How Good Is Synthetic Data for Social Media Texts? A Study on Fine-Tuning Low-Resource Language Models for Vietnamese

Luan Thanh Nguyen[1,2]
[1]Faculty of Information Science and Engineering, University of Information Technology, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
luannt@uit.edu.vn

Abstract

Recent advancements in natural language processing (NLP) have demonstrated the remarkable performance of large language models (LLMs). Leveraging these LLMs to generate synthetic data has emerged as a promising solution to address the scarcity of training data for specific tasks, particularly in low-resource languages. However, LLMs often generate overly formal synthetic texts that do not accurately reproduce the informal nature of spoken language and social media texts, resulting in outputs that poorly represent human-generated content online. Furthermore, LLMs may be limited in generating data for tasks involving harmful content. In this research, we introduce LoSo, which utilizes LLMs to generate social media-like texts in low-resource language settings. Our approach aims to bridge the gap between synthetic and authentic human-generated text, making the output more representative of real-world online content. Additionally, we conduct thorough experiments and comparisons focusing on specific characteristics of social media tasks. The materials used in this study will be made available for research purposes[1].

Warning: The study examines actual social media content that could be viewed as offensive and hateful.

1 Introduction

Social media data has gained significant attention in the NLP community due to its unique characteristics and potential applications in areas such as sentiment analysis, hate speech detection, and crisis management (Neri et al., 2012; Balahur, 2013; Zhang et al., 2018). However, the informal and noisy nature of social media text poses challenges for traditional NLP models trained on well-formed text sources (Han, 2014). This has led to a growing interest in developing specialized models and techniques tailored for social media data processing (Farzindar et al., 2015; Stieglitz et al., 2018). The data scarcity problem is amplified for low-resource languages, as large-scale annotation efforts are often hindered by the lack of resources and linguistic expertise (Magueresse et al., 2020; King, 2015; Nguyen et al., 2022). Consequently, many low-resource languages still need to be studied in the social media domain, limiting the development of robust NLP systems for these languages.

The rise of large language models (LLMs) has opened up new avenues for generating synthetic data, potentially alleviating the data shortage. However, these models are primarily pre-trained on formal text sources, such as books and websites, and may need help to capture the nuances and idiosyncrasies of social media language (Myers et al., 2024; Schramowski et al., 2022). As a result, LLM-based approaches for generating human-like textual data still need to improve in mimicking human behavior in expressing feelings and thoughts through texts.

This paper details experiments focused on synthetic data creation, empirically for Vietnamese, a language with limited resources. The key contributions of this work are as follows:

- First, we analyze the characteristics of benchmark datasets in the social media domain. This analysis is crucial for developing systems that can generate realistic, human-like data reflecting actual content on the internet.
- Second, we introduce LoSo, an AI-driven dataset creation system that combines large language models (LLMs) and small language models (SLMs) to generate synthetic social media texts. Our results show that LoSo produces AI-generated datasets comparable to human-annotated ones.
- Third, we conduct in-depth analyses regarding

[1]https://github.com/tarudesu/LoSo

spoken text rate and hate speech percentage in both original and analysis. The obtained results give us an overview of critical factors that contribute to the distinction of social media data.

2 Related Work

In the era of machine learning, data is the critical factor contributing to developing robust and high-performing models (Sun et al., 2017). However, obtaining high-quality labeled data can be challenging, especially for low-resource languages and domains such as social media text. Researchers have explored various approaches for generating synthetic data to deal with this issue.

2.1 Traditional Data Augmentation Approaches

Traditional data augmentation techniques in natural language processing (NLP) involve transforming existing text data through back-translation, token manipulation, and rule-based perturbations (Feng et al., 2021; Wei and Zou, 2019). These techniques can increase the size and diversity of training datasets. However, they often need help capturing the nuances and complexities of social media language, characterized by informal tone, slang, and misspellings.

2.2 Using Small Language Models

An alternative approach involves using small language models (SLMs) to generate labeled data automatically. In this method, an SLM is first fine-tuned on a subset of human-labeled data for a specific task, such as text classification or named entity recognition. The fine-tuned SLM is then used to classify unlabeled text data, effectively generating labeled synthetic data (Chen et al., 2023; Meng et al., 2022). While this approach can be more efficient than manual annotation, it still requires some initial human-labeled data for fine-tuning, and the performance of the SLM may limit the quality of the synthetic data.

