A SYSTEM FOR DETECTING NON-STANDARD VIETNAMESE LANGUAGE ON SOCIAL MEDIA: DEVELOPMENT AND PRACTICAL IMPLEMENTATION

Pham Quoc An[1,*], Bui Khac Khanh[1], Pham Ngoc Dong[1]

DOI: https://doi.org/10.57001/huih5804.2025.415

ABSTRACT

The rapid growth of social media in Vietnam intensified the spread of toxic content, including hate speech, offensive language, and spam, posing threats to online safety. This study developed an automated system for detecting toxic Vietnamese language on social media. We evaluated traditional deep learning models (TextCNN, GRU, LSTM) against pre-trained Vietnamese Transformer models (PhoBERT, BERT4News) using a dataset labeled into four classes: clean, offensive, hate, and spam. The results showed that BERT4News achieved strong performance on imbalanced data, with F1-scores of 95.38% (clean), 58.28% (hate), 66.32% (offensive), and 76.95% (spam). On the combined ViHSD and ViSpam datasets, our BERT4News model reached 90.99% accuracy and 74.23% F1-macro, significantly surpassing the multilingual BERT baseline from the original study by 4.11% in accuracy and 11.54% in F1-macro. However, after balancing with SMOTE, traditional models such as LSTM (56.39% accuracy) outperformed Transformer-based models, suggesting different optimal strategies for balanced and imbalanced data. Finally, the system was deployed as a browser extension and a web-based dashboard, providing practical tools for automated moderation and enhancing digital communication safety in Vietnam.

Keywords: *Deep learning models; Hate speech detection; SMOTE; Transformer-based models; Vietnamese social media.*

[1]Faculty of Information and Communication Technology, CMC University, Vietnam

[*]Email: bit220006@st.cmcu.edu.vn

Received: 15/8/2025

Revised: 15/11/2025

Accepted: 28/11/2025

1. INTRODUCTION

In recent years, social media platforms such as Facebook, TikTok, and YouTube have become integral to daily life in Vietnam, transforming communication and information dissemination. However, this digital freedom has also led to a surge in harmful content [7], including personal attacks, hate speech, discrimination, and online scams. These issues pose a significant societal challenge, particularly as the legal framework for information control is still evolving within the context of national digital transformation.

From a technical standpoint, detecting inappropriate language in Vietnamese is fraught with difficulty. The language is characterized by its rich expressiveness, nuanced semantics, and the prevalent use of slang, regional dialects, sarcasm, and metaphors [14], which complicates natural language processing. Furthermore, the availability of high-quality, labeled datasets for Vietnamese is limited [16], hindering the effective training of machine learning models.

To address these challenges, this study focuses on developing an automated system to detect inappropriate language on social media by leveraging and fine-tuning modern deep learning models for Vietnamese. The primary objective is not only to evaluate various language processing techniques but also to build a practical tool that can assist in content moderation and foster a safer online environment in Vietnam.

The main contributions of this paper are threefold:

- **A Unified Labeled Dataset:** We constructed a comprehensive dataset by merging two existing Vietnamese resources, ViHSD [19] and ViSpamReviews [20]. This combined dataset was standardized into four labels (clean, offensive, hate, spam) and processed using SMOTE to address data imbalance.

- **Comparative Model Evaluation:** We conducted a systematic evaluation of various deep learning architectures, including traditional models (TextCNN, GRU, LSTM) and Transformer-based models pre-trained specifically for Vietnamese (PhoBERT, BERT4News) [2].
- **Practical System Implementation:** A significant contribution is the development of an integrated, real-world application system, featuring a browser extension for end-users and a web dashboard for administrators, demonstrating a viable pathway from research to practical deployment.

