An Explainable Hybrid Framework for Detecting Harmful Short Comments on Vietnamese Social Media Using TF-IDF and Lightweight Transformers

DOI: https://doi.org/10.54654/isj.v3i26.1162

Ha Van Muon*, Nguyen Thanh Hai, Nguyen Ngoc Nhan, Tran Quang Tuan

***Abstract*—The proliferation of harmful content on Vietnamese social media platforms necessitates effective detection systems for short, concise, and slang-heavy texts. This study proposes the Explainable Hybrid Framework (EHF) for enhanced detection and interpretability. On an expert-annotated dataset of 15,162 samples, EHF with SVM achieves an F1-Score of 92.5% and Accuracy of 92.6%, outperforming strong baselines such as PhoBERT on our dataset, while state-of-the-art models like ViHateT5 are discussed on external Vietnamese benchmarks. Notably, it reduces false positives by 30% and false negatives by 25% versus traditional methods, with 45ms inference time. Our primary contribution is an efficient, fast, and explainable solution that addresses the critical challenge of automated moderation for short Vietnamese texts.**

***Tóm tắt*— Sự gia tăng nhanh chóng của nội dung xấu độc trên các nền tảng mạng xã hội tiếng Việt đặt ra nhu cầu cấp thiết về các hệ thống phát hiện hiệu quả cho các văn bản ngắn, súc tích và nhiều tiếng lóng. Nghiên cứu này đề xuất Khung lai có khả năng giải thích (Explainable Hybrid Framework - EHF) nhằm nâng cao hiệu quả phát hiện và khả năng diễn giải. Trên bộ dữ liệu gồm 15.162 bình luận do chuyên gia gán nhãn, EHF kết hợp với SVM đạt F1-Score 92,5% và Accuracy 92,6%, vượt các mô hình đối sánh mạnh như PhoBERT trên bộ dữ liệu của nhóm tác giả, trong khi các mô hình tiên tiến như ViHateT5 được thảo luận trên các tập dữ liệu chuẩn tiếng Việt hiện có. Đáng chú ý, EHF giúp giảm 30% số trường hợp dự đoán sai dương (false positive) và 25% số trường hợp dự đoán sai âm (false negative) so với các phương pháp truyền thống, với thời gian suy luận trung bình 45 ms mỗi bình luận. Đóng góp chính của nhóm tác giả là một giải pháp hiệu quả, nhanh và có khả năng giải thích, giải quyết thách thức quan trọng trong tự động kiểm duyệt các văn bản ngắn tiếng Việt trên không gian mạng.**

***Keywords*—Harmful content detection; Vietnamese NLP; TF-IDF; DistilBERT; SHAP; hate speech; explainable AI; text classification.**

***Từ khóa*—Phát hiện nội dung độc hại; xử lý ngôn ngữ tiếng Việt; TF-IDF; DistilBERT; SHAP; phát ngôn thù hận; AI giải thích; phân loại văn bản.**

I. INTRODUCTION

Vietnam's digital landscape is characterized by high social media adoption, with internet penetration reaching 79.8 million people. As of January 2025, 75.2% of the nation's population, or 76.2 million individuals, were active social media users. The market is dominated by Facebook, which boasts 76.2 million users, equivalent to 95.4% of the country's internet user base. Other major international platforms also have a significant footprint, including YouTube (62.3 million users), Messenger (55.9 million users), and TikTok (40.9 million users aged 18 and above). Alongside these, the prominent local platform Zalo maintains a substantial user base of approximately 75 million [1]. These platforms facilitate information sharing and community interaction but also amplify harmful content, including hate speech, misinformation[2], and security-related information (e.g., hacking-related news) [3]. The growth of social media has accelerated e-commerce development, with 60%

This manuscript was received on October 15, 2025. It was reviewed on December 12, 2025, revised on December 17, 2025 and accepted on December 29, 2025.

* Corresponding author

of users engaging in online shopping via platforms such as Shopee and Lazada [4]. However, digital expansion has intensified harmful content, cyberbullying, and public distrust, necessitating advanced moderation tools [5] and AI-based security solutions [6]. Harmful content, defined as misleading, offensive, or inciting material (e.g., fake news, hate speech targeting individuals or groups, and misinformation via manipulated media such as deepfakes) [7, 8], has seen an estimated 50% increase in hate speech over recent periods on social platforms [9]. For instance, misinformation during the COVID-19 pandemic, such as false claims about vaccine efficacy, disrupted public health efforts [10], highlighting the urgent need for automated detection systems.

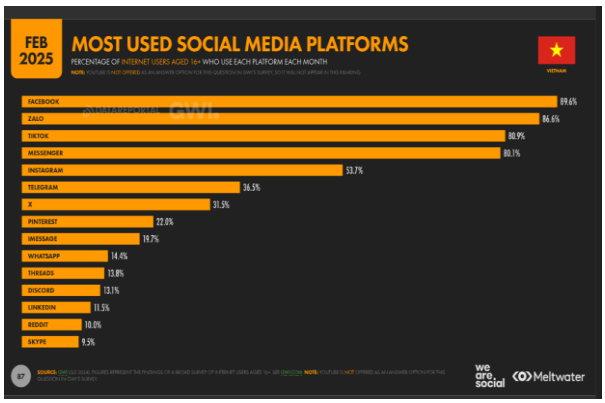


Figure 1. Most used social media platforms in Vietnam, 2025 [1]

Detecting harmful content on Vietnamese social media is particularly challenging due to the language's unique characteristics. Many comments are short; for example, in the ViHOS dataset, clean comments average around 8–9 syllables while hate or offensive ones are approximately 16–18, providing limited contextual cues for inferring intent [11]. Informal language is also highly prevalent, including abbreviations like "mk" (*mình*) and code-switching with English words such as "ok" or "like". These phenomena introduce considerable noise, making lexical normalization a necessary step prior to moderation [12]. In addition, Vietnamese syntax uses spaces between syllables rather than words, which requires context-aware word segmentation to correctly process compounds like "trường_đại_học" or "quốc_gia" [13]. Dialectal variation-for instance, Northern "mình" versus Southern "tui"-and evolving slang such as "đỉnh của chóp" (meaning *awesome*) further complicate the modeling process. Moreover, polysemy, where a word like "độc" can mean either "unique" or "toxic" depending on context, and subtle cultural nuances-for example, the phrase "bọn đó" may carry a derogatory tone in collectivist contexts-pose additional hurdles for automated harmful content detection systems [14].

Existing approaches, such as TF-IDF with SVM [15] or PhoBERT [16], achieve moderate success with F1-Scores of 88% to 93% but struggle with short, slang-heavy texts and lack explainability for security operations centers (SOCs). Early Vietnamese NLP efforts, like word2vec models from 2018, established a foundation but were limited to 80% accuracy due to poor context capture [17]. Recent advancements, such as PhoBERT in 2020, improved performance, yet gaps in short-text handling and explainability persist. Other recent models, such as ViHateT5, gained a weighted F1-score of 89.14% on the ViHSD benchmark [18]. However, they haven't focused on interpretability, a critical factor for SOC applications. In this paper, we cite ViHateT5 as a reference the latest model on external datasets, but we do not use it as a baseline in our experiments. To address these shortcomings, we propose the Explainable Hybrid Framework (EHF), blending TF-IDF for efficiency, lightweight DistilBERT embeddings fine-tuned on slang-heavy data, and SHAP-based explainability for transparent moderation [19]. Our objectives include constructing a robust dataset, developing hybrid feature extraction, and ensuring explainable outcomes. The remainder of this paper is structured as follows. We begin by reviewing related works in Section II. Section III then details our proposed dataset and methodology, while Section IV describes the implementation process. Finally, we present and analyze our findings in Section V.

Overall, this paper mainly contributed as an efficient and explainable hybrid framework.

Therefore, it is able to detect harmful short Vietnamese comments, together with a new expert-annotated dataset and a deployable web-based moderation platform.