2.3 Using Large Language Models

The release of large language models, for example, GPT-3 (Brown et al., 2020) and LLaMA (Touvron et al., 2023), has opened new possibilities for synthetic data generation. LLMs can be used as labelers by fine-tuning them on a small set of labeled data, similar to the SLM approach. However, this can be expensive in computation due to the large size of LLMs.

Alternatively, LLMs can be used as generators to create synthetic text data from scratch (Keskar et al., 2019; Li et al., 2023; Kholodna et al., 2024). This approach leverages the LLM's ability to generate coherent and diverse text samples based on prompts or conditioning. While LLMs have shown impressive text generation capabilities, their outputs may still need to include the distinctive characteristics of social media language when directly applied to this domain, as they are primarily trained on formal text sources.

3 Methodology

The LoSo system consists of two main components: an LLM for generating initial text drafts and an SLM for refining and filtering these drafts to enhance alignment with social media data characteristics. By leveraging the complementary capabilities of these two models, LoSo aims to produce synthetic data that is diverse and reflective of the target domain. The following sections provide a detailed description of the LoSo system, its components, and our evaluation methodology.

3.1 LoSo: An End-to-End Synthetic Data Generation System

LoSo is a specialized end-to-end synthetic data generation system for text-based social media tasks. It comprises two primary components, targeting to generate and label data, culminating in a high-fidelity AI-generated dataset.

3.1.1 LLM-based Generator

The LLM-based Generator is the core of our system, tasked with creating synthetic text tailored to specific domains and labels. By harnessing the capabilities of LLMs, it produces human-like text samples guided by a clearly crafted prompt structure. This structure ensures that the generated text aligns with the target domain, adheres to label criteria, and emulates real-world linguistic diversity.

The proposed prompt structure, depicted in Figure 2, consists of five main components designed to effectively guide a large language model in generating high-quality, domain-specific text data.

1. **Role Assignment**: Defines the model's assumed role or perspective for generating text, ensuring it aligns with the task or domain.

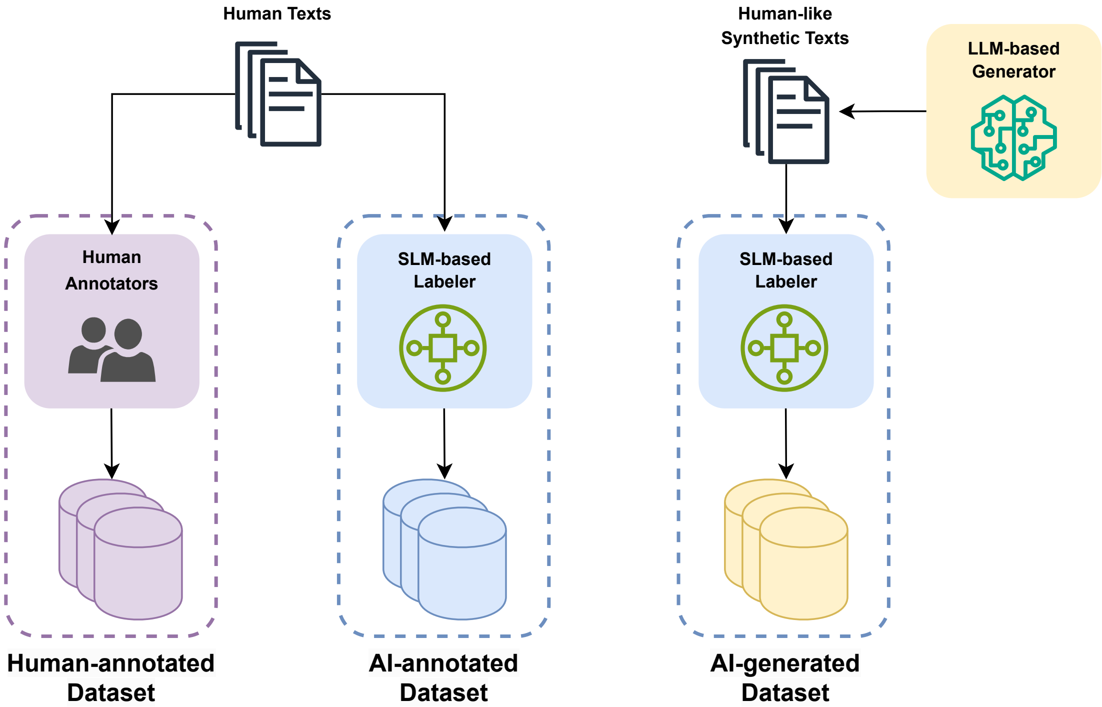

Figure 1: An overview of three data creation approaches.