2. METHODOLOGY

2.1. Data Collection and Preprocessing

Table 1. Some analysis and statistics of the dataset

free_text	label_id
"tình hình tham nhũng đang ổn định :3" (English: The corruption situation is stable :3)	0
"Trong lúc nước sôi lửa bỏng này mà vẫn có kẻ nhận thức kém vậy" (English: Even in this critical moment, there are still people with poor awareness)	1
*"Em lại xạo l*n" (English: You're lying again, f*ck)*	2
"https://www.facebook.com/..." (English: Link to Facebook...)	3

The foundation of our research is a composite dataset created by merging two public Vietnamese datasets: ViHSD [19], which contains 33,400 social media comments labeled as CLEAN, OFFENSIVE, or HATE, and ViSpam [20], which includes 19,868 e-commerce reviews labeled as spam or legitimate [12]. This merged dataset was then unified under a four-label scheme: *clean, offensive, hate,* and *spam*. Table 1 provides examples from the dataset.

The data underwent a rigorous preprocessing pipeline [1], which included removing noise (special characters, emojis, URLs), normalizing text (lowercase conversion, handling common abbreviations), and tokenizing the text using the Underthesea library [21], which is optimized for Vietnamese word segmentation [21].

The visualization of dataset distribution changes is shown in Figure 1. The corresponding shift in word prominence for each label before and after balancing is shown in Figure 2 and Figure 3. To address class imbalance, we applied the Synthetic Minority Oversampling Technique (SMOTE) combined with undersampling [24] to create a more balanced dataset. The final dataset was split into training (70%), validation (20%), and testing (10%) sets.

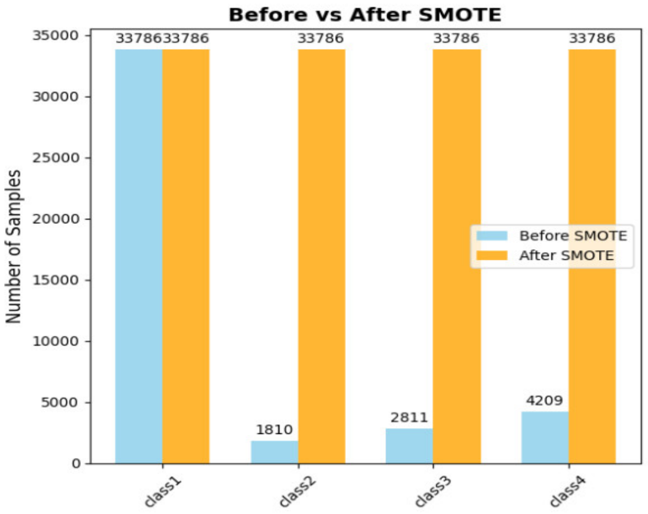


Figure 1. Dataset distribution comparison before and after applying the SMOTE technique for class balancing

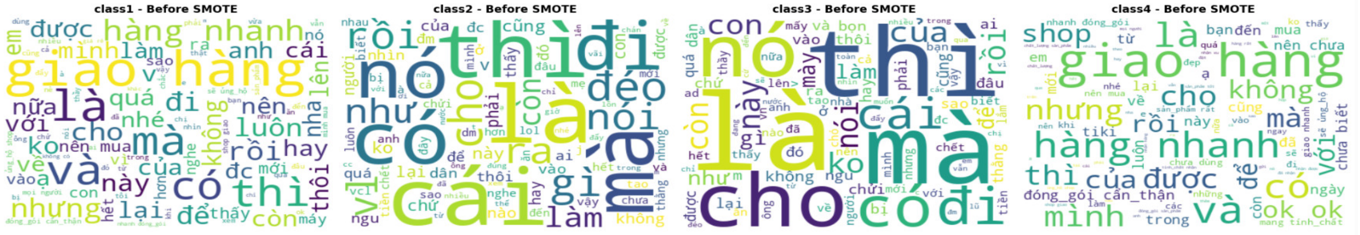


Figure 2. Word clouds of the four label categories before applying the SMOTE technique