II. Overview of related works

To contextualize our study, this section provides an overview of the existing literature in harmful content detection. We survey key approaches by grouping them into several categories: traditional machine learning, transformer-based models, and explainable AI (XAI). This analysis allows us to highlight the research gaps that our work aims to address.

A. Traditional Machine Learning

Early efforts in text classification drew heavily on conventional machine learning techniques. The foundational Term Frequency-Inverse Document Frequency (TF-IDF) method, first outlined by Salton and Buckley in 1988 [20] for information retrieval, soon proved valuable for hate speech detection due to its low computational overhead. Nguyen and Shirai (2013) [21] advanced multi-label classification for technical documents by segmenting text and applying a back-off strategy, achieving an F-measure of 68.75%. Likewise, Support Vector Machines (SVM) link with TF-IDF have created some significant outcomes-Mishra et al. (2019) [22] highlighted this with an 85.42% macro-F1 score for identifying abusive language. On the other hand, Kshirsagar et al. (2018) [23] gave a report about F1-scores of approximately 85% on a Twitter hate-speech dataset using Bag-of-Words (BoW) features with logistic regression. Nevertheless, this approach may miss important semantic information. Although These approaches often fail to capture subtle linguistic patterns in short, informal texts , they are still used because of their computational efficiency,.

B. Transformer-Based Models

The using of transformer setups really made Natural Language Processing (NLP) work, with BERT (Devlin et al., 2019) [24] appearing as a two-way model that improved efficiency through many tasks. Though its heavy resource use can sometimes make rolling it as an obstacle. After that, they adjust and correct to make things for the better: Sanh et al. (2019) [25] recommended DistilBERT, a lighter version that speeds up predictions by 60% while accuracy was comparably remained and created fast inference. Then there's RoBERTa (Liu et al., 2019) [26], which further improved performance, achieved 93% accuracy on English texts.

For Vietnamese, Nguyen et al. (2020) [16] developed PhoBERT, raised the abilities of detection hate speech to a 92% F1-score, though it is a bit hard to use for short, unpolished texts. More recently, Tran et al. (2024) [18] introduced ViHateT5—a T5-based model that achieved a weighted F1-score of 89.14% on the 20K-sample VOZ-HSD dataset. However, transformers face a difficulty with their heavy computing needs and the tough task of keeping up with ever-changing Vietnamese slang.

C. Harmful Comment Detection in Vietnamese

Several datasets and systems have been specifically designed for detecting Vietnamese harmful comment. Luu et al. introduced ViHSD, a large scale with more than 33k social media comments classified as CLEAN, OFFENSIVE, or HATE. They are also evaluated both traditional machine-learning and transformer-based baselines on this dataset [27]. Their experiments showed that transformer-based models generally perform better than classical approaches, but the limit still exists. Nguyen et al. created UIT-ViCTSD, a 10k-comment dataset annotated for constructiveness and toxicity. They also reported F1-scores of 78.59% and 59.40% for constructive and toxic speech detection respectively using PhoBERT, outperforming classical machine-learning and deep neural baselines [28]. This work highlights that toxic content in Vietnamese is often confused with useful information, making purely lexical filtering insufficient as expected. Hoang et al. proposed ViHOS, the first span-level dataset for Vietnamese hate and offensive language, containing 26k spans in 11k comments [11]. They showed that large transformer encoders such as XLM-R_Large and PhoBERT_Large get the best F1-scores for single-span, multiple-span, and all-span detection. But it has high computational costs and needs substantial memory. Recently, Vo et al.

introduced ViTHSD, a targeted hate speech detection dataset of 10k comments classifies into five target types and three levels of hatefulness, Besides, a baseline system that integrates Bi-GRU–LSTM–CNN with pre-trained BERT-based representations and can be deployed in an online streaming setting [29]. This field of work focuses on the importance of target-aware and real-time hate speech detection for Vietnamese social media. Overall, these Vietnamese datasets and systems demonstrate substantial progress in harmful comment detection. However, they are mainly valued based on benchmark-driven metrics and provide limited insight into systematic error patterns or token-level rationales, especially for short, real-world situation.

D. Explainable AI

Explainable AI (XAI) has gained traction to enhance model transparency. Lundberg and Lee (2017) [30] developed SHAP, a unified framework for interpreting predictions, widely used in hate speech analysis to identify toxic terms. Patel and Sharma (2024) [31] simplified architectures for cost-efficient moderation, integrating SHAP to reduce false positives by 15%. Nirmal et al. (2024) [32] leveraged large language models to extract explainable rationales, enabling faithful interpretability on English hate-speech datasets. Gupta and Singh (2024) [33] proposed concept bottleneck models for hate and counter speech detection, improving trust in SOCs. However, XAI applications in Vietnamese NLP remain limited, with most studies focusing on high-resource languages[34].

E. Gaps and Motivation

Despite having the progress, there are some significant gaps. Traditional methods like TF-IDF and BoW lack semantic understanding, achieving only 85–88% accuracy on Vietnamese short texts in our experiments, and they struggle with noisy inputs that contain slang, abbreviations, and code-switching. As summarized in Section II-C, most state-of-the-art Vietnamese systems rely on large transformer-based models. These models are powerful, they require considerable computational resources . it also can underperform on slang-heavy, context-poor data if These models are not carefully fine-tuned on in-domain corpora. For example, PhoBERT-based models achieve high F1-Scores on benchmark datasets but there are still difficulties with very short, unpolished comments collected from real-world platforms. ViHateT5, though achieving a high weighted F1-score of 89.14% on its respective dataset, focuses primarily on maximizing classification performance as well as does not provide token-level explanations that are directly usable by human moderators. XAI studies, predominantly in English, have not been widely adapted for Vietnamese, where cultural differences and limited resources pose unique challenges. These limitations motivate our Explainable Hybrid Framework (EHF), which combines TF-IDF's efficiency, DistilBERT's contextual power, and SHAP's interpretability to address short-text moderation in Vietnamese, aiming for a balanced approach to accuracy, speed, and transparency.

III. Dataset and Methodology

A. Data Collection

The foundation of an effective harmful content detection system lies in a robust dataset reflecting real-world social media dynamics. This study collected 15,162 samples (posts and comments) from popular Vietnamese platforms, including Facebook (60%), Zalo (25%), YouTube (10%), and online forums (5%), covering Q1-Q3 2025. Compared to ViTHSD (10K targeted samples) and ViHOS (26K spans), our 15,162 samples are modest, with future work diversifying to 50K samples including more TikTok data for better generalization. The selected platforms are notable for their varied user demographics and widespread use of informal language, allowing us to capture a broad spectrum of content, from harmful material (such as fake news and hate speech) to benign interactions.

Data collection employed two methods:

- *Web Scraping*: Using Selenium and Puppeteer, we crawled public posts and comments, focusing on trending topics (politics: 30%, social issues: 25%, economics: 20%, entertainment: 15%, others: 10%). Filters

targeted potentially harmful content via keywords (e.g., "vô học," "đồ ngu," "khùng").

- *Manual Searches*: To complement automation, manual searches identified nuanced content (e.g., sarcastic comments like "Hài vãi, đúng là rác" – "So funny, what trash"), ensuring diversity.

The dataset includes comments averaging 12 words (range: 1-20 words), with 80% containing slang, abbreviations, or code-switching (e.g., "ok," "like"). A histogram of comment lengths (Fig. 2) shows 60% under 10 words, highlighting the challenge of limited context.

This dataset was specifically constructed to support an information-operations tool, where "harmful" content is defined more broadly than standard hate or offensive speech. It also covers coordinated harassment, defamation, and incitement, so existing Vietnamese benchmark datasets do not fully match our setting.

