Abstract

# **1. Introduction**

Nowadays, social networks and online communication platforms have become essential spaces for interaction, information sharing, and community engagement. While these platforms offer numerous benefits, they also pose significant challenges in terms of content moderation and maintaining a healthy online environment.

One increasingly prevalent issue is the rise of toxic comments—remarks that may contain offensive language, hate speech, discriminatory content, or personal attacks. Such comments often reflect negative personal opinions and can contribute to hostility and polarization in online communities. The spread of toxic content not only undermines constructive discourse but also adversely affects the psychological well-being of individuals involved in these interactions.

To address this problem, the task of toxic comment detection has emerged as a critical area of research. The goal is to develop models capable of automatically identifying and classifying harmful comments, thereby assisting in content moderation efforts. Effective toxic comment detection plays a vital role in limiting the spread of harmful content and fostering safer, more respectful online communities.

# 2. Relate Work

## 2.1. ViHOS: Hate Speech Spans Detection for Vietnamese

"ViHOS: Hate Speech Spans Detection for Vietnamese" was conducted by Phu Gia Hoang, Canh Duc Luu, Khanh Quoc Tran, Kiet Van Nguyen, and Ngan Luu Thuy Nguyen from the University of Information Technology, VNU-HCM. This paper focuses on the task of hate and offensive comment classification using the ViHOS dataset.

The ViHOS dataset contains 11,056 Vietnamese comments, each annotated with one of two labels: Clean Comment or Hate/Offensive Comment. To tackle this classification task, the authors employed a combination of pre-trained language models (such as XLM-R and PhoBERT) along with deep learning architectures, notably BiLSTM-CRF, for improved sequence labeling and detection of hate speech spans.

The experimental results presented in the paper demonstrate the effectiveness of integrating contextualized embeddings with sequential models for detecting toxic spans in Vietnamese texts. This work lays a foundational benchmark for Vietnamese hate speech detection and highlights the importance of domain-specific datasets and language models.

# 3. Dataset Creation

## 3.1. Dataset Source

### 

- craw data on social media

## 3.2. Annotation Guidelines

In this paper, we classify comments into three toxicity levels based on their linguistic characteristics and intent: Neutral, Offensive, and Hate. Each label is defined as follows:

* Neutral: Comments do not contain any negative, aggressive, or insulting language. These comments are respectful, constructive, and often contribute meaningfully to the discussion. They may express agreement, disagreement, or inquiry in a civil manner without attacking others.
* Offensive: Comments include informal, inappropriate, or disrespectful language that targets an individual or idea, often with the intent to insult or provoke. However, they do not incite violence or target a specific group. These comments may involve profanity, name-calling, or sarcasm, and typically lack constructive value.
* Hate: Comments contain highly aggressive, vulgar, or hostile language that promotes hatred or discrimination against individuals or groups based on identity-related attributes such as race, gender, religion, or region. These comments often include threats, slurs, or dehumanizing language and may advocate violence or illegal behavior.

To ensure consistent labeling across all annotators, we developed a detailed annotation guideline. The purpose of this annotation guideline is to provide annotators with a clear and consistent framework for labeling dataset comments into the three predefined categories: *Hate comments*, *Offensive comments*, and *Neutral comments*. The Annotation Guideline was revised and refined after each round of test annotation. Figure 3.2.1 illustrates the complete Annotation Guideline in detail as a flowchart.