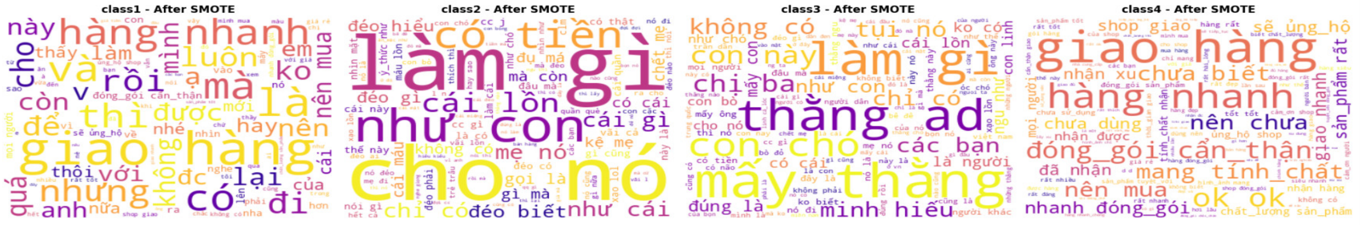


Figure 3. Word clouds of the four label categories after applying the SMOTE technique

2.2. Model Architectures

We evaluated two main classes of models: traditional deep learning architectures and Transformer-based models.

2.2.1. Deep Learning Models

We implemented TextCNN [11], LSTM, and GRU [3] as baselines.

- **LSTM-based Model:** Long Short-Term Memory (LSTM) is a variant of Recurrent Neural Networks (RNN) designed to overcome the vanishing gradient problem using gating mechanisms (forget, input, and output gates) to control information flow, making it effective for capturing long-term dependencies in sequential data.

- **CNN-based Model:** Convolutional Neural Networks (CNNs) are effective for extracting local features (n-grams) from text. Our architecture consists of an embedding layer, a convolutional layer to capture local patterns, a max-pooling layer to extract salient features, and a fully connected layer for classification.

- **GRU-based Model:** A Gated Recurrent Unit (GRU) is a simplified variant of LSTM that combines the forget and input gates into a single update gate, reducing model complexity while maintaining performance in capturing sequential dependencies.

These models were trained for 10 - 15 epochs using the Adam optimizer [22] and categorical cross-entropy loss function.

2.2.2. Transformer-based Models

Transformers [15] utilize self-attention mechanisms to capture relationships between all tokens in a sequence, enabling parallel training and superior context modeling. We employed PhoBERT, based on RoBERTa and tailored for Vietnamese, and BERT4News [7], a variant optimized for Vietnamese news and social media. These models were fine-tuned for 3 - 5 epochs with a batch size of 16, using the AdamW optimizer [23] and a learning rate of 2e-5.

2.3. Evaluation Metrics

In this experiment, the models are trained and evaluated to verify the effectiveness of each approach in addressing the task of detecting harmful content in Vietnamese text. All models are assessed using the same set of standard evaluation metrics [4, 8] to ensure objectivity and comparability across methods.

Model performance was assessed using standard classification metrics: Accuracy, Precision, Recall, and F1-score. These foundational metrics are essential for classification tasks. Given the class imbalance in the dataset, the Weighted F1-score was particularly important. AUC-ROC [25] was also used for analysis (Table 4). The F1-score is calculated as the harmonic mean of Precision and Recall:

$$F1 - score = 2\ x\ \frac{Precision\ x\ Recall}{Precision + Recall} \quad (1)$$

F1-score: The harmonic mean of Precision and Recall.

To evaluate model performance, we utilized common metrics in classification tasks. These metrics are calculated based on values from the confusion matrix, including:

- True Positives (TP): Correctly predicted positive cases.

- True Negatives (TN): Correctly predicted negative cases.

- False Positives (FP): Incorrectly predicted positive cases (Type I error).

- False Negatives (FN): Incorrectly predicted negative cases (Type II error).

