is to calculate a character-level-based Kappa. Still, it has two associated problems: (1) annotators need to look at sequences of one or more characters instead of characters alone, causing the Kappa not to reflect the annotation task well; and (2) the "O"-labeled characters (the negative cases) outnumber the hate and offensive ones (the positive cases), provoking the Kappa to be computed on highly imbalanced data. For these reasons, the F1-score calculated without the negative cases is usually the measure for calculating IAA for the NLP tagging tasks like NER (Deleger et al., 2012). In this paper, IAA based on both F1-score (macro average) and character-level-based Kappa are calculated, while the former is the primary measure.

Phase 1: Pilot Annotation

Six undergraduate students were hired for our annotation tasks. The primary purpose of this pilot annotation phase was to familiarize our annotators with this task before entering the Main Annotation phase. We then developed an initial version of annotation guidelines with examples and distributed them to annotators. All annotators were required to carefully study the guidelines and give feedback before annotating the same 100 random samples from the 5,528 HATE, OFFENSIVE comments from ViHSD. This process was conducted four times with the F1-score and the Kappa for measuring IAA, which was calculated by averaging the results of pairwise comparisons across all annotators, shown in Table 1. All annotators were qualified as there was no F1-score of pairwise comparisons below 0.8.

Phase 2: Ground Truth Annotation

We randomly sampled a Ground Truth set of 600 comments from the 5,528 HATE, OFFENSIVE comments for this phase. Two guideline developers annotated the Ground Truth set separately using the well-developed guidelines from the former phase, resulting in an F1-score of 0.86 and Kappa (Cohen's Kappa) of 0.72. Afterward, we hosted a discussion to deal with annotation conflicts and update the annotation guidelines.

Phase 3: Main Annotation

We split the remaining HATE and OFFENSIVE

	Train	Dev	Test
Number of clean comments	4,552	569	575
Number of Ha/Off comments	4,422	553	553
Average clean comments length	8.69	8.50	9.04
Average Ha/Off comments length	16.81	17.68	16.13
Clean comments vocabulary size	4,234	1,423	1,400
Ha/Off comments vocabulary size	5,162	2,089	2,013
Number of multi-span comments (%)	2,322 (26.26)	308 (27.85)	296 (26.76)
Number of single-span comments (%)	1,970 (22.27)	229 (20.70)	235 (21.25)
Number of non-span comments (%)	4,552 (51.47)	569 (51.45)	575 (51.99)
Average number of spans	2.10	2.09	2.00

Table 2: ViHOS statistics. Vocabularies size and comments length are calculated at the syllable level.

comments (Luu et al., 2021) (4,928 comments left) into six non-overlapping and equal subsets. We also divided the 600-sample Ground Truth set from Phase 2 into six equal 100-sample smaller sets to insert into each subset. Each well-trained annotator from Phase 1 received a subset to annotate. Their annotation performance was assessed by calculating the F1 score and the Kappa score of the 100-sample Ground Truth sets in their subset. If any score is below 0.81 in terms of the F1 score, its corresponding annotator has to annotate again until it meets the requirement. This process was completed with an F1-score of 0.86 in the mean.

Furthermore, our annotators manually annotated CLEAN comments from ViHSD to spot any hate and offensive spans before being added to our dataset. This process collected 5,528 additional clean comments that met our requirements of having no hate and offensive spans.

3.4 Dataset Statistics

Before conducting dataset analysis and experiments, ViHOS has a total of 11,056 comments after the annotation process and is divided into three subsets: train, development, and test, with an 8:1:1 ratio. In detail, ViHOS has 5,360 comments with hate and offensive spans and 5,696 clean comments without in which 5,528 comments were additionally added and 168 comments have no hate and offensive spans after Phase 3 in the annotation process. Table 2 contains more information on the ViHOS statistics. It is apparent that the vocabulary of ViHOS is medium-sized, which is due to the small number of words in comments and comments in our dataset. In addition, more statistics about spans in ViHOS are shown in Table 3.

		Train	Dev	Test
Spans Quantity	0 span (%)	4,552 (51.47)	569 (51.45)	575 (51.99)
	1 span (%)	1,970 (22.27)	229 (20.71)	235 (21.25)
	2 - 3 spans (%)	1,527 (17.27)	207 (18.72)	202 (18.26)
	4 - 6 spans (%)	601 (6.80)	75 (6.78)	68 (6.15)
	7 - 10 spans (%)	164 (1.85)	18 (1.63)	21 (1.90)
	>10 spans (%)	30 (0.34)	8 (0.72)	5 (0.45)
Spans Length	1 syllable (%)	5,253 (52.03)	699 (52.48)	647 (52.77)
	2 - 3 syllables (%)	3,554 (35.20)	466 (34.98)	474 (38.66)
	4 - 6 syllables (%)	916 (9.07)	122 (9.16)	112 (9.14)
	7 - 10 syllables (%)	259 (2.57)	31 (2.33)	19 (1.55)
	>10 syllables (%)	114 (1.13)	14 (1.05)	14 (1.14)

Table 3: Spans quantity and length statistics.