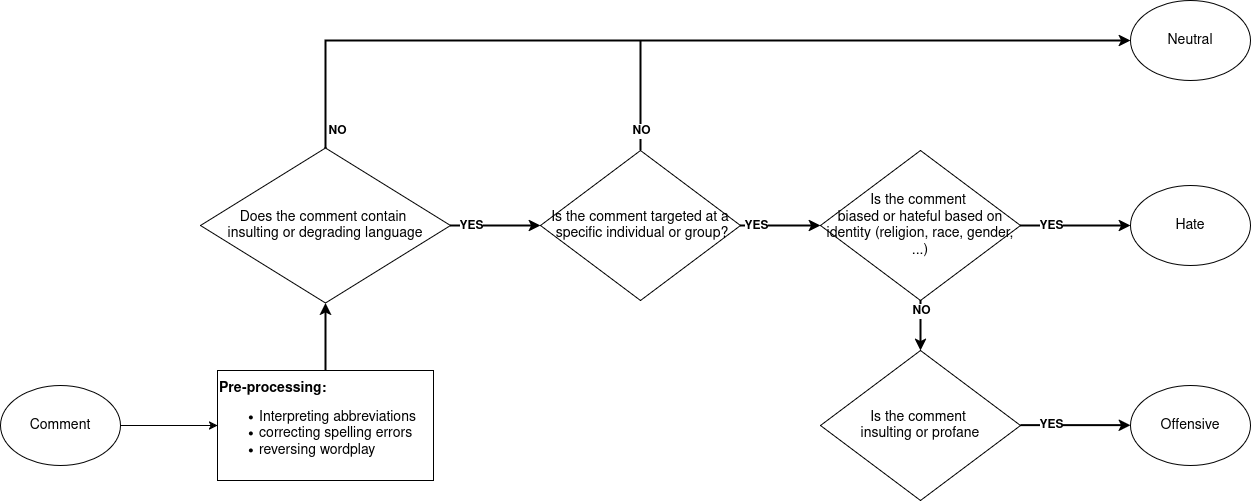
**

Figure 3.2.1. Flowchart of Annotation Guidelines

## 3.2. Dataset Construction Process

For the dataset construction, we adopted a three-phase annotation process inspired by the methodology used in ViHOS: Hate Speech Spans Detection for Vietnamese. These phases include Pilot Annotation, and Main Annotation

### 3.2.1. Pilot Annotation

In the pilot phase, we conducted a trial annotation on a sample of 100 comments. Team members independently labeled the data to assess inter-annotator agreement and to identify ambiguities in the definitions of toxic content. we revised the annotation guideline and based on the disagreements observed.

We use the Fleiss' Kappa coefficient to measure the level of agreement among the members.

### 3.2.2. Main Annotation

After finalizing the annotation guidelines, the full dataset was divided among the annotators for labeling. During this phase, annotators not only labeled the comments but also documented frequently occurring insulting and degrading expressions as well as common abbreviations used in Vietnamese online discourse. These lists were later used for preprocessing tasks such as normalization and filtering.

In addition, comments deemed irrelevant to the classification task were excluded from the dataset. These include:

* Spam or nonsensical comments (e.g., repeated characters or symbols)
* Comments consisting solely of emojis or punctuation marks.
* Tag-only comments (e.g., comments that contain only tagged names without meaningful content)

## 3.3. Dataset Statistics

# 4. Experiments and Results

## 

## 4.1. Classification Model

In this paper, we approach the task of toxic comment classification as a text classification problem with three labels: Neutral, Offensive, and Hate. We employ pre-trained language models to build our classification system, including PhoBERT, ELECTRA, BART, ViT5 and MT5.

### 4.1.1. PhoBERT

PhoBERT is a pre-trained language model based on the RoBERTa architecture, specifically trained on large-scale Vietnamese corpora. It has demonstrated state-of-the-art performance on various Vietnamese NLP tasks, making it an ideal choice for toxic comment classification.

In our implementation, we fine-tune the *vinai/phobert-base* model for a three-class classification task (neutral, offensive, and hate comments). We employ the corresponding PhoBERT tokenizer to preprocess the input texts by tokenizing, adding special tokens, and padding or truncating sentences to a fixed maximum length.

### 4.1.2 ELECTRA -Hoành

### 4.1.3. BART - Khoa

### 4.1.4. ViT5 - Cù ngọc hoàng

### 4.1.5. MT5 - Duy

## 4.2. Experiments and Results

## 4.3. Results Analysis

# 5. Conclusion and Future Work

*Conclusion*

Based on the results obtained from our toxic comment classification task, we propose several directions for future improvement:

First, we aim to refine the definitions and annotation guidelines for each comment category. Conducting multiple rounds of annotation and improving inter-annotator agreement will help enhance label consistency and dataset quality.

Second, we plan to expand the dataset by diversifying social media topics and collecting more toxic comments from posts that tend to provoke conflict or controversy. This will also help mitigate class imbalance, which remains a common issue in toxic language datasets.

Third, we intend to enhance the preprocessing pipeline by building a more comprehensive dictionary that includes abbreviations, slang, puns, and reversed or altered word forms commonly used in online communication. This will improve the model’s ability to understand informal or implicit toxic content.

Finally, we suggest increasing the training time for PhoBERT and exploring additional fine-tuning techniques to further boost classification performance.