Accuracy, Precision, and Recall are defined based on the components of the confusion matrix. Accuracy: The ratio of correct predictions over the total number of samples.

$$Accuracy\ = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

Precision: The proportion of correctly predicted samples for a class out of all samples predicted as that class.

$$Precision\ = \frac{TP}{TP + FP} \quad (3)$$

Recall: The proportion of actual class samples correctly identified by the model.

$$Recall\ = \frac{TP}{TP + FN} \quad (4)$$

MacroF1-score: The unweighted average of F1-scores across all classes.

$$Macro\ F1\ = \frac{F1_{class1} + F1_{class2} + \cdots + F1_{classN}}{N} \quad (5)$$

WeightedF1-score: The average F1-score weighted by the number of samples in each class.

$$Weighted\ F1\ = \frac{\sum_{i=1}^{N} w_i.F1_i}{\sum_{i=1}^{N} w_i} \quad (6)$$

MicroF1-score: Computed using the total TP, FP, and FN across all classes.

$$Micro\ F1\ = \frac{Micro\ Precision\ x\ Micro\ Recall}{Micro\ Precision + Micro\ Recall} \quad (7)$$

3. RESULTS AND DISCUSSION

3.1. Performance on Imbalanced Data (Before SMOTE)

As presented in Table 2, the Transformer-based models significantly outperformed traditional deep learning architectures on the original, imbalanced dataset [18]. BERT4News emerged as the top-performing model, achieving the highest Accuracy (90.99%) and Weighted F1-score (91.36%). Its ability to understand context was particularly evident in its strong performance on the more nuanced hate and spam categories, where it achieved F1-scores of 58.28% and 76.95%, respectively.

Compared to traditional architectures such as CNN, GRU, and LSTM, which reached Weighted F1-scores between 81.19% and 83.46%, while Transformer-based models demonstrated a notable improvement of approximately 7 - 10% in overall performance. PhoBERT also showed strong results (Accuracy 90.42%, Weighted F1 90.09%), confirming the effectiveness of pre-trained language models tailored for Vietnamese.

The superiority of BERT4News over PhoBERT can be attributed to its training on diverse Vietnamese news and social media corpora, which enhanced its generalization across both formal and informal language patterns. These findings indicate that language-specific Transformers not only handle imbalanced distributions effectively but also better differentiate between subtle linguistic expressions of toxicity, such as implicit hate or coded offensive terms.

Furthermore, Table 3 shows that our BERT4News model achieved gains of 4.11% in accuracy and 11.54% in F1-macro compared to the original multilingual BERT baseline [19], underscoring the advantage of using language-specific pretrained models.

Table 3. Comparison of Transformer Models on Combined ViHSD and ViSpam Datasets

Model	Accuracy (%)	F1-macro (%)
Bert-base-multilingual-cased (Original)	86.88	62.69
BERT4News (Ours)	**90.99**	**74.23**

3.2. Performance on Balanced Data (After SMOTE)

The application of SMOTE to create a balanced dataset revealed a significant shift in model performance, as detailed in Table 4. While overall accuracy decreased for all models, this suggests that the high accuracy on the original dataset may have been inflated due to the model's bias towards the majority 'clean' class. A critical insight emerged: under balanced data conditions, the LSTM model became the top performer with an accuracy of 56.39%, surpassing the Transformer-based models. This result suggests that traditional architectures like LSTM may exhibit more robust generalization on artificially balanced, multi-class text classification tasks.

Table 4. Performance of models after SMOTE application

Model	Accuracy (%)	F1-micro (%)	F1-macro (%)	AUC-ROC	Best Threshold	Best F1-macro
CNN	53.29	53.29	42.31	82.44	0.18	50.21
LSTM	**56.39**	**56.39**	**45.08**	82.24	0.83	48.55
GRU	54.45	54.45	42.87	81.56	0.19	47.72
PhoBERT	55.81	55.81	44.66	81.23	0.89	46.53
BERT	53.97	53.97	43.10	79.34	0.89	45.50
BERT4News	42.63	42.63	35.84	77.11	0.89	38.17

3.3. Deployment Feasibility

As shown in Table 5, a critical trade-off exists between accuracy and computational cost. Traditional models are