4 Experiments and Results

4.1 Baseline Models

We treat the task of detecting hate and offensive spans as a task of sequence tagging. As a result, we make use of IOB format (Ramshaw and Marcus, 1995) to tag characters for model training, and testing. We conduct experiments on a set of solid baseline models, including BiLSTM-CRF and two pre-trained language models, XLM-R and PhoBERT, to assess the difficulty of our dataset.

BiLSTM-CRF: We use BiLSTM-CRF (Lample et al., 2016), a model that achieves high performance in the span detection tasks (Pavlopoulos et al., 2021; Nguyen et al., 2021a). We implemented this model with three main layers: (1) The word embedding layer using pre-trained PhoW2V (Nguyen et al., 2020), (2) The BiLSTM layer, and (3) the Conditional Random Field (CRF).

XLM-R: XLM-RoBERTa (Conneau et al., 2019) is a multilingual language model and a variant of RoBERTa, pre-trained on 2.5T of data across 100 languages containing 137GB of Vietnamese texts. On several cross-lingual benchmarks, XLM-R outperforms mBERT.

	BiLSTM-CRF + Pho2W$_{syllable}$	BiLSTM-CRF + Pho2W$_{word}$	XLM-R$_{Base}$	XLM-R$_{Large}$	PhoBERT$_{Base}$	PhoBERT$_{Large}$
Full Data	0.7453	0.7036	0.7467	**0.7770**	0.7569	0.7716
W/o additional clean comments	0.6241	0.6244	0.6479	0.6756	0.6738	**0.6867**

Table 4: Experimental results on Full Data versus Without additional clean comments.

	Model	Single span			Multiple spans			All spans		
		P	R	F1	P	R	F1	P	R	F1
Syllable	BiLSTM-CRF + Pho2W$_{syllable}$	0.4222	0.5009	0.4329	0.5134	0.5712	0.5068	0.7452	0.7769	0.7453
	XLM-R$_{Base}$	0.7604	0.7653	0.7203	0.7927	0.7574	0.7327	0.7766	0.7574	0.7467
	XLM-R$_{Large}$	**0.7577**	**0.7679**	**0.7214**	**0.7829**	**0.7569**	**0.7357**	**0.8071**	**0.7887**	**0.7770**
Word	BiLSTM-CRF + Pho2W$_{word}$	0.3196	0.4468	0.3594	0.3533	0.5001	0.4013	0.6823	0.7489	0.7036
	PhoBERT$_{Base}$	0.7392	0.7485	0.7016	0.7761	0.7329	0.7092	0.7870	0.7680	0.7569
	PhoBERT$_{Large}$	**0.7435**	**0.7567**	**0.7067**	**0.7878**	**0.7557**	**0.7321**	**0.8028**	**0.7835**	**0.7716**

Table 5: Experimental results on Single span, Multiple spans, and All spans subsets.

PhoBERT: PhoBERT (Nguyen and Nguyen, 2020) is a monolingual language model which is pre-trained on a 20GB Vietnamese dataset and has the same architecture and approach as RoBERTa. PhoBERT is proven as a state-of-the-art method in multiple Vietnamese-specific NLP tasks such as Part-Of-Speech Tagging, Dependency Parsing, and NER (Truong et al., 2021; Nguyen and Nguyen, 2020).

4.2 Experimental Settings

We empirically fine-tuned all pre-trained language models using *simpletransformers*[2]. For the tokenizer, each comment was tokenized using VnCoreNLP (Vu et al., 2018) in word-level and syllable-level for fine-tuning the PhoBERT and the XLM-R, respectively. In addition, we used Adam optimizer with a learning rate of 2e-5, a batch size of 8, and trained with 10 epochs.

We utilized a pre-trained word embedding - PhoW2V both syllable-level and word-level settings (Nguyen et al., 2020) with 100 dims to implement the BiLSTM-CRF model. The optimal hyper-parameters of BiLSTM-CRF are described in Table 6. All baseline models were trained on a system having 26GB RAM and an NVIDIA Tesla P100 GPU.

4.3 Evaluation Metrics

The macro-average F1-score (F1) is used to evaluate our models. For each pair of gold-predicted spans, we compute F1 and then calculate the arithmetic mean of F1 for each of these cases. It should be noted that the final F1-score, Accuracy, and Precision reported are an average of more than ten runs with various random seeds.