	Human-annotated Data	SLM-based Classifier	LLM-based Generator
Data	Human Resources	Human Resources	Synthetic Data
Human Costs	High	Low	Low
Compute Costs	Low	Medium - High	High
Time	Long	Medium - Short	Short

Table 1: The comparison of three data creation approaches regarding data source, human costs, compute costs, and time. Note that "time" indicates the time to build a completed dataset for a specific task.

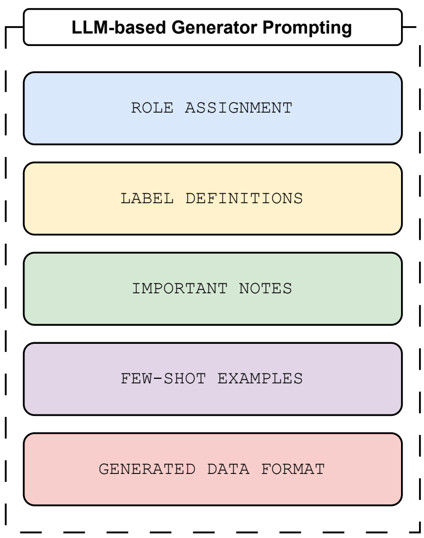

Figure 2: The prompt structure used to generate synthetic human-like texts for each task and label in the LoSo system, which is based on an LLM.

2. **Label Definition**: Clearly outlines criteria defining the target label or category for generated text, crucial for accuracy.

3. **Important Notes**: Provides guidelines and constraints for generating text, ensuring diversity, style, and avoiding biases.

4. **Few-shot Examples**: Representative examples illustrating desired characteristics, helping the model understand patterns and content.

5. **Generated Data Format**: Specifies the required format for presenting generated text data, ensuring consistency and structure.

This decomposed prompt structure equips the LLM with clear guidance, rich context, and well-defined constraints. Consequently, it enables the model to harness its linguistic prowess for generating high-quality, task-specific text data.

3.1.2 SLM-based Labeler

The SLM-based Labeler component in our LoSo system serves as an AI-driven classifier that assigns more accurate labels to the generated data, thereby enhancing the quality and relevance of the synthetic dataset. By leveraging the inherent strengths of SLMs, which are adept at capturing domain-specific nuances and linguistic patterns, we aim to improve the accuracy of labeling while maintaining computational efficiency.

The effectiveness of using an SLM as a classifier lies in its ability to learn from a limited amount of in-domain data. Unlike their larger counterparts, SLMs show great ability in the fine-tuning stage on task-specific datasets, allowing them to develop a focused understanding of the target domain. This specialization enables the SLM-based Labeler to discern subtle differences between classes and assign more precise labels.

3.2 Social Media Text Classification Evaluation Benchmark

To assess LoSo's efficacy, we utilize a comprehensive benchmark comprising three Vietnamese social media datasets. These datasets encapsulate diverse task complexities, label distributions, and linguistic characteristics. The statistics of these datasets are recorded in Table 2.

Sentiment Analysis. The VLSP-SA dataset (Nguyen et al., 2018) evaluates sentiment analysis models for Vietnamese text using user reviews about technological devices. It categorizes 5,100 sentences into positive, neutral, and negative sentiments. These reviews offer concise opinions on specific objects, providing a practical context for sentiment analysis tasks.

Emotion Recognition. The VSMEC (Ho et al., 2020) facilitates emotion recognition in Vietnamese social media text. It features annotated posts categorized into emotions such as joy, sadness, anger, fear, and surprise. This dataset serves as a valuable resource for developing and assessing models to understand and classify emotions expressed in Vietnamese social media content.

Hate Speech Detection. The ViHSD (Luu et al., 2021) dataset focuses on detecting hate speech in Vietnamese social media. It includes annotated comments and posts, identifying offensive language and more severe forms of hate speech directed towards individuals or groups based on attributes like race, gender, or religion. This dataset is essential for creating automated systems that can identify and mitigate hate speech, promoting a safer and more inclusive digital environment.