Table 2. Performance of models before SMOTE application

Model	Data Preprocess	Accuracy (%)	F1 (Clean)	F1 (Offensive)	F1 (Hate)	F1 (Spam)	WeightedF1 (%)
CNN	Raw	80.55	90.88	41.47	42.62	47.67	81.19
GRU	Raw	81.60	89.25	44.80	41.43	54.54	82.26
CNN	Preprocessed	82.56	91.01	43.75	51.77	52.11	81.88
GRU	Preprocessed	81.90	90.12	44.32	45.28	53.45	82.17
LSTM	Preprocessed	82.45	91.03	43.10	49.00	60.00	83.46
PhoBERT	Preprocessed	90.42	95.12	64.90	52.97	72.47	90.09
Bert4News	**Preprocessed**	**90.99**	**95.38**	**66.32**	**58.28**	**76.95**	**91.36**

lightweight and fast, whereas Transformer models require significantly more resources [13]. The superior accuracy of models like BERT4News comes at the cost of larger size and slower inference times, a key consideration for real-world applications.

Table 5. Deployment feasibility analysis

Model	Model Size (MB)	Avg. Inference Time (ms/sample)
TextCNN	35.5	45
GRU	36.5	50
LSTM	38.1	55
PhoBERT	515	120
BERT4News	515	115

4. SYSTEM IMPLEMENTATION

The most significant contribution of this research is the translation of our findings into a tangible, user-centric system. The system is built on a Client-Server architecture and consists of two main components designed for different user roles: end users and administrators.

The overall system architecture, illustrating the integration between the browser extension, backend API, and web dashboard, is shown in Figure 4.

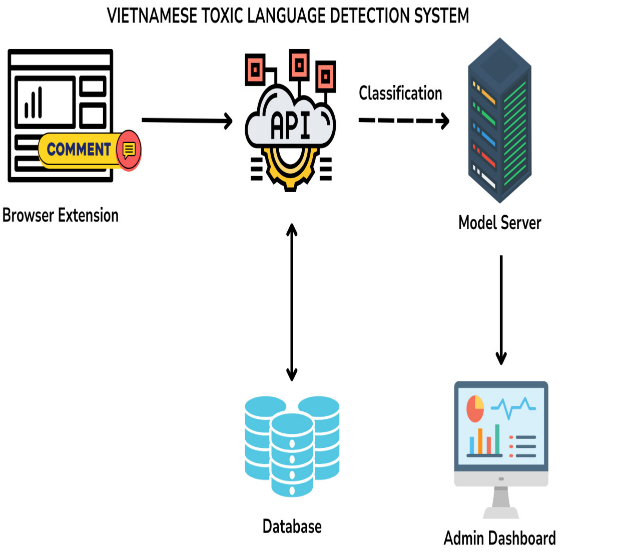
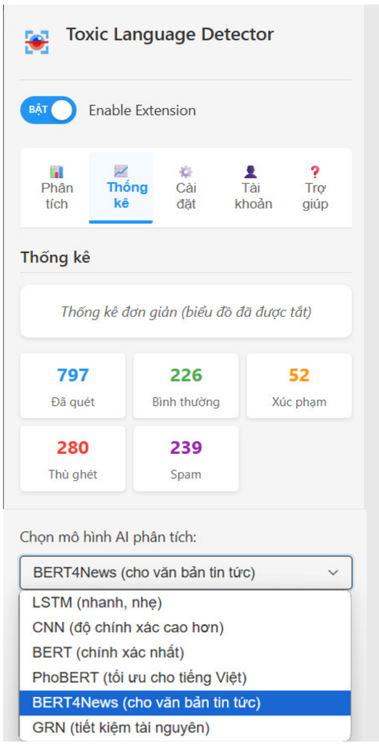


Figure 4. Overall system architecture showing the integration between browser extension, backend API, and web dashboard

- Browser Extension: For end-users, we developed a browser extension that automatically scans comments on social media platforms. The extension sends comment text to our backend API, which hosts the trained BERT4News model. The classification result is then displayed directly on the user's screen, providing immediate feedback. The visual representations of the browser extension and the corresponding classification labels are displayed in Figure 5 and Figure 6.