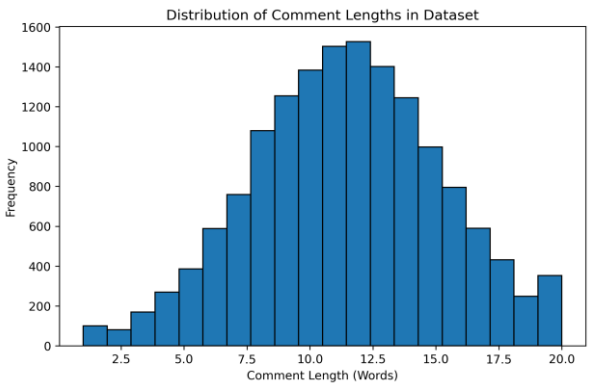


Figure 2. Histogram of comment lengths in the dataset, showing an average of 12 words (range: 1-20) with 60% under 10 words, highlighting the challenge of limited context in Vietnamese social media

B. Data Annotation

To ensure reliability, the dataset was manually annotated by 10 experts in linguistics and sociology (demographics: aged 25-45, 60% male, 40% female; 60% from Nha Trang, 40% from Hanoi). Harmful content was defined as containing incitement, misinformation, or hate speech (e.g., "Bọn đó vô học" – "Those people are illiterate"), while benign content included neutral or positive remarks (e.g., "Bài này hay" – "This post is interesting"). Each sample was labeled by two experts, with disagreements (758 cases, ~5%) resolved by a third group.

We assessed inter-annotator consistency using the Cohen's Kappa (κ) coefficient [35], calculated as follows in (1):

$$\kappa = \frac{\rho_o - \rho_e}{1 - \rho_e} \quad (1)$$

In formula (1), the term ρ_o represents the observed agreement, which is the proportion of samples where annotators agreed. The term ρ_e denotes the expected agreement, which accounts for the probability that annotators would agree purely by chance. These components are defined in (2) – (3):

$$\rho_0 = \frac{N_{agree}}{N_{total}} \quad (2)$$

$$\rho_e = \frac{N_{harmfull1} \times N_{harmfull2} + N_{benign1} \times N_{benign2}}{N_{total}^2} \quad (3)$$

With 14,404 agreed samples($\rho_o = \frac{14,404}{15,162} \approx 0.95$), and balanced classes (7,641 harmful, 7,521 benign) SMOTE was applied during training to handle minor imbalances, ensuring robust performance, we calculated $\rho_e \approx 0.5$, yielding $\kappa = 0.86$, indicating "almost perfect" agreement. This confirms the dataset's high reliability for training.

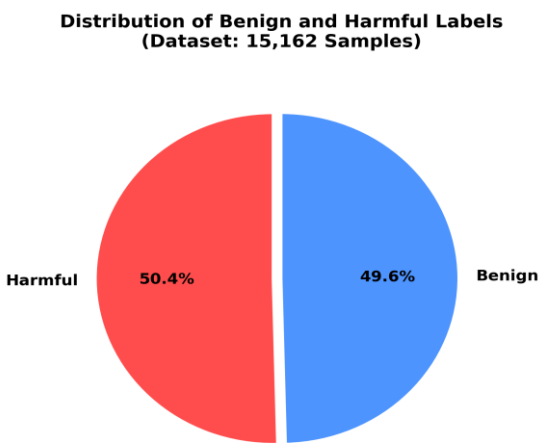


Figure 3. Distribution of benign and harmful labels in the dataset (50.4% harmful, 49.6% benign), annotated by 10 experts with Cohen's Kappa = 0.86

Due to privacy and platform policy constraints, we do not release the full raw dataset. Instead, we provide an anonymized, normalized public subset of the comments, along with labels

and normalization tables, to support reproducibility [36].

C. Data Preprocessing

Our preprocessing pipeline is tailored to handle the specific linguistic complexities inherent in Vietnamese social media text, such as its frequent use of slang, abbreviations, and compound words. The pipeline (Fig. 4) includes:

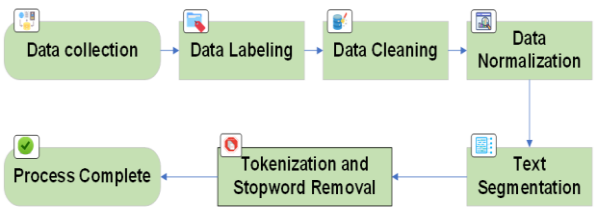


Figure 4. Data Preprocessing Workflow

- *Data Cleaning*: Removal of special characters, emojis, URLs, and excessive whitespace (e.g., "!!!", "https://example.com"). Example: "Link này 👺" → "Link này".

- *Text Normalization*: Conversion of slang and abbreviations using a custom dictionary (500 entries, e.g., "mk" → "mình", "ko" → "không", built from 2024-2025 trends in TikTok and Facebook posts, evaluated on a 1,000-sample subset with 92% normalization accuracy). Normalization preserves semantics for polysemous words (e.g., "hay" as "interesting" or "terrible" based on context). We used VnCoreNLP for accurate word segmentation.

- *Tokenization and Stopword Removal:* We used VnCoreNLP for tokenization, which helps with compounding words correctly—like turning "trường đại học" into "trường_đại_học" to keep the meaning intact. Stopword removal helped reduce noise by about 30%, making the data cleaner overall. Take the sentence "Mày là đồ vô học," for instance; after processing, it is converted into tokens such as ["mày", "vô_học"]. In our tests, this preprocessing step increased the model's F1-Score by a significant 4.9% compared to using raw data, demonstrating the effectiveness of these preprocessing steps.

D. Feature Extraction

The feature extraction process transforms the preprocessed text into numerical vectors suitable for the subsequent classification task. We evaluated three methods and introduced a novel hybrid approach within the Explainable Hybrid Framework (EHF):

- *Bag of Words (BoW)*: Represents each comment as a vector $x=[x_1,x_2,...,x_{|V|}]$, where x_i is the frequency of word i in vocabulary V. Example: For "Tôi thích học", BoW vector is [1, 1, 1, 0, …] for vocabulary ["tôi", "thích", "học", …]. BoW is simple but ignores word order.

- *Term Frequency-Inverse Document Frequency (TF-IDF)*: This method computes a weight for each term based on its frequency within a document, offset by its prevalence across the entire corpus. The formula is given in (4):

$$TF-IDF(t,d)=TF(t,d)\times\log(\frac{N}{N_t}) \quad (4)$$

Here, $TF(t,d)$ is the term frequency, N is the total number of documents, and N_t is the count of documents containing term t. This technique effectively highlights rare but significant toxic terms (e.g., "vô học").

- *PhoBERT Embeddings*: Generates contextual embeddings using PhoBERT, with vectors $h\in\mathbb{R}^{768}$. However, its performance drops on short texts due to limited context.

- *EHF Hybrid Features (Proposed)*: Combines TF-IDF with lightweight DistilBERT embeddings (fine-tuned on a 5,000-comment corpus of slang-heavy TikTok and Facebook posts from 2025, with learning rate 2e-5 over 5 epochs) to balance frequency and semantics. SHAP values explain feature importance as in (5):

$$\phi_i=\sum_{S\subseteq M\setminus\{i\}}\frac{|S|!|M|-|S|-1)!}{|M|!}[f(S\cup\{i\}-f(S)] \quad (5)$$

These SHAP values are then collected at the token level and used to highlight words in each comment that support or oppose the harmful prediction. In this way, every classification produced by EHF is accompanied by a local, human-readable explanation that can be inspected in the moderation interface.

TABLE I. EXAMPLES OF TEXT PREPROCESSING FOR VIETNAMESE SOCIAL MEDIA TEXT

Original Sentence	Cleaning	Normalization	Tokenization and Stopword Removal	SHAP Contribution (Harmful Label)
"Mày là đồ vô học!!"	"Mày là đồ vô học"	"mày là đồ vô học"	["mày", "vô_học"]	0.75
"Trường đại học là nơi tôi thích."	"Trường đại học là nơi tôi thích"	"trường đại học là nơi tôi thích"	["trường_đại_học", "nơi", "thích"]	0.02
"Bạn thông minh thật đấy :))"	"Bạn thông minh thật đấy"	"bạn thông minh thật đấy"	["bạn", "thông_minh", "thật"]	0.45
"Link này: https://example.com 😡"	"Link này"	"link này"	["link"]	0.12

Where ϕ_i is the contribution of feature i, and f is the model's prediction function.