4 Experiments and Results

In this Section, we conduct multiple experiments to assess the proposed LoSo system's performance in generating social media synthetic texts and serving benchmark classification tasks in Vietnamese. The experiments go through different data conditions and are then evaluated by the performance of the fine-tuned ViSoBERT on these datasets.

4.1 Data

We mainly conduct settings with three primary categories of data, including (1) Original, the top line with data labeled manually by humans and (2) Synthetic, the baseline with data generated and labeled by only an LLM, and (3) The proposed end-to-end synthetic data by LoSo system which leverages LLMs and SLMs in order to generate texts their corresponding labels, respectively. It is worth noting that all types of datasets described below have the same number of samples for each label[2] and each split to ensure the fairness.

Topline With Human-annotated Data. Original datasets from three chosen tasks are used as the topline of this study. As described in Table 1 and in previous studies, they show their effectiveness in solving specific problems but are still costly and time-consuming.

Baseline with Generated Text-Label Data. For the baseline, we use the GPT-3.5-turbo model for generating texts and their corresponding labels for each task. First, we follow the prompt designation (mentioned in Section 3.1.1), aiming to create the exact texts for each label. Then, several minor pre-processing techniques are applied to clean the outputs, including removing unnecessary strings, normalizing labels, and removing users' identities.

End-to-End Synthetic Data Generation. In this approach, we follow the process to create AI-generated Data, depicted in Figure 1 to generate human-like texts by an LLM and re-label them by a specific SLM. The SLM used in our system is ViSoBERT-LoSo, chosen by conducting experiments with multiple pre-trained language models on three selected tasks (mentioned in Appendix C,

[2]The number of samples for each label of data generated by LoSo may be a bit different from the others due to the re-labeling progress.

	VLSP-SA	VSMEC	ViHSD
Task	Sentiment Analysis	Emotion Recognition	Hate Speech Detection
N.o. Labels	3	7	3
Data Source	Users' Reviews	Facebook	Facebook, Youtube
Average Spoken Text Rate	33.94	15.81	51.30
Average Hate Speech Percentage	0.32	13.55	14.67
Average Sequence Length	127.45	55.95	48.92

Table 2: Statistics of three Vietnamese social media benchmark datasets detailing the number of labels, data sources, average spoken text rate (%), hate speech percentage (%), and sequence length (words) across three splits for each dataset.

which outperforms other ones in classification performance. Note that we reuse textual data created from the baseline to adopt this proposed system.

4.2 Model Settings

For the use of LLM, we use the GPT-3.5-Turbo by OpenAI API[3] to generate texts for experiments. For the SLM-based Labeler in the LoSo system, we use several settings and illustrate in detail in Appendix C.

For all main evaluations of data types in three social media tasks, we fine-tune ViSoBERT, one with the settings of 4 epochs, 16 batch size, learning rate 2e-5, and the max sequence length of 128. This study only uses a single NVIDIA A100 GPU for all experiments.

4.3 Evaluation Metrics

In this research, downstream tasks are evaluated with metrics that align with those used in previous studies, namely accuracy score (Acc), weighted F1-score (WF1), and macro F1-score (MF1). MF1 is the primary evaluation metric for each task, as the original research indicates. Furthermore, we determine the Average Macro F1-Score (AF1) by averaging the MF1 scores across three benchmark datasets. This metric reflects the overall performance of each type of training data for the various tasks.

4.4 Experimental Results

Table 3 presents the performance of various data types across three Vietnamese social media text classification tasks. The results demonstrate the effectiveness of our proposed LoSo system in generating high-quality synthetic data for training robust models.

The human-annotated data establishes a strong topline, achieving the highest AF1 of 68.10% across the three tasks. This performance highlights the resource-intensive nature of obtaining such datasets. In contrast, the synthetic data generated solely by the LLM shows a significant performance drop, with an AF1 of 45.07%. This decline is particularly pronounced in the Emotion Recognition and Hate Speech Detection tasks, where the LLM-generated data leads to models with substantially lower accuracy and F1 scores than those trained on human-annotated data.

Remarkably, our proposed LoSo system, which combines LLM-generated texts with SLM-based labeling, significantly narrows the performance gap. The LoSo-generated data achieves an AF1 of 60.48%, a 15.41 percentage point improvement over the LLM-only baseline. This improvement is consistent across all three tasks, with particularly notable gains in Sentiment Analysis and Emotion Recognition.