Toxic Language Detector

BẬT Enable Extension

Phân tích | Thống kê | Cài đặt | Tài khoản | Trợ giúp

Thống kê

Thống kê đơn giản (biểu đồ đã được tắt)

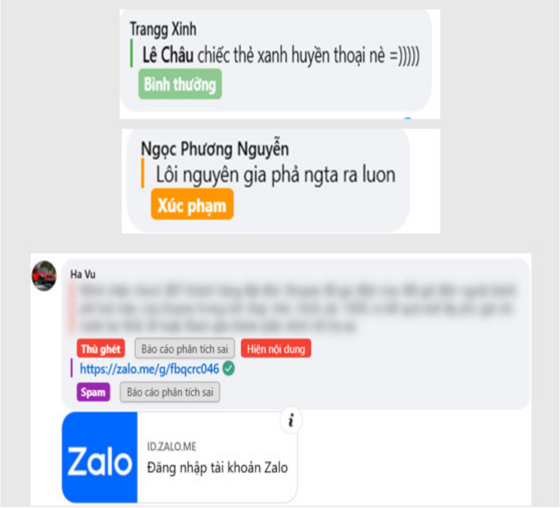
797 Đã quét	226 Bình thường	52 Xúc phạm
280 Thù ghét	239 Spam	

Chọn mô hình AI phân tích:

BERT4News (cho văn bản tin tức)

LSTM (nhanh, nhẹ)
CNN (độ chính xác cao hơn)
BERT (chính xác nhất)
PhoBERT (tối ưu cho tiếng Việt)
BERT4News (cho văn bản tin tức)
GRN (tiết kiệm tài nguyên)

Figure 5. Browser extension


Trangg Xinh
Lê Châu chiếc thẻ xanh huyền thoại nè =)))))
Bình thường

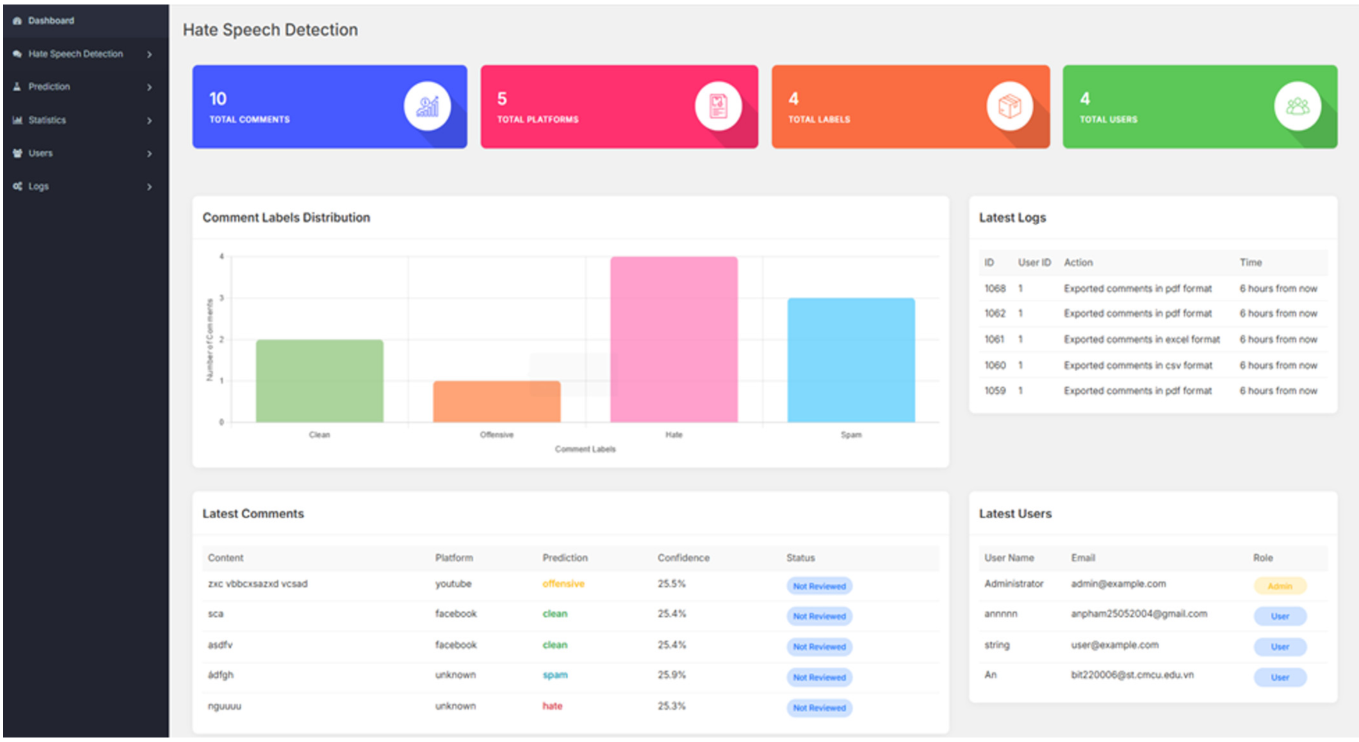
Ngọc Phương Nguyễn
Lôi nguyên gia phả ngta ra luon
Xúc phạm