[2]https://simpletransformers.ai/ (ver.0.63.3)

Hyper-parameters	Values
Optimizer	Adam
Learning rate	0.001
Mini-batch size	64
LSTM hidden state size	60
Embedding size	100
Dropout	[0.1, 0.1]
Epochs	10

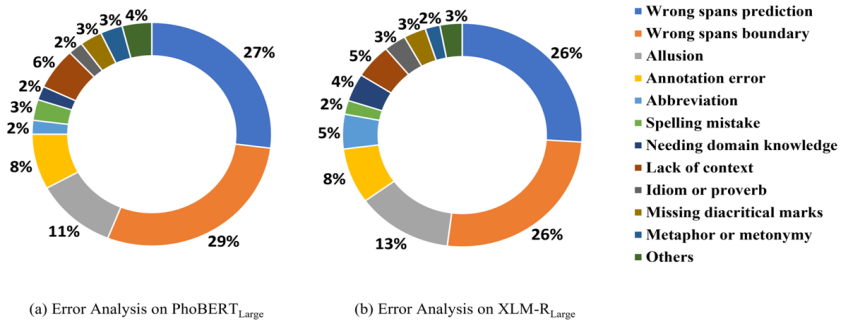
Table 6: Hyper-parameters of the BiLSTM-CRF.

4.4 Experiments and Results

Table 4 reports the baseline results before and after adding the 5,528 additional clean comments. We discover that after the addition, the performances improve 0.1002±0.0210. Specifically, PhoBERT$_{Large}$ considerably outperforms other models in the dataset without additional clean data, achieving 0.6867 in F1-score. In addition, the best model trained on Full data is XLM-R$_{Large}$, which has an F1-score of 0.7770. We find that XLM-R$_{Large}$ increased by 0.1014 and PhoBERT$_{Large}$ increased by 0.0849. These results demonstrate that the appearance of the additional clean comments successfully reduces model bias and improves performance.

Table 5 reports our results in three subsets corresponding to Single, Multiple, and All spans. Both Single and Multiple spans subsets are made by the process of splitting the All spans, which also known as the test set, based on the number of spans in each comment. Their results are described as follows:

Single span: We experimented with both



(a) Error Analysis on PhoBERT$_{\text{Large}}$ (b) Error Analysis on XLM-R$_{\text{Large}}$

Figure 3: Error analysis conducted on prediction on dev set made by PhoBERT$_{Large}$ and XLM-R$_{Large}$. We divide error cases into 12 categories including wrong spans detection, wrong spans boundary, allusion, annotation error, abbreviation, spelling mistake, needing domain knowledge, lack of context, idiom or proverb, missing diacritical marks, metaphor or metonymy, and others (rare characters, mixing other languages, all words stick together, etc.). These error cases are defined in Appendix C.

syllable-level and word-level language models. We discover that the pre-trained language models outperform the BiLSTM-CRF model by 0.3521±0.0099 in F1. This significant gap proves the fact that word embedding and features extraction of the pre-trained language models on the social media texts are superior to the BiLSTM-CRF. The XLM-R$_{Large}$ model achieves the best performance with a 0.7214 in F1-score. On the other hand, PhoBERT$_{Large}$ achieves a 0.7067 in F1-score. These results show no significant difference in performances among the multilingual and monolingual pre-trained models in the Single span.

Multiple spans: We experimented with the syllable-level and word-level and found that the pre-trained language models beat the BiLSTM-CRF model by 0.3212±0.0132. In addition, the performance of XLM-R$_{Large}$ is slightly better than the PhoBERT$_{Large}$ by 0.0036 in F1-score. The results on the Multiple spans are always better than the Single span, which might be explained by the fact that data in Multiple spans comprise more hate and offensive spans that can assist the models in learning more features of the data.

All spans: The results of the experiments on the All data are higher than the Single span and the Multiple spans. Specifically, in terms of F1-score, results of the XLM-R$_{Large}$ model are higher by 0.0556 and 0.0413 than the highest in the Single span and the Multiple spans, respectively while the figures for the PhoBERT$_{Large}$ are 0.0649 and 0.0395, respectively.

4.5 Results Analysis

We choose two best models: PhoBERT$_{Large}$ and XLM-R$_{Large}$ to conduct error analysis. As shown in Figure 3, we report the statistics of the ratio of various types of error cases[3] of 100 random samples in the dev set. We notice that *wrong spans prediction*[4], *wrong spans boundary*[5], *allusion*[6], *annotation error*[7], and *lack of context*[8] are major types of prediction failure of the PhoBERT$_{Large}$ and the XLM-R$_{Large}$.

We show some cases from the ViHOS development set in Table 7. In the first case of "Ns vậy lại xúc phạm cái đầu b**i *neutral face emoji*" (**Eng**: *If you said so, you insult the d**k head *neutral face emoji**), we notice that the PhoBERT$_{Large}$ could only predict "đầu"$_{head}$ as a hate and offensive span, whereas XLM-R$_{Large}$ predicts "b**i"$_{b**i}$. Both fail to predict the full boundary of "đầu b**i"$_{d**k\ head}$. The reason is that asterisks exist in the text ("b**i"). As defined in Subsection 4.2, the PhoBERT$_{Large}$, which was fine-tuned on the word-level data, cannot identify these tokens, but the XLM-R$_{Large}$, which was fine-tuned on the syllable-level data, can somewhat predict more accurate hate and offensive span even if it has asterisks.