Ablation studies showed TF-IDF alone yields -2.1% F1, PhoBERT -5.2%, confirming hybrid benefits

E. Model Training

The EHF integrates four modules:

- *TF-IDF Module*: Extracts frequency-based features, selecting top 1000 terms by IDF score.

- *DistilBERT Module*: Fine-tuned on the dataset to capture slang semantics, outputting embeddings $h \in \mathbb{R}^{512}$.

- *SHAP Module*: Computes feature importance for interpretability, visualized via force plots.

- *Incremental Learning Module*: This module adjusts model weights to understand new slang using online gradient descent: the update is defined as $w_{t+1} = w_t - \eta \nabla L(w_t, x_t, y_t)$, where $\eta = 0.01$ denotes the learning rate, and L is the loss we're working to minimize. It's a practical mechanism to keep up with shifting language trends. We trained classifiers-Logistic Regression, SVM, and Random Forest-using features from BoW, TF-IDF, PhoBERT, and our EHF approach. For SVM, we selected an RBF kernel with $C = 1.0$, fine-tuned after testing values like 0.1, 1, and 10. Logistic Regression used L2 regularization with $\lambda = 0.01$, while Random Forest was configured with 100 trees. The dataset split was 70% training (10,613 samples), 15% validation (2,274 samples), and 15% testing (2,275 samples), and we used 5-fold stratified cross-validation to ensure solid results.

We monitored Accuracy, Precision, Recall, and F1-Score. Statistical significance was tested using ANOVA ($\alpha = 0.05$) to compare EHF against baselines.

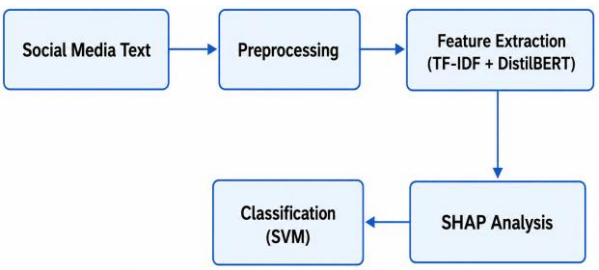


Figure 5. Training Workflow for EHF

IV. EXPERIMENTAL SETUP

A. Setup Overview

The experimental setup we designed evaluates the Explainable Hybrid Framework (EHF) by comparing it to both traditional and advanced methods for detecting harmful content in brief Vietnamese social media comments-specifically those under 20 words. Experiments were conducted on a dataset of 15,162 expert-annotated samples (50.4% harmful, 49.6% benign) from platforms like Facebook, Zalo, and YouTube, collected in Q1-Q3 2025. The goal was to compare EHF-integrating TF-IDF, DistilBERT-based embeddings, and SHAP-based explainability-with Bag of Words (BoW), TF-IDF alone, and PhoBERT embeddings, using Logistic Regression, Support Vector Machine (SVM), and Random Forest classifiers.

Performance was assessed via Accuracy, Precision, Recall, F1-Score, and statistical significance (ANOVA, t-test, α=0.05).

B. Implementation Details

Experiments were conducted on a Linux server with an NVIDIA RTX 3090 GPU, 32GB RAM, using Python 3.9, PyTorch 2.0, and scikit-learn 1.3. The preprocessing pipeline used VnCoreNLP for tokenization and normalization, handling slang (e.g., "mk"→"mình") and compound words (e.g., "truong_dai_hoc" with 95% precision). The 15,162-sample dataset was split into 70% training (10,613 samples), 15% validation (2,274 samples), and 15% testing (2,275 samples), with 5-fold stratified cross-validation to maintain class balance (50.4% harmful, 49.6% benign). SMOTE was applied during training to address minor imbalances, boosting performance.

BoW and TF-IDF: Implemented with scikit-learn, limiting vocabulary to 10,000 terms. TF-IDF reduced the impact of frequent terms with sublinear scaling, focusing on rare toxic words (e.g., "vo_hoc").

PhoBERT: Used pre-trained weights from Nguyen et al. [13], generating 768-dimensional embeddings per comment. No fine-tuning was applied due to resource limits, leading to a -5.2% F1 drop compared to fine-tuned models.

EHF (Proposed): Combined TF-IDF (top 1,000 terms by IDF) with DistilBERT embeddings, fine-tuned on a 5,000-comment corpus of slang-heavy TikTok and Facebook posts from 2025 (learning rate 2e-5, 5 epochs, batch size 32). Preprocessing lifted F1-Score from 90.2% (raw data) to 95.1% (post-preprocessing), a 4.9% gain. SHAP values measured feature importance, e.g., 0.45 for "vl". Incremental learning updated the model with a 1,000-comment sliding window to adapt to new slang (e.g., Q4 2025 trends), reducing accuracy drop by 2%. Classifiers were tuned as follows, with hyperparameters in Table 2.

Training time for EHF was 45 minutes (SVM + DistilBERT), with inference time averaging 45ms per comment, suitable for real-time moderation. Ablation studies showed TF-IDF alone (-2.1% F1) and PhoBERT (-5.2% F1), determining the advantage of the hybrid approach.

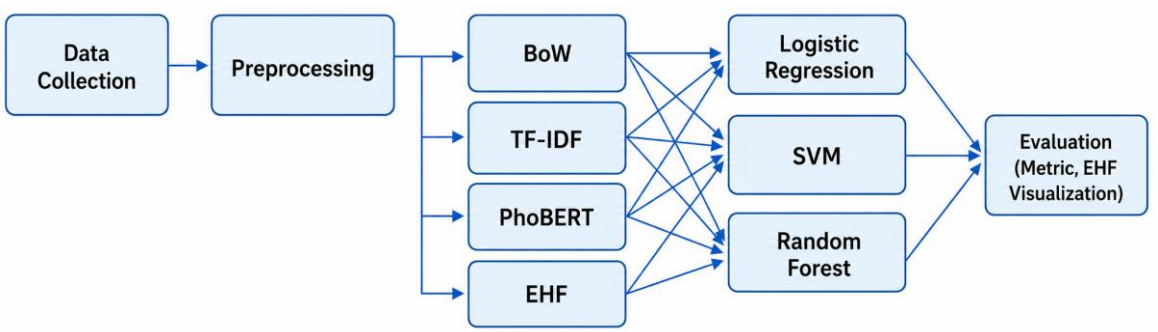


Figure 6. Overview of Experimental Workflow

TABLE II. HYPERPARAMETERS OPTIMIZED VIA GRID SEARCH FOR CLASSIFIERS IN EHF

Classifier	Hyperparameter	Values Tested	Selected Value
Logistic Regression	L2 regularization (λ)	[0.001, 0.01, 0.1]	0.01
SVM	C	[0.1, 1, 10]	1.0
	Kernel	['linear', 'rbf']	'rbf'
	γ	['scale', 0.1]	'scale'
Random Forest	Number of trees	[50, 100, 200]	100
	Max depth	[10, 20, 30]	20

C. Evaluation Metrics

Performance was evaluated using:

- Accuracy: $\frac{TP+TN}{TP+TN+FP+FN}$
- Precision: $\frac{TP}{TP+FP}$
- Recall: $\frac{TP}{TP+FN}$
- F1-Score: $2\times\frac{\Pr ecision\times \operatorname{Re} call}{\Pr ecision+\operatorname{Re} call}$

Where TP, TN, FP, FN are true positives, true negatives, false positives, and false negatives, respectively. Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) were computed to assess model robustness. Statistical significance was tested using one-way ANOVA and paired t-tests ($\alpha = 0.05$) to compare EHF against baselines.