Ha Vu
[redacted]
Thù ghét | Báo cáo phân tích sai | Hiện nội dung
https://zalo.me/g/fbqcrc046
Spam | Báo cáo phân tích sai

Zalo
ID.ZALO.ME
Đăng nhập tài khoản Zalo

Figure 6. Classification labels

- Web Dashboard: For administrators, we created a comprehensive web dashboard that provides a high-level overview of detected toxic language, with statistics and tools for large-scale content moderation. The backend is powered by Laravel for user management and a separate FastAPI server to serve the AI model via a RESTful API. The interface of the administrative web dashboard, which provides comprehensive analytics and content moderation tools, is presented in Figure 7.



Latest Logs

ID	User ID	Action	Time
1068	1	Exported comments in pdf format	6 hours from now
1062	1	Exported comments in pdf format	6 hours from now
1061	1	Exported comments in excel format	6 hours from now
1060	1	Exported comments in csv format	6 hours from now
1059	1	Exported comments in pdf format	6 hours from now

Latest Comments

Content	Platform	Prediction	Confidence	Status
zxc vbbcxsazxd vcsad	youtube	offensive	25.5%	Not Reviewed
sca	facebook	clean	25.4%	Not Reviewed
asdfv	facebook	clean	25.4%	Not Reviewed
ádfgh	unknown	spam	25.9%	Not Reviewed
nguuuu	unknown	hate	25.3%	Not Reviewed

Latest Users

User Name	Email	Role
Administrator	admin@example.com	Admin
annnnn	anpham25052004@gmail.com	User
string	user@example.com	User
An	bit220006@st.cmcu.edu.vn	User

Figure 7. Administrative web dashboard interface providing comprehensive analytics and content moderation tools

5. CONCLUSION

This research successfully developed and validated a high-performance system for detecting toxic language in Vietnamese social media content. Our results demonstrate that model selection depends critically on data balance: BERT4News achieves state-of-the-art accuracy (90.99%) on naturally imbalanced data, while LSTM models perform better (56.39%) on datasets artificially balanced via SMOTE.

More importantly, we demonstrated the practical applicability of our work by deploying the model in a functional system composed of a user-facing browser extension and an administrative web dashboard. This flexible and scalable architecture has significant potential for integration into various platforms, contributing to a more civilized and secure online environment in Vietnam.

Despite these achievements, we acknowledge certain limitations. Our training data may not fully represent language used in other contexts [9, 12], and the models face challenges with highly ambiguous or metaphorical language [6, 17]. Future work will focus on incorporating more diverse data sources and exploring advanced multilingual Transformer models [5, 10] to enhance accuracy further.

REFERENCES

[1]. C. C. Aggarwal, C. Zhai, *A Survey of Text Classification Algorithms*. Boston, MA: Springer US, 2012.