Furthermore, in the second sampled comment, both best models failed to predict "đám cờ vàng" (**Eng**: *those yellow flag ones*) as a hate and offensive span, owing to the fact that this phase is an offensive metaphor for a rival party to Vietnamese

[3]Definition of errors are explained in the Appendix C.

[4]The model predicts clean spans as hateful and offensive.

[5]The model predicts inadequate boundary or fails to predict correctly.

[6]The comment refers to another person or subject indirectly and disrespectfully.

[7]The annotated span is wrong in terms of our guidelines. There is no later annotation modification in ViHOS.

[8]The comment has multiple meanings in different contexts, which mislead the prediction.

Ground truth spans	Models	
	PhoBERT$_{Large}$	XLM-R$_{Large}$
Ns vậy lại xúc phạm cái đầu b**i *neutral face emoji* (**Eng**: *If you said so, you insult the d**k head *neutral face emoji**)	["xúc", "đầu"]	["b**i"]
@username nhân dân VN tức là cờ đỏ. Còn đám cờ vàng là lũ súc vật lưu vong. Hiểu hông? (**Eng**: *@username Vietnamese people are red flag. The yellow flag ones are animals and exiled. Understand?*)	["đỏ.", "đám", "lũ", "súc"]	["đám", "lũ", "súc", "vật", "vong."]
Cướp đêm là giặc, cướp ngày là quan. 24/7 lúc nào cũng phải nơm nớp. Than ôi cái đất nước hạnh phúc (**Eng**: *Night thieves are enemies, day thieves are bureaucrats. 24/7, we are in a perpetual state of fear and anxiety. What a happy country*)	["Cướp", "cướp", "nớp."]	["Cướp", "giặc,", "cướp"]
Đúng rồi đéo thể tin được. Đáng lẽ phải cào bằng ra mới đúng. Thật không thể tin nổi. Phải cào bằng ra mới được (**Eng**: *Yes, fucking unbelievable. They should dig until it comes out. Unbelievable. Should dig till it comes out*)	["đéo", "cào", "cào"]	["đéo"]

Table 7: Case studies in the dev set from ViHOS that are complicated for the PhoBERT$_{Large}$ and XLM-R$_{Large}$. The highlighted spans in first column are Ground truth spans associated with their comments. **Eng** refers to the English meaning of associated comments.

people. In the third instance, the phrase "cướp đêm là giặc, cướp ngày là quan," (**Eng**: *night thieves are enemies, day thieves are bureaucrats*) which is an idiom that originated from folk poetry, also misleads the prediction. In the final example, the verb "cào"$_{dig}$ has no object and must be comprehended in context. These intriguing and challenging linguistic phenomena encourage more research into more robust models and methods in this field.

5 Conclusion and Future Work

We presented ViHOS, a new Vietnamese dataset for evaluating hate and offensive spans detection models. ViHOS includes 26,467 human-annotated spans on 11,056 comments. In addition, state-of-the-art models are conducted as the first baseline models, including BiLSTM-CRF and pre-trained language models such as XLM-R$_{Base}$, XLM-R$_{Large}$, PhoBERT$_{Base}$, and PhoBERT$_{Large}$. As a result, the XLM-R$_{Large}$ model achieves the best performance, with an F1-score of 0.7770. Furthermore, we discover that the performance when detecting multiple spans is better than the performance in detecting single spans in Vietnamese hate and offensive spans detection. Our dataset is available publicly at the GitHub link[9].

Despite the study's many promising contributions, the proposed research work still has several potential concerns, especially since the performance is still modest, and incorrect predictions could harm users' reputations if they rely heavily on our method. We intend to expand the dataset size and diversity of hate and offensive context for Vietnamese in the future to address this shortcoming. Furthermore, pre-and post-processing techniques will be used to standardize social networking texts (Clark and Araki, 2011) and deal with complex cases (as discussed in Subsection 4.5) to improve model performance(Suman and Jain, 2021; Kotyushev et al., 2021; Chhablani et al., 2021), particularly for Vietnamese pre-trained language models.

[9]https://github.com/phusroyal/ViHOS

Limitations, Social Impacts, and Ethical Considerations

Limitations and Social Impacts

There are numerous incomprehensible comments in our dataset due to the lack of context. Consequently, our annotators had to place themselves in imaginary contexts in order to annotate those comments (see Table 9 for more details about our solution). This shortcoming combined with the limitations of the neural networks in terms of understanding various linguistic phenomena (see Table 8 for more details about nine different linguistic phenomena) caused their performances of this task still insufficient to become practical.

We also acknowledge the risk associated with publicizing a dataset of hate and offensive spans (e.g. utilizing ours as a source for building abusive chatbots). However, we firmly believe that our proposed benchmark creates more value than risks.