[3]https://platform.openai.com/

5 Discussion

5.1 How Similar Synthetic Data Is?

The duplicates in synthetic data generation are also a challenging obstacle we need to consider. Thus, we define a Corpus Similarity Score to compute the similarity between each sample pair per each label in the dataset, followed by the Formula 1.

$$\bar{S} = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} S_{ij} \quad (1)$$

Here, $\bar{S}$ denotes the average similarity computed over all unique pairs of sentences. S_{ij} represents the cosine similarity between the embeddings of the i-th and j-th sentences, which is obtained by feeding them into a Sentence Transformer (Reimers and Gurevych, 2019) model. The variable n signifies the total number of sentences in the input list. $\binom{n}{2}$ represents the number of unique pairs that

Data Type	Data Source		Sentiment Analysis			Emotion Recognition			Hate Speech Detection			AF1
	Text	**Label**	**Acc**	**WF1**	**MF1**	**Acc**	**WF1**	**MF1**	**Acc**	**WF1**	**MF1**	
Original	Human	Human	83.79	85.29	65.48	74.95	74.41	74.41	66.23	66.41	64.41	**68.10**
Synthetic	LLM	LLM	65.23	71.36	48.23	52.00	49.97	49.97	38.53	36.07	37.02	45.07
	LLM	SLM	86.39	86.15	63.68	65.05	64.87	64.87	56.71	55.93	52.89	**60.48**

Table 3: Experimental results of multiple training data types, including human-annotated and AI-generated datasets. Note that all these datasets are validated by fine-tuning the ViSoBERT on them, evaluated by accuracy (Acc), weighted F1-score (WF1), macro F1-score (MF1), and average macro F1-score on three tasks (AF1) (%).

can be formed from n items without repetition, ensuring each sentence is compared with all others exactly once.

Following that, we assess the corpus similarity score between the raw texts in the original and those generated by the LLM-based Generator. Here, we use the Vietnamese-SBERT[4] as the Sentence Transformer model to extract text embeddings. Table 4 shows us the overview of the similarity score in three textual data types on each label per each split.

Table 4 shows significant differences in corpus similarity between original and synthetic datasets across three tasks. Synthetic data consistently scores higher, increasing by 14.51 to 27.11 percentage points, indicating the LLM-based Generator produces more homogeneous text within class labels. In emotion recognition, synthetic data averages 46.71% similarity compared to 20.04% for original data, suggesting less diverse emotional expressions. Similar trends are seen in sentiment analysis and hate speech detection. These findings highlight the need for diverse training data and reveal a potential drawback of LLM-based text generation in overfitting specific patterns, urging future research to balance variability and semantic coherence in synthetic data generation.

5.2 Informal Texts in Social Media Data

One of the essential characteristics of social media texts, a challenging model in capturing semantic characteristics, is using informal texts, also known as spoken language form. In this section, we conduct experiments with different data conditions regarding spoken text rate scores.

5.2.1 Spoken Text Rate Score

We define the Spoken Text Rate (STR) score to analyze the proportion of text classified as spoken language. We fine-tune a model to distinguish between spoken and formal Vietnamese using ViSpoChek, detailed in Appendix A. This binary classification task labels texts from ViLexNorm (Nguyen et al., 2024a), combining human-written and normalized versions. The STR score averages these labels across all samples:

$$\text{STR} = \frac{\sum_{i=1}^{n} C(s_i)}{n} \tag{2}$$

where n is the total number of text samples, and $C(s_i)$ is the ViSpoChek Classifier that labels each sample s_i as '0' (non-spoken) or '1' (spoken). Thus, the STR score represents the average rate of samples classified as spoken text.

5.2.2 Data Analysis

Analysis of STR scores across datasets reveals significant differences in language formality, which is crucial for NLP tasks. Figure 3 and Table 5 summarize these differences in original versus synthetic texts.

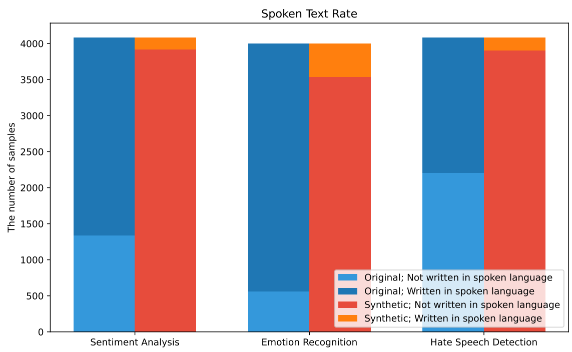

Figure 3: The analysis of spoken text rate in the dataset.