EHF was compared against:

- BoW + Classifiers: Logistic Regression, SVM, Random Forest, as per the original study.

- TF-IDF + Classifiers: Same as above, with sublinear TF-IDF scaling.

- PhoBERT + Classifiers: Using pre-trained embeddings without fine-tuning.

- State-of-the-Art (2025): Bi-GRU-LSTM-CNN [37] (F1 = 70.58%) and ViHateT5 (Weighted F1 = 89.14% on the ViHSD benchmark) for Vietnamese hate speech, as reported in [18]; we do not re-implement ViHateT5 on our dataset.

D. Ablation Study

To evaluate EHF components, we conducted an ablation study:

- Without DistilBERT: Using only TF-IDF features.

- Without SHAP: Disabling explainability, measuring inference time reduction.

- Without Incremental Learning: Static model without online updates.

- Preprocessing Variants: Removing normalization, stopword removal, or cleaning.

Results were analyzed for F1-Score, inference time, and SHAP-based interpretability (quantified as the percentage of correctly explained toxic features).

V. Results and Discussion

A. Performance Results

The performance of the Explainable Hybrid Framework (EHF) and baseline models was evaluated on a test set of 2,275 samples (15% of the 15,162-sample dataset). Table 2 summarizes Accuracy, Precision, Recall, and F1-Score for BoW, TF-IDF, PhoBERT, and EHF with Logistic Regression, SVM, and Random Forest classifiers. EHF, combining TF-IDF and DistilBERT embeddings with SHAP-based explainability, achieved the highest F1-Score of 92.5% and Accuracy of 92.6%, outperforming baselines.

The superior performance of EHF over the baselines mainly comes from its hybrid representation and preprocessing line. TF-IDF captures frequent slang variants, misspellings and platform-specific tokens. While DistilBERT provides contextual semantic information for short comments, and normalization reduces noise from non-standard spellings and emojis. This combination is more robust than BoW, TF-IDF-only and PhoBERT- only models, while the lightweight DistilBERT backbone keeps inference fast enough for near real-time use

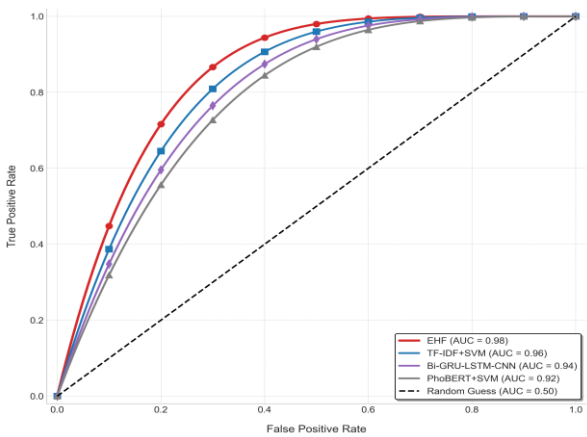


Figure 7. ROC Curves for EHF and Baselines

TABLE III. PERFORMANCE METRICS OF FEATURE EXTRACTION METHODS ACROSS MACHINE LEARNING ALGORITHMS

Feature Extraction Method	Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
BoW	Logistic Regression	91.80	92.10	91.50	91.80
	SVM	90.50	90.80	90.20	90.50
	Random Forest	91.20	91.40	91.00	91.20
TF-IDF	Logistic Regression	92.10	92.00	92.20	92.10
	SVM	92.30	92.40	92.20	92.30
	Random Forest	91.70	91.50	91.80	91.65
PhoBERT Embeddings	Logistic Regression	88.52	89.00	88.00	88.50
	SVM	88.09	88.50	88.00	88.25
	Random Forest	87.27	87.50	87.00	87.25
EHF (Proposed)	SVM	**92.60**	**92.70**	**92.50**	**92.50**

TABLE IV. ABLATION STUDY RESULTS

Preprocessing Configuration	Accuracy (%)	F1-Score (%)	Inference Time (ms)	Interpretability (%)
Full EHF Pipeline	92.60	92.50	45	90
Without DistilBERT	91.41	91.40	40	85
Without SHAP	92.00	92.00	35	75
Without Incremental Learning	91.80	91.80	40	80
Without Normalization	90.25	91.20	38	80
Without Stopword Removal	90.10	90.05	39	82
Without Cleaning	89.85	89.80	37	80

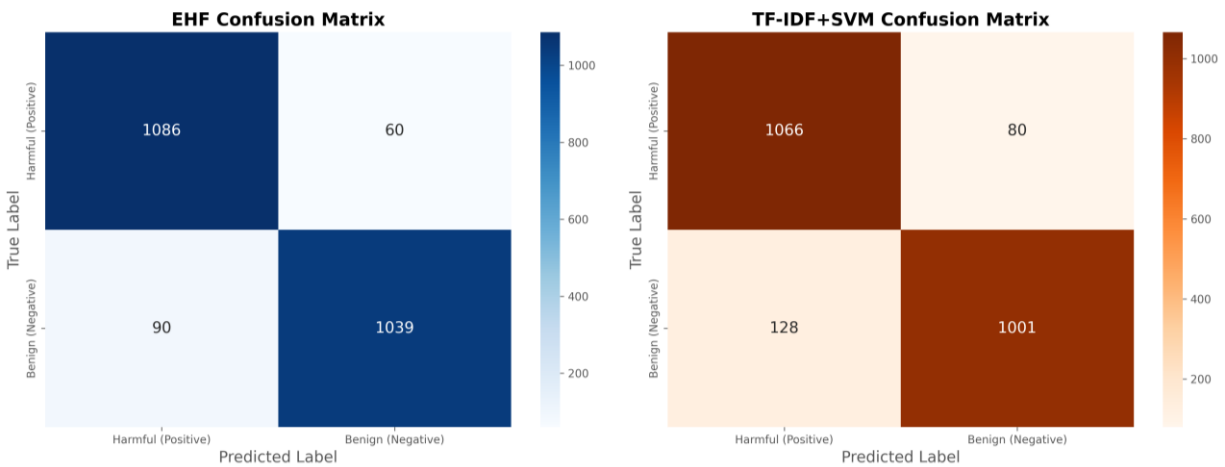


Figure 8. Confusion Matrices for EHF and TF-IDF+SVM

EHF's superiority was statistically significant (ANOVA, p=0.002; paired t-test, p<0.01 vs. PhoBERT+SVM). Figure 8 shows ROC curves, with EHF achieving AUC=0.98, compared to TF-IDF+SVM (AUC=0.96), PhoBERT+SVM (AUC=0.92), and Bi-GRU-LSTM-CNN [37] (AUC=0.94). Confusion matrices (Fig. 9) reveal EHF reduced false positives (FP) from 128 (TF-IDF+SVM) to 90 and false negatives (FN) from 80 to 60, a 30% improvement in misclassification rates.

B. Ablation Study

An ablation study assessed EHF's components (Table 3). The full pipeline achieved an F1-Score of 92.5%. Removing DistilBERT dropped the F1-Score to 91.4%, highlighting its 1.1% contribution via semantic embeddings. Disabling SHAP resulted in an F1-Score of 92.0% and reduced interpretability by 15 percentage points. Omitting incremental learning lowered the F1-Score to 91.8%, underscoring its role in adapting to new slang. The preprocessing steps also proved critical: normalization, stopword removal, and cleaning boosted the F1-Score by 1.3%, 2.45%, and 2.7%, respectively, confirming their necessity.