Ethical Considerations

The undergraduate students in the annotation process are Vietnamese native speakers; have at least 12 years of studying Vietnamese with average scores on the Vietnam National Exam on Literature of 6.5; have at least three years of using social network platforms. They were explicitly warned that their tasks will display hateful and offensive content and if they became overwhelmed, they were also urged to stop labeling. These undergraduate students were paid $0.1 per comment, which takes an average of 6.44 seconds to complete (excluding the time used by workers who took exceptionally lengthy comments).

All the comments in ViHOS originated in the study of Luu et al., 2021, which preserved users' anonymity by removing all of them when creating the ViHSD. As a result, the comments in our dataset do NOT reflect our thoughts or viewpoints. ViHOS is available to the public under a usage agreement for research and related purposes only.

Acknowledgments

This work has been funded by The VNUHCM-University of Information Technology's Scientific Research Support Fund.

References

Amna Anjum, Xu Ming, Ahmed Faisal Siddiqi, and Samma Faiz Rasool. 2018. An empirical study analyzing job productivity in toxic workplace environments. *International journal of environmental research and public health*, 15(5):1035.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500.

Gunjan Chhablani, Abheesht Sharma, Harshit Pandey, Yash Bhartia, and Shan Suthaharan. 2021. Nlrg at semeval-2021 task 5: Toxic spans detection leveraging bert-based token classification and span prediction techniques. *arXiv preprint arXiv:2102.12254*.

Patricia Chiril, Farah Benamara, Véronique Moriceau, Marlène Coulomb-Gully, and Abhishek Kumar. 2019. Multilingual and multitarget hate speech detection in tweets. In *Conférence sur le Traitement Automatique des Langues Naturelles (TALN-PFIA 2019)*, pages 351–360. ATALA.

Eleanor Clark and Kenji Araki. 2011. Text normalization in social media: progress, problems and applications for a pre-processing system of casual english. *Procedia-Social and Behavioral Sciences*, 27:2–11.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

Louise Deleger, Qi Li, Todd Lingren, Megan Kaiser, Katalin Molnar, et al. 2012. Building gold standard corpora for medical natural language processing tasks. In *AMIA Annual Symposium Proceedings*, volume 2012, page 144. American Medical Informatics Association.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*.

Cyril Grouin, Sophie Rosset, Pierre Zweigenbaum, Karën Fort, Olivier Galibert, and Ludovic Quintard. 2011. Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview. In *Proceedings of the 5th linguistic annotation workshop*, pages 92–100.

Hoang Phe. 1983. *Vietnamese Dictionary*. Hong Duc, Vietnam.

George Hripcsak and Adam S Rothschild. 2005. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American medical informatics association*, 12(3):296–298.

Mikhail Kotyushev, Anna Glazkova, and Dmitry Morozov. 2021. Mipt-nsu-utmn at semeval-2021 task 5: Ensembling learning with pre-trained language models for toxic spans detection. *arXiv preprint arXiv:2104.04739*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *CoRR*, abs/1603.01360.

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. *arXiv preprint arXiv:2103.11528*.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2020. Hatexplain: A benchmark dataset for explainable hate speech detection. *arXiv preprint arXiv:2012.10289*.

Puneet Mathur, Rajiv Shah, Ramit Sawhney, and Debanjan Mahata. 2018. Detecting offensive tweets in Hindi-English code-switched language. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 18–26, Melbourne, Australia. Association for Computational Linguistics.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Shruthi Mohan, Apala Guha, Michael Harris, Fred Popowich, Ashley Schuster, and Chris Priebe. 2017. The impact of toxic language on the health of reddit communities. In *Canadian Conference on Artificial Intelligence*, pages 51–56. Springer.

Hamdy Mubarak, Ammar Rashed, Kareem Darwish, Younes Samih, and Ahmed Abdelali. 2020. Arabic offensive language on twitter: Analysis and experiments. *arXiv preprint arXiv:2004.02192*.

Anh Tuan Nguyen, Mai Hoang Dao, and Dat Quoc Nguyen. 2020. A Pilot Study of Text-to-SQL Semantic Parsing for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4079–4085.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. Phobert: Pre-trained language models for vietnamese. *arXiv preprint arXiv:2003.00744*.

Kim Thi-Thanh Nguyen, Sieu Khai Huynh, Luong Luc Phan, Phuc Huynh Pham, Duc-Vu Nguyen, and Kiet Van Nguyen. 2021a. Span detection for aspect-based sentiment analysis in vietnamese. *arXiv preprint arXiv:2110.07833*.

Luan Thanh Nguyen, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2021b. Constructive and toxic speech detection for open-domain social media comments in vietnamese. *arXiv preprint arXiv:2103.10069*.