Task/Dataset	Spoken Text Rate	
	Original	**Synthetic**
Sentiment Analysis (VLSP-SA)	32.77	4.04
Emotion Recognition (VSMEC)	14.08	11.58
Hate Speech Detection (ViHSD)	53.97	4.36

Table 5: The spoken text rate for each dataset of each data type across the training set(%).

The task of hate speech detection exhibits the

[4] https://huggingface.co/keepitreal/vietnamese-sbert

Task	Labels	Original		Synthetic	
		Train	Validation	Train	Validation
Sentiment Analysis	NEUTRAL	25.22	25.22	28.24	28.65
	POSITIVE	23.02	21.85	41.46	41.57
	NEGATIVE	25.05	24.24	47.12	45.89
	Average Score	24.43	23.77	38.94	38.70
Emotion Recognition	OTHER	15.04	15.89	25.49	25.07
	DISGUST	20.04	19.26	48.89	49.97
	ENJOYMENT	18.00	17.78	48.40	48.37
	ANGER	25.23	25.23	51.92	51.69
	SADNESS	20.95	20.53	52.81	53.06
	FEAR	22.32	22.10	57.44	58.21
	SURPRISE	18.70	19.90	41.99	44.07
	Average Score	20.04	20.10	46.71	47.21
Hate Speech Detection	CLEAN	14.91	15.37	28.95	29.03
	OFFENSIVE	18.12	18.23	36.75	36.73
	HATE	21.32	21.04	46.27	46.42
	Average Score	18.12	18.21	37.32	37.39

Table 4: The corpus similarity score (%) of three textual data types (lower is better).

Data Type	Data Text	Average STR	Average AF1
Original	Human	33.61	68.10
Synthetic	LLM + ViDenormalizer	57.42	48.82
	LLM	6.66	60.48

Table 6: The comparison between original and synthetic training data with different data forms. The average STR and AF1 scores are calculated by the average of all STR scores (in the training part) and the AF1 scores of each dataset.

highest original spoken text rate (53.97%), reflecting its informal social media origins. However, synthetic data for this task shows a markedly lower rate (4.36%), suggesting challenges in replicating informal language. Similarly, the sentiment analysis task sees a drop from 32.77% (original) to 4.04% (synthetic) in spoken text rate, indicating a shift towards more formal language by the Generator. Meanwhile, the emotion recognition task shows a relatively minor difference (14.08% original compared with 11.58% synthetic), indicating better preservation of informal language style.

5.2.3 Results

Here, we experiment with two main categories, shown in Table 6, to demonstrate how text data form for training affects model performance.

The results in Table 6 demonstrate how text formality impacts model performance across diverse data types. Human-authored data, characterized by an average Spoken Text Rate (STR) of 33.61%, achieves the highest AF1 score at 68.10%, effectively capturing nuances typical of social media discourse. In contrast, synthetic data from the LLM exhibits a low average STR of 6.66% and a reduced AF1 score of 60.48, indicating a bias towards formal language unsuited for social media contexts. Applying the ViDenormalizer to LLM-generated data notably increases STR to 57.42%, surpassing original data informality levels, but this adjustment correlates with a significant AF1 score decline to 48.82%. These findings underscore the challenge of balancing natural language informality with semantic integrity in synthetic data generation for social media analysis, necessitating further exploration of advanced techniques to achieve this balance effectively.

5.3 Hate Speech in Social Media Texts

Besides spoken-language form, toxicity or hate speech in texts is also a crucial characteristic that differentiates social media texts from formal ones. Here, we conduct statistics regarding the hate speech percentage of each dataset in both original and generated texts.

5.3.1 Hate Speech Percentage

First, we use the Hate Speech Percentage (HSP) score, defined in the work of Thanh Nguyen (2024), which refers to how many hateful samples are occupied in the dataset. This progress reveals the

Task/Dataset	Hate Speech Percentage	
	Original	Synthetic
Sentiment Analysis (VLSP-SA)	0.34	5.42
Emotion Recognition (VSMEC)	14.63	13.88
Hate Speech Detection (ViHSD)	45.93	60.61

Table 7: The hate speech percentage for each dataset of each data type across the training set (%).

utilization of a machine learning classifier[5] to detect whether a text is hateful or not. The final score is computed by dividing the number of hateful samples by the number of all data samples.