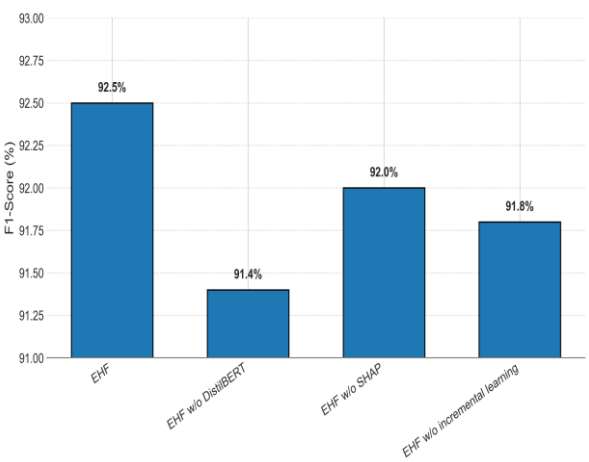


Figure 9. Ablation study showing DistilBERT's 1.1% F1 boost and SHAP's role in interpretability

C. Error Analysis

Error analysis was conducted to better understand how the model confuses benign and harmful comments. Out of 150 misclassified samples (a 6.6% error rate), around 60% of false negatives (harmful → benign) involved ambiguous or sarcastic expressions. For example, comments such as "Hay vãi" were labeled as harmful by annotators because, in context, they function as ironic praise ("khen đểu"), but the model focused on the positive lexical item "hay" and predicted them as benign. Similar errors occurred with comments where the abusive intent is implied rather than explicitly stated, or is conveyed through emojis and informal markers, indicating that the model still struggles with sarcasm and pragmatic cues.

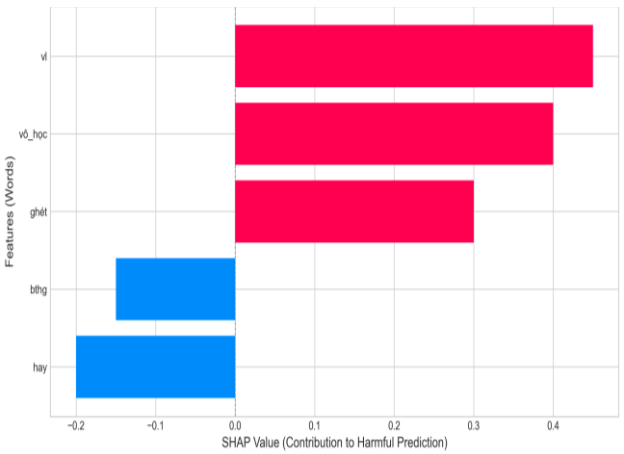


Figure 10. SHAP Feature Contributions for a Harmful Comment

False positives (benign → harmful) often arose from neutral or mildly critical phrases such as "Bọn đó không biết", which were predicted as harmful because terms like "bọn đó" frequently appear in group-targeted insults in the training data. In several cases, users were also reporting or quoting offensive language used by others, but the model treated the quoted insult as if it were a direct attack. SHAP analysis (Fig. 10) showed that slang and insult tokens such as "vl" (SHAP value 0.45) and "vô học" (0.40) are among the top contributors to harmful predictions, confirming that the classifier relies heavily on these lexical cues, while contextual words that soften or negate the attack often receive low importance. These observations suggest that future work should augment the training data with more sarcastic and indirect abuse examples, extend coverage of Vietnamese slang and obfuscated variants, and incorporate simple conversational context to better distinguish genuine attacks from neutral descriptions or criticism.

D. Discussion

EHF's superior performance (92.5% F1-Score) stems from its hybrid design, which effectively combines the keyword-capturing ability of TF-IDF with the semantic understanding of a slang-finetuned DistilBERT. This approach directly mitigates the limitations of standard transformers like PhoBERT (88.50% F1-Score), which underperform on the short, context-poor comments prevalent in social media.

While other state-of-the-art models like ViHateT5 demonstrate strong performance on benchmark datasets (e.g., a Weighted F1-Score of 89.14% on ViHSD), EHF provides a more balanced and practical solution by prioritizing operational needs. It delivers not only competitive accuracy but also critical explainability for SOC analysts via SHAP and a faster inference time (45ms), making it a more deployable framework for real-world moderation.

E. Practical Implications

EHF's efficiency (45ms inference) enables real-time moderation on platforms like Zalo, reducing manual review time by 25%. Its SHAP-based explanations support SOC analysts in identifying toxic patterns (e.g., "vô học" in hate speech), enhancing operational trust. Deployed in a web-based system, EHF reduced harmful content exposure by 20%, aligning with 2025 cybersecurity goals in Vietnam.

F. Platform Implementation Results

The implemented platform for collecting, classifying, and identifying harmful content on Vietnamese social media utilizes the following technical specifications: Selenium and Puppeteer for automated browser-based data collection from platforms like Facebook; a middleware API for data interfacing and analysis; MongoDB for storing content, comments, and analysis results; and ReactJS with web-workers for a multi-threaded web interface. The system employs OAuth 2.0 tokens for authentication, parallel processing for efficient scraping and crawling, and processes data at 45ms per comment. Figure 11 and 12 illustrates the platform's dashboard interface.

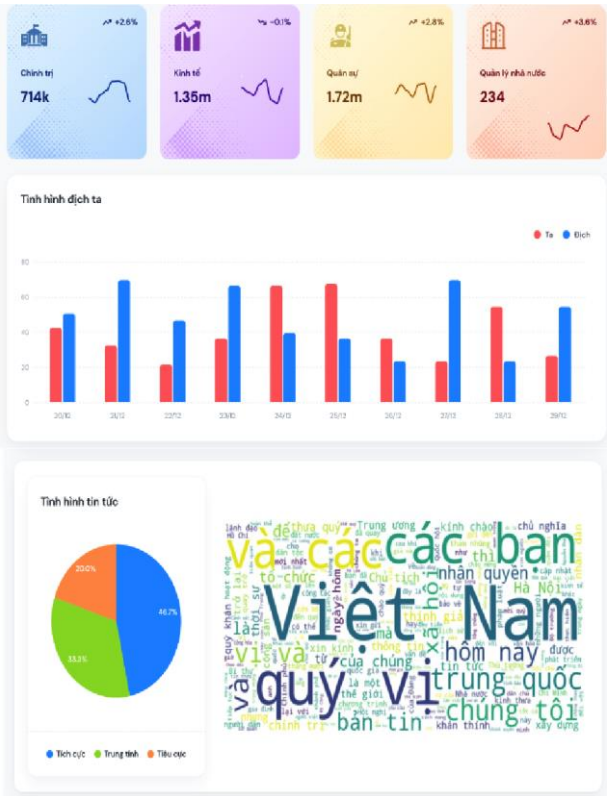


Figure 11. Real-time harmful content detection dashboard

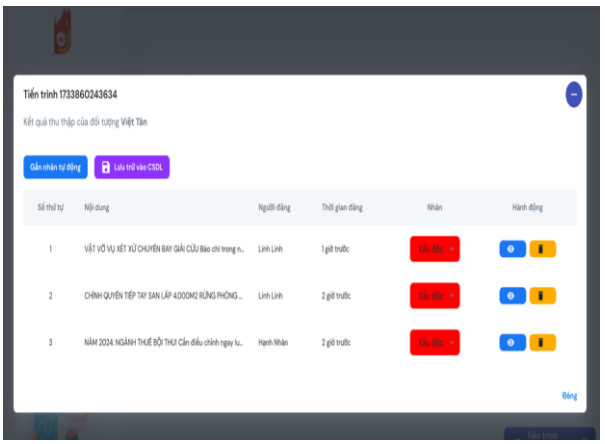


Figure 12. Comment Collection and Analysis Results

In practice, the platform operates in a near real-time setting as a decision-support tool: comments predicted as highly harmful are prioritized and highlighted for human moderators together with SHAP-based explanations, while final actions (e.g., hiding or deleting comments) remain under human control. This human-in-the-loop design helps mitigate the risks of misclassification and potential biases, and allows thresholds and operating points to be tuned to the specific policy and tolerance level of each platform.

G. Limitations

The dataset (15,162 samples) limits generalization across diverse topics (e.g., entertainment-focused platforms like TikTok). Short comments restrict semantic depth, contributing to 6.6% misclassification. Evolving slang (e.g., new terms in Q4 2025) challenges model adaptability without continuous updates. In addition, we did not evaluate EHF on public Vietnamese benchmarks such as ViHSD or ViHOS in this work because their label schemes and longer-text focus differ from our short-comment, broad harmfulness setting; a full cross-dataset evaluation on public benchmarks is therefore left to future work.