Nguyen Hoai Nguyen. 2010. *Phonological characteristics of Nghe-Tinh dialect with the study of Vietnamese language history.* Vinh University, Vietnam.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. Semeval-2021 task 5: Toxic spans detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 59–69.

Tal Perry. 2021. LightTag: Text annotation platform. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 20–27, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Sebastian Ruder, Matthew E Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18.

Thakur Ashutosh Suman and Abhinav Jain. 2021. Astartwice at semeval-2021 task 5: Toxic span detection using roberta-crf, domain specific pre-training and self-training. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 875–880.

Thinh Hung Truong, Mai Hoang Dao, and Dat Quoc Nguyen. 2021. COVID-19 named entity recognition for Vietnamese. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2146–2153, Online. Association for Computational Linguistics.

Stéphan Tulkens, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven, and Walter Daelemans. 2016. A dictionary-based approach to racism detection in dutch social media. *arXiv preprint arXiv:1608.08738*.

Thanh Vu, Dat Quoc Nguyen, Dai Quoc Nguyen, Mark Dras, and Mark Johnson. 2018. Vncorenlp: A vietnamese natural language processing toolkit. *arXiv preprint arXiv:1801.01331*.

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*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. 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 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). *arXiv preprint arXiv:2006.07235*.

A Abusive Language Characteristics

Table 8: Characteristics of abusive language in ViHOS with examples, explanations, and solutions for annotators.

Abusive language characteristics	**Examples**	**Explanations and Solutions**
Non-diacritical marks comments	(1) Dit me cai quy trinh,vao cap cuu deo co tien,deo bao hiem thi nam do di. (**Eng:** *Fuck the procedure, without money or insurance, you could just lay there and no one cares about your emergency* (2) Dung la con linh dien dien vua thoi chang qua nt goi de choc my dien thoi "con ng" nhu linh dien thi ai them (**Eng:** *That must be the Crazy Linh, so crazy! I just call to tease the Crazy My. No one gonna love Crazy/The person like Crazy Linh!*	**Explanation:** some of these non-diacritical marks comments might trick annotators a little bit. (1): this is a non-diacritical marks comment but still able to understand. (2): there are some problems causing annotators to re-read multiple times as no punctuation, diacritic, and the text "con ng" could be considered as "crazy girl" or "the type (of human)" and both of these meanings is inappropriate. **Solution:** Non-diacritical marks comments are annotated as others. Annotators have to re-read until they fully understand the texts if needed. Those examples are annotated as follows: (1): ["Dit me", "deo", "deo"] (**Eng:** ["*Fuck*", "*fuck*", "*fuck*"]) (2): ["con", "dien", "dien", "con ng", "dien", "ai them"] (**Eng:** ["*con*", "*crazy*", "*crazy*", "*crazy/the person like*", "*crazy*", "*No one gonna love*"] in which the word "con" is an inappropriate way to call a woman.)

Table 8 continued from the previous page.

Metaphors, metonymies	(1) cái miệng rộng quá đẻ con ra còn lọt (**Eng:** *The mouth is too big that a baby could even be born through it*) (2) Dm! qua đợt dịch này thì thằng sống ích kỷ này chắc sớm gia nhập juventuts (**Eng:** *Fuck it! After the pandemic, this selfish boy will soon join Juventuts*)	**Explanation:** in our dataset, many comments use metaphors or metonymy to convey their hate or offensiveness in another way. (1): this is a metaphor of a mouth as a vagina. (2): this is a metonymy of Juventus' jersey as prison shirts (a common metonymy in Vietnamese). **Solution:** Annotators are required to annotate the whole ideas of metaphors, metonymy. (1): ["cái miệng rộng quá đẻ con ra còn lọt"] (2): ["Dm", "thằng sống ích kỷ này chắc sớm gia nhập juventuts"] (**Eng:** ["*Fuck*", "*this selfish boy will soon join Juventuts*"])
Puns	(1) Ad đăng bài này cũng là Bồn Kỳ Lắc nè nè nè :) (**Eng:** *The admin, who posts this, is also a "Bồn Kỳ Lắc")*	**Explanation:** Some comments use phrases that only read them backwards, they make sense. As in the example, "Bồn Kỳ Lắc", if this phrase is read backwards, it is "Bắc Kỳ Lồn" (pussy north). **Solution:** Annotators are required to annotate these puns too. (1): ["Bồn Kỳ Lắc"]
Using non-words characters to form hieroglyphs	(1) có tin t lấy *knife symbol* xiên chết cụ m ko :))) (**Eng:** *Do you belive that I could get a *knife symbol* to fucking kill you?*) (2) Đâm vào () (**Eng:** *stab in the ()*	**Explanation:** (1): the *knife symbol* is used instead of the word knife. (2): the existence of "()" can be considered as pussy in this context. **Solution:** Annotating the non-words only if they can convey a full meaning of hate or offensiveness, and the whole phase if they can not. (1): ["t", "*knife symbol*", "xiên chết cụ m""] (**Eng:** ["*t*", "**knife symbol**", "*fucking kill you*"]) in which "t", "m" are inappropriate pronouns.) (2): ["Đâm vào ()"]

Table 8 continued from the previous page.