[2]. P. Badjatiya, S. Gupta, M. Gupta, V. Varma, "Deep learning for hate speech detection in tweets," in *Proc. 26th Int. Conf. World Wide Web Companion*, Republic and Canton of Geneva, CHE, pp. 759-760, 2017.

[3]. K. Cho, et al., "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. 2014 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, pp. 1724-1734, 2014.

[4]. J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, 20, 1, 37-46, 1960.

[5]. A. Conneau, et al., "Unsupervised cross-lingual representation learning at scale," in *Proc. 58th Annual Meeting of the Association for Computational Linguistics*, Online, pp. 8440-8451, 2020.

[6]. T. Davidson, D. Warmsley, M. Macy, I. Weber, "Automated hate speech detection and the problem of offensive language," [Online]. Available: arXiv:1703.04009, 2017.

[7]. J. Devlin, M. W. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Minneapolis, MN, 2019.

[8]. B. Di Eugenio, "On the usage of kappa to evaluate agreement on coding tasks," in *Proc. 2nd Int. Conf. Language Resources and Evaluation (LREC'00)*, Athens, Greece, May 2000.

[9]. P. Fortuna, J. Rocha da Silva, J. Soler-Company, L. Wanner, S. Nunes, "A hierarchically-labeled Portuguese hate speech dataset," in *Proc. 3rd Workshop on Abusive Language Online*, Florence, Italy, pp. 94–104, 2019.

[10]. E. Grave, P. Bojanowski, P. Gupta, A. Joulin, T. Mikolov, "Learning word vectors for 157 languages," in *Proc. 11th Int. Conf. Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, 2018.

[11]. Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. 2014 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014.

[12]. T. N. Nguyen, M. McDonald, T. H. T. Nguyen, B. McCauley, "Gender relations and social media: A grounded theory inquiry of young Vietnamese women's self-presentations on Facebook," *Gender, Technology and Development*, pp. 1-20, 2020.

[13]. V. Sanh, L. Debut, J. Chaumond, T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," [Online]. Available: arXiv:1910.01108, 2020.

[14]. A. Schmidt, M. Wiegand, "A survey on hate speech detection using natural language processing," in *Proc. 5th Int. Workshop on Natural Language Processing for Social Media*, pp. 1-10, 2017.

[15]. A. Vaswani, et al., "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 5998-6008, 2017,

[16]. X. S. Vu, T. Vu, M. V. Tran, T. Le-Cong, H. T. M. Nguyen, "HSD shared task in VLSP campaign 2019: Hate speech detection for social good," in *Proc. VLSP 2019*, 2019.

[17]. Z. Waseem, D. Hovy, "Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter," in *Proc. NAACL Student Research Workshop*, San Diego, CA, pp. 88-93, 2016.

[18]. X. Yang, L. Yang, R. Bi, H. Lin, "A comprehensive verification of transformer in text classification," in *Proc. China National Conf. on Chinese Computational Linguistics*, pp. 207–218, 2019.

[19]. S. T. Luu, K. V. Nguyen, N. L. T. Nguyen, "A Large-scale Dataset for Hate Speech Detection on Vietnamese Social Media Texts," in *Proc. Int. Conf. on Computational Linguistics and Intelligent Text Processing (CICLing)*, 2021. Available: arXiv:2103.11528.

[20]. C. Van Dinh, S. T. Luu, A. G. T. Nguyen, "Detecting spam reviews on Vietnamese e-commerce websites," *arXiv preprint*, arXiv:2207.14636, 2022.

[21]. V. T. Nguyen, et al., "Underthesea: A Vietnamese natural language processing toolkit," in *Proc. VLSP Workshop*, 2018.

[22]. D. P. Kingma, J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint*, arXiv:1412.6980, 2014.

[23]. I. Loshchilov, F. Hutter, "Decoupled weight decay regularization," *arXiv preprint*, arXiv:1711.05101, 2019.

[24]. H. He, E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, 21, 9, 1263-1284, 2009.

[25]. T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, 27, 8, 861-874, 2006.