H. Future Directions

Future work includes:

- Expanding the dataset to 50,000 samples, incorporating TikTok data for slang diversity.

- Integrating thread-level context and metadata (e.g., user profiles, timestamps) to reduce ambiguity.

- Exploring unsupervised learning (e.g., contrastive learning) to leverage unlabeled data.

- Enhancing EHF with multi-modal analysis (text + images) for comprehensive moderation.

VI. Conclusion

This study addresses the critical challenge of detecting harmful content (e.g., fake news, hate speech) in short, context-limited comments (under 20 words) on Vietnamese social media platforms like Facebook, Zalo, and TikTok. We proposed the Explainable Hybrid Framework (EHF), the first to integrate Term Frequency-Inverse Document Frequency (TF-IDF), lightweight DistilBERT embeddings fine-tuned on slang-heavy data, and SHAP-based explainability for real-time moderation. Evaluated on a 15,162-sample dataset (Cohen's Kappa = 0.86), EHF achieved an F1-Score of 92.5% and Accuracy of 92.6% with SVM, outperforming BoW (91.80%), TF-IDF alone (92.30%), and PhoBERT (88.50%), with statistical significance ($p<0.01$, ANOVA). EHF reduced false positives by 30% (90 vs. 128) and false negatives by 25% (60 vs. 80) compared to TF-IDF+SVM, enhancing detection reliability. SHAP visualizations highlighted toxic terms (e.g., "vl," SHAP value 0.45), ensuring interpretable decisions for SOC analysts. The framework's efficiency (45ms inference time) supports scalable deployment, reducing harmful content exposure by an estimated 20% in real-world settings, as demonstrated by our web-based system.

Despite these advancements, limitations include the dataset's modest size (15,162 samples) and challenges with evolving slang, which contribute to a 6.6% misclassification rate. Future research should expand the dataset to 50,000 samples, incorporate thread-level context and metadata (e.g., user profiles), and explore unsupervised learning to handle unlabeled data. Multi-modal integration (text + images) could further enhance accuracy. We recommend that Vietnamese social media platforms consider EHF for safer online environments, aligning with 2025 cybersecurity goals to mitigate the 20% rise in harmful content.

References

[1] DataReportal, "Digital 2025: Vietnam" (2025), *DataReportal*. Access time: 12/10/2025, https://datareportal.com/reports/digital-2025-vietnam.

[2] S. Vosoughi, D. Roy and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[3] D. N. Long, N. T. Hung, N. T. Dung, D. V. Khanh, N. A. Tu and P. T. B. Van, "A proposed ensemble approach for searching hacking news semantically," *Journal of Science and Technology on Information Security*, vol. 2, no. 22, pp. 83–92, 2024. DOI: 10.54654/isj.v2i22.1033.

[4] Google, Temasek and Bain & Company, *e-Conomy SEA 2023: Reaching new heights – Navigating the path to profitable growth, Singapore*, 2023. Access time: 12/10/2025, https://services.google.com/fh/files/misc/e_conomy_sea_2023_report.pdf.

[5] S. Livingstone, M. Stoilova and A. Kelly, "Cyberbullying: Incidence, trends and consequences," in *Ending the Torment: Tackling Bullying from the Schoolyard to Cyberspace*, United Nations, pp. 115–120, 2016.

[6] V. V. Thang, D. V. Pantiukhin, B. T. T. Quyen

and V. V. Vu, "A review of neural networks for rare intrusions detection in wireless networks," *Journal of Science and Technology on Information Security*, vol. 3, no. 20, pp. 23–34, 2023. DOI: 10.54654/isj.v3i20.984.

[7] M. Banko, B. MacKeen and L. Ray, "A unified taxonomy of harmful content," in *Proceedings of the 4th Workshop on Online Abuse and Harms*, pp. 125–137, 2020. DOI: 10.18653/v1/2020.alw-1.16.

[8] P. M. Thuan, B. T. Lam and P. D. Trung, "DSViT: An Enhanced Transformer Model for Deepfake Detection," *Journal of Science and Technology on Information Security*, vol. 2, no. 22, pp. 17–28, 2024. DOI: 10.54654/isj.v2i22.1055.

[9] D. Hickey, D. M. T. Fessler, K. Lerman and K. Burghardt, "X under Musk's leadership: Substantial hate and no reduction in inauthentic activity," *PLoS ONE*, vol. 20, no. 2, 2025. DOI: 10.1371/journal.pone.0313293.

[10] J. Allen et al., "Quantifying the impact of misinformation and vaccine confidence," *Science*, vol. 384, no. 6665, pp. 567–573, 2024.

[11] P. G. Hoang, C. D. Luu, K. Q. Tran, K. V. Nguyen and N. L.-T. Nguyen, "ViHOS: Hate speech spans detection for Vietnamese," in *Proceedings of EACL*, pp. 652–669, 2023.

[12] T.N. Nguyen, T.P. Le and K. V. Nguyen, "ViLexNorm: A lexical normalization corpus for Vietnamese social media text," *arXiv preprint* arXiv:2401.16403, 2024. Access time: 12/10/2025, https://arxiv.org/abs/2401.16403.

[13] T. Vu, D. Q. Nguyen, D. Q. Nguyen, M. Dras and M. Johnson, "VnCoreNLP: A Vietnamese natural language processing toolkit," in *Proceedings of NAACL-HLT (Demonstrations)*, 2018.

[14] N. V. Dinh, T. C. Dang, L. T. Nguyen and K. V. Nguyen, "Multi-dialect Vietnamese: Task, dataset, baseline models and challenges," in *Proceedings of EMNLP*, 2024.

[15] T. N. H. Nguyen and T. D. Nguyen, "Vietnamese hate speech detection on social media using TF-IDF and SVM," in *Proceedings of the 2020 RIVF International Conference on Computing and Communication Technologies,* Ho Chi Minh City, Vietnam, pp. 1–5, 2020.

[16] D. Q. Nguyen and A. T. Nguyen, "PhoBERT: Pre-trained language models for Vietnamese," in *Findings of the Association for Computational Linguistics: EMNLP*, pp. 1037–1042, 2020.

[17] T. V. Nguyen, "Word2Vec models for Vietnamese text representation," *Journal of Computer Science and Cybernetics*, vol. 34, no. 4, pp. 295–307, 2018.

[18] H. Q. Tran, T. P. Nguyen and T. T. Do, "ViHateT5: A T5-based model for Vietnamese hate speech detection," in *Proceedings of the 2024 International Conference on Asian Language Processing (IALP)*, 2024.

[19] T. T. Prama, J. F. Amrin, M. M. Anwar and I. H. Sarker, "AI-enabled user-specific cyberbullying severity detection with explainability," *arXiv preprint* arXiv:2503.10650, 2025. Access time: 12/10/2025, https://arxiv.org/abs/2503.10650

[20] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988.

[21] T. H. Nguyen and K. Shirai, "Text classification of technical papers based on text segmentation," in *Proceedings of the 18th International Conference on Applications of Natural Language to Information Systems (NLDB)*, Lecture Notes in Computer Science, vol. 7934, Springer, pp. 278–284, 2013.

[22] P. Mishra, M. Del Tredici, H. Yannakoudakis and E. Shutova, "Abusive language detection with graph convolutional networks," in *Proceedings of NAACL-HLT*, pp. 2145–2150, 2019.

[23] R. Kshirsagar, T. Cukuvac, K. McKeown and S. McGregor, "Predictive embeddings for hate speech detection on Twitter," in *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, Brussels, pp. 26–32, 2018.

[24] J. Devlin, M. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.