Spelling mistakes	(1) Tôi đeo hiểu bạn noi cai gì luôn a :))) (**Eng:** *I have no fucking idea about what you saying*) (2) Mà dx cái chữi ngta nghe hài vcl bù lại củng đở (**Eng:** *Those fucking curses is so funny that can even refill that*) (3)Khi bạn xai trính tã nhưng cá ghén vít đún trính tã :)) (**Eng:** *When you make spelling mistakes but trying to fix it :)))*)	**Explanation:** (1): the words "đeo," "noi," "cai," "a" are spelling mistakes. The words can be understood as "đéo," "nói," "á". As that, these words are the same in accidentally missing acute marks. (2): the words: "chữi," "củng," "đở" are spelling mistakes is a phenomenon of mistaking tilde mark for hook above mark, and this often happens in some parts of Vietnam (Nguyen Hoai Nguyen, 2010). There are also familiar phenomena of mistaking marks such as tilde mark for underdot mark, acute mark for hook above mark, and so on (Nguyen Hoai Nguyen, 2010). (3): the phases: "xai trính tã," "cá ghén," "vít đúng trính tã" are spelling mistakes but on purpose. This comment utilizes spelling mistakes to attack opponents who also have spelling mistakes. Furthermore, these spelling mistakes also abuse opponents based on regional distinctions in accent, which cause some phenomenon of mistaking diacritical marks as in Example (2). **Solution:** The same as dealing with non-diacritical marks comments, annotators work as usual. (1): ["deo"] (**Eng:** ["*fuck*"]) (2): ["chữi", "vcl"] (**Eng:** ["*curse*", "*fuck*"]) (3): ["bạn xai trính tã nhưng cá ghén vít đún trính tã :))"]

Table 8 continued from the previous page.

Allusive language	(1) ra gì thì toi rồi, nó khác gì cách đảng CS chọn người, hồng hơn chuyên. Suy nghĩ kĩ đi. Giải độc cộng sản đã khó, giải độc tư tưởng cánh tả còn khó hơn. (**Eng:** *If it had something, it was done! It is just like the way CS (stands for Communism) chooses people, beauty over the profession. Detoxifying Communism is hard; detoxifying the left-wing political ideologies is even harder.*) (2) Rút kinh nghiệm lại được xài nghìn tỷ (**Eng:** *Just say learned and then can use trillion VND*)	**Solution:** Annotators are required to annotate the whole profound abuse. (1): ["toi", "hồng hơn chuyên", "Giải độc cộng sản đã khó, giải độc tư tưởng cánh tả còn khó hơn"] (**Eng:** ["*die*", "*beauty over the profession*", "*Detoxifying Communism is hard; detoxify the left-wing political ideologies even harder*"]) (2): ["Rút kinh nghiệm lại được xài nghìn tỷ"]
Homonym	(1) coin card ("coin card" is a homonym of "con cặc", which means *dick*)	**Solution:** annotate the whole abusive homonym. (1): ["coin card"]
Mixing languages	(1) Vl fake **Eng:** Fuck! Fake) (2) phe *X* compat tổng *Y* =)) (**Eng:** *X's side combats Y's side*) (3) Tối Thầy stream đá phò đi thầy ơi (**Eng:** *You should stream fucking some whores tonight, please!*)	**Solution:** treat foreign words as others and annotate if they meet the definition of hate and offensive spans. (1): ["Vl", "fake"] (**Eng:** ["*fuck*", "*fake*"]) (2): ["compat"] (**Eng:** ["*combat*"]) (3): ["đá phò"] (**Eng:** ["*fuck some whores*"])

Table 8 continued from the previous page.

Trick hate speech detecting systems on purpose	(1) Đấy Ông già xạo l*n của các bạn được tắm bồn Thái Lan đấy=))) (**Eng:** *See, your fucking old liar is Thai bathing again =))))*) (2) thằng c h.ó ngu (**Eng:** *Fucking stupid guy*) (3) vualonemlabaonhieu cm (**Eng:** *How long in cm to fit your pussy?*)	**Explanation**: some hate or offensive comments use punctuation to censor their inappropriate words (as in (1), asterisk is used in the word l*n (pussy)) or to disunite characters in words (as in (2), dot and space are used to disunite the words chó (dog) into c h.ó). These efforts actually can trick many hate speech detecting systems, or to put all words together (as in (3), vualonemlabaonhieu cm should be "vừa lồn em là bao nhiêu cm"). **Solution:** we annotate all characters which can be form into hate or offensive phase. (1): ["xạo l*n"] (**Eng:** ["*fucking lie*"]) (2): ["thằng c h.ó", "ngu"] (3): ["lon"] (**Eng:** ["*pussy*"])

B Notices for annotators

Table 9: Additional notices for annotators.

Notices	**Example**	**Explanation**
Try to figure out and consider as in the original context of the comments.		This could help annotators understand complex and non-context hate, offensive comments.
Do not let emotion affect the annotating process.		Annotators exposed in a long time to toxic comments are reported to get used to the frequently appearing hate, offensiveness.
Check the provided Vietnamese Dictionary if there is any uncertainty in being sure a word is hate or offensive.		We use the most reputable Vietnamese dictionary (Hoang Phe, 1983) to provide to annotators.
Should span the object is compared to in an inappropriate comparison.	"Ăn cơm nhìn như chó" (**Eng:** *You eat like a dog*) **Spans:** ["nhìn như chó"]	We consider this is an inappropriate comparison in which annotators must span "chó"$_{dog}$.

Table 9 continued from the previous page.

However, we should span the whole comparison if spanning only the object is compared to might not convey the complete hate or offensive idea	"ý thức như trẻ lớp mầm" **Spans:** ["ý thức như trẻ lớp mầm"] (**Eng:** *Your awareness is just like a kindergarten kid*)	If we only span "trẻ lớp mầm" (kindergarten kid), it will not convey the complete offensive idea. As that, annotators are encouraged to span the whole text instead.
Do not span conjunctions; exceptional Vietnamese cases; standard ways to call LGBTQ+.	(1) Vì thế, nên, nhưng, mà, etc. (**Eng:** *so, so, but, but, etc.*) (2) Gay, les, etc (3) Bóng, bê đê, ái nam ái nữ, etc.	(1) Some conjunctions in Vietnamese. (2) Appropriate ways to specify LGBTQ+. (3) Inappropriate ways to specify LGBTQ+ that need to be spanned in comments.
Blatant hate and offensive words prioritize being spanned over the others, especially in sentences without diacritics and could not be understanded.	May Cha H O an Tro. Cap. Ranh c. Di ngoi le. Nhieu chuyen. Do Cai Thu , do tam than **Spans:** ["Cap", "ngoi le", "Nhieu chuyen", "do tam than"] (**Eng:** ["*steal*", "*gossip*", "*talkative*", "*psycho*"])	Similar to this non-diacritical and incomprehensible comment, words as highlighted ones are more straightforward to be spanned.
Span the whole phase violating human rights.	"Về nước anh nên vào tù ở trước thay vì đi cách ly." (**Eng:** *When you return to Vietnam, you should go to jail first instead of going to isolation*) **Spans:** ["Về nước anh nên vào tù ở trước thay vì đi cách ly"]	Comments violate human rights, usually complex to specify hate, offensive words to span. As in this example, a Vietnamese citizen comes back from a foreign country has a right to have isolated healthcare firstly.
Span the whole obscene acronyms.	(1) clgv (**Eng:** *wtf is that?*) (2) clmn (**Eng:** *your mom's pussy*) (3) cmn (**Eng:** *your mother*).	
Follow the rule with words that do not have diacritics and conjoin to span out the hate, offensive spans.	cailongithe (**Eng:** *wtf is that?*)	Some comments have strings being constructed by many words missing diacritical marks, but still able to understand. The annotators should only span the hate and offensive characters set out of the string as in the example.
Span hate, offensiveness separately.	"Cả đám dlv là nhóm ngu dốt, trẻ trâu potay vcl luôn." (**Eng:** *The whole dlv crew are stupid, bull-headed kids (so I have) no fucking thing to say.*) **Spans:** ["đám", "ngu dốt", "trẻ trâu"] (**Eng:** ["*crew*", "*stupid*", "*bull-headed kids*"])	This comment must be spanned as in the example, but not as "đám," "ngu dốt, trẻ trâu," "vcl," "NGU."

C Definition of Error Cases for Error Analysis

We introduce 12 error definitions as follows:

1. **Wrong spans prediction**: The model predicts clean spans as hateful and offensive.
2. **Wrong spans boundary**: The model predicts inadequate boundary or fails to predict correctly.
3. **Allusion**: The comment refers to another person or subject in an indirect and disrespectful way.
4. **Annotation error**: Annotators have improperly annotated the span. The reason might be that they somehow do not follow the provided guidelines. However, there is no modification in the final dataset.
5. **Abbreviation**: The comment contain short forms of words.
6. **Spelling mistake**: The comment is spelling mistake.
7. **Needing domain knowledge**: Dialect and professional expertise are required to detect span in comments.
8. **Lack of context**: In different contexts, the comment could be understand in multiple meaning.
9. **Idiom or proverb**: The comment contains idiom or proverb.
10. **Missing diacritical marks**: Words in the comment do not have diacritical marks.
11. **Metaphor or metonymy**: The comment contains metaphor or metonymy.
12. **Others**: The comment contains rare characters, other languages, words in it are stick together, etc.