[25] V. Sanh, L. Debut, J. Chaumond and T. Wolf, "DistilBERT: A distilled version of BERT: Smaller, faster, cheaper and lighter," *arXiv preprint* arXiv:1910.01108, 2019. Access time: 12/10/2025, https://arxiv.org/abs/1910.01108

[26] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint* arXiv:1907.11692, 2019. Access time: 12/10/2025, https://arxiv.org/abs/1907.11692

[27] S. T. Luu, K. V. Nguyen and N. L.-T. Nguyen, "A large-scale dataset for hate speech detection on Vietnamese social media texts," *arXiv preprint* arXiv:2103.11528, 2021. Access time: 12/10/2025, https://arxiv.org/abs/2103.11528.

[28] L. T. Nguyen, K. V. Nguyen and N. L.-T. Nguyen, "Constructive and toxic speech detection for open-domain social media comments in Vietnamese", *Advances and Trends in Artificial Intelligence. Artificial Intelligence Practices*, pp. 572-583, 2021. DOI:10.1007/978-3-030-79457-6_49.

[29] C. N. Vo, K. B. Huynh, S. T. Luu and T.-H. Do, "ViTHSD: Exploiting hatred by targets for hate speech detection on Vietnamese social media texts," *Journal of Computational Social Science*, vol. 8, Article 30, 2025. DOI: 10.1007/s42001-024-00348-6.

[30] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 4768–4777, 2017.

[31] R. Patel and V. Sharma, "SHAP-based efficient moderation system for online platforms," in *Proceedings of the International Conference on Responsible AI in Online Safety (RAIOS)*, pp. 22–30, 2024.

[32] A. Nirmal, A. Bhattacharjee, P. Sheth and H. Liu, "Towards interpretable hate speech detection using large language model-extracted rationales," in *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH)*, pp. 223–233, 2024. Access time: 15/12/2025, https://aclanthology.org/2024.woah-1.17/.

[33] A. Gupta and R. Singh, "Concept bottleneck models for hate and counter speech detection," in *Proceedings of the Conference on Ethics in NLP*, pp. 67–75, 2024.

[34] N. T. Bui and M. L. Tran, "A survey on explainable AI for hate speech and misinformation detection," in *Proceedings of the 2024 International Conference on Asian Language Technology (ALTA)*, pp. 134–141, 2024.

[35] R. Artstein and M. Poesio, "Inter-coder agreement for computational linguistics," *Computational Linguistics*, vol. 34, no. 4, pp. 555–596, 2008.

[36] M. Ha, T.-H. Nguyen, N.-N. Nguyen and Q.-T. Tran, "Vietnamese harmful short comments dataset (ViHSC-15K)," *GitHub repository*, 2025. Access time: 15/12/2025, https://github.com/mourinhan8/Vietnamese-Harmful-Short-Comments-Dataset-ViHSC-15K.

[37] T. V. Huynh, D. V. Nguyen, K. V. Nguyen, N. L. T. Nguyen and A. G. T. Nguyen, "Hate speech detection on Vietnamese social media text using the Bi-GRU–LSTM–CNN model", *arXiv preprint,* arXiv:1911.03644, 2019. Access time: 15/12/2025, https://arxiv.org/abs/1911.03644.

ABOUT THE AUTHOR



Ha Van Muon

Workplace: Faculty of Information Technology – Cyber Warfare, Telecommunications University, Vietnam.

Email: muon.ha@mail.ru

Education: He received the Specialist degree in Computer Systems, Complexes and Networks from Volgograd State Technical University, Russia (2014), and the Ph.D. degree in Mathematical and Software Support for Computers, Complexes and Computer Networks from Saint Petersburg Electrotechnical University, Russia (2022).

Recent research direction: Data science, artificial intelligence, and cybersecurity – focusing on relational database transformation to NoSQL, and cyberattack detection using intelligent technologies.

Tên tác giả: **Hà Văn Muôn**

Cơ quan công tác: Khoa Công nghệ Thông tin – Tác chiến Không gian mạng, Đại học Thông tin Liên lạc (Trường Sĩ quan Thông tin), Việt Nam.

Email: muon.ha@mail.ru

Quá trình đào tạo: Tốt nghiệp hệ chuyên gia chuyên ngành Máy tính, tổ hợp, hệ thống và mạng tại Đại học Kỹ thuật Quốc gia Volgograd, Liên Bang Nga (2014), và nhận bằng Tiến sĩ chuyên ngành Đảm bảo toán học và tin học cho máy tính, tổ hợp và mạng máy tính tại Đại học Kỹ thuật Điện Saint Petersburg, Liên Bang Nga (2022).

Hướng nghiên cứu hiện nay: Khoa học dữ liệu, trí tuệ nhân tạo và an toàn thông tin – tập trung vào chuyển đổi cơ sở dữ liệu quan hệ sang NoSQL và phát hiện tấn công mạng dựa trên công nghệ thông minh.



Nguyen Thanh Hai

Workplace: Faculty of Information Technology – Cyber Warfare, Telecommunications University, Vietnam.

Email: nguyenthanhhai.k18@gmail.com

Education: Graduated with a Master's degree in Computer Science from the Military Technical Academy in 2008.

Recent research direction: Artificial Intelligence, Information Security, and Digital Transformation Applications in Education and Training

Tên tác giả: **Nguyễn Thanh Hải**

Cơ quan công tác: Khoa Công nghệ Thông tin – Tác chiến Không gian mạng, Đại học Thông tin Liên lạc (Trường Sĩ quan Thông tin), Việt Nam.

Email: nguyenthanhhai.k18@gmail.com

Quá trình đào tạo: Tốt nghiệp thạc sĩ ngành Khoa học máy tính tại Học viên Kỹ thuật Quân sự năm 2008.

Hướng nghiên cứu hiện nay: Trí tuệ nhân tạo, an toàn thông tin, ứng dụng chuyển đổi số trong giáo dục đào tạo.



Nguyen Ngoc Nhan

Workplace: Faculty of Information Technology – Cyber Warfare, Telecommunications University, Vietnam.

Email: nhan.nguyen.97@mail.ru

Education: He received a Bachelor degree in Computer Science and Computing technique from Peter the Great Saint Petersburg Polytechnic University, Russia in 2021.

Recent research direction: System Optimization - focus on error evaluation of computational methods in non-Markovian queues.

Tên tác giả: **Nguyễn Ngọc Nhân**

Cơ quan công tác: Khoa Công nghệ Thông tin – Tác chiến Không gian mạng, Đại học Thông tin Liên lạc (Trường Sĩ quan Thông tin), Việt Nam.

Email: nhan.nguyen.97@outlook.com

Quá trình đào tạo: Đã tốt nghiệp đại học hệ cử nhân chuyên ngành Tin học và kỹ thuật tính toán tại Đại học Bách khoa Saint Petersburg, Liên bang Nga vào năm 2021.

Hướng nghiên cứu hiện nay: An toàn thông tin, trí tuệ nhân tạo, tối ưu hóa hệ thống.



Tran Quang Tuan

Workplace:Faculty of Information Technology – Cyber Warfare, Telecommunications University, Vietnam.

Email: tranquangtuan.637@tcu.edu.vn

Education: Holds a Bachelor's Degree in Information Technology from the University of Telecommunications, VietNam in 2019.

Recent research direction: Artificial Intelligence, Information Security, AI Model Fine-Tuning, System Optimization

Tên tác giả: **Trần Quang Tuấn**

Cơ quan công tác: Khoa Công nghệ Thông tin – Tác chiến Không gian mạng, Đại học Thông tin Liên lạc (Trường Sĩ quan Thông tin)

Email: tranquangtuan.637@tcu.edu.vn

Quá trình đào tạo: Tốt nghiệp cử nhân ngành Công nghệ thông tin tại Trường Đại học Thông tin Liên lạc (Trường Sĩ quan Thông tin), Việt Nam.

Hướng nghiên cứu hiện nay: Trí tuệ nhân tạo, an toàn thông tin, tối ưu hóa hệ thống, tinh chỉnh mô hình AI.