5.3.2 Data Analysis

We also calculated the HSP score based on the original and the generated texts in this study. Figure 3 and Table 7 demonstrate the achieved analysis.

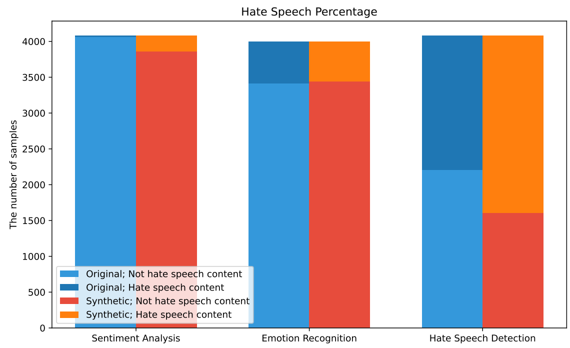

Figure 4: The analysis of hate speech percentage in texts per dataset.

The analysis of hate speech percentages across datasets reveals significant differences between original and synthetic data. Figure 4 and Table 7 illustrate these findings. In the sentiment analysis task, the original data exhibits minimal hate speech (0.34%), whereas synthetic data shows a higher percentage (5.42%). Similarly, in the task of emotion recognition, hate speech percentages are comparable between original (14.63%) and synthetic (13.88%) data, indicating successful replication of emotionally charged language. Most notably, the Hate Speech Detection (ViHSD) dataset displays a substantial increase in hate speech percentage from the original (45.93%) to synthetic (60.61%) data. This suggests the potential amplification of hateful characteristics during data generation, thanks to the well-designed and constrained prompt in generating data.

These findings underscore the importance of considering hate speech prevalence in synthetic data generation, offering insights for refining NLP models to mitigate unintended biases and toxicity.

[5]https://huggingface.co/tarudesu/ViSoBERT-HSD

6 Conclusions

This study introduces LoSo, a potential system for generating synthetic data to enhance social media text classification in Vietnamese, a low-resource language. LoSo combines large language models (LLMs) for text generation and small language models (SLMs) for labeling, effectively mitigating data scarcity while capturing social media language nuances. Experiments on Vietnamese datasets demonstrate that LoSo-generated data achieves performance levels comparable to human-annotated data in sentiment analysis and emotion recognition tasks.

However, the analysis reveals challenges: LLMs tend to produce more formal language than authentic social media text, impacting model performance on real-world data. Moreover, LLMs can inadvertently amplify hate speech when trained on datasets with high hate content. These findings underscore the need for balancing informal language accuracy with semantic fidelity in synthetic data creation, particularly in addressing sensitive issues like hate speech.

Acknowledgement

This research is funded by the University of Information Technology-Vietnam National University HoChiMinh City under grant number D1-2024-58.

We are grateful to the anonymous reviewers for their insightful and constructive comments. Their input has greatly improved the quality and depth of our work.

References

Alexandra Balahur. 2013. Sentiment analysis in social media texts. In *Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 120–128.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. 2023. An empirical survey of data augmentation for limited data learning in nlp. *Transactions of the Association for Computational Linguistics*, 11:191–211.

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

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

Phong Do, Son Tran, Phu Hoang, Kiet Nguyen, and Ngan Nguyen. 2024. VLUE: A new benchmark and multi-task knowledge transfer learning for Vietnamese natural language understanding. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 211–222, Mexico City, Mexico. Association for Computational Linguistics.

Atefeh Farzindar, Diana Inkpen, and Graeme Hirst. 2015. *Natural language processing for social media*. Springer.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

Bo Han. 2014. *Improving the utility of social media with natural language processing*. Ph.D. thesis, University of Melbourne, Department of Computing and Information Systems.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Vong Anh Ho, Duong Huynh-Cong Nguyen, Danh Hoang Nguyen, Linh Thi-Van Pham, Duc-Vu Nguyen, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2020. Emotion recognition for vietnamese social media text. In *Computational Linguistics: 16th International Conference of the Pacific Association for Computational Linguistics, PACLING 2019, Hanoi, Vietnam, October 11–13, 2019, Revised Selected Papers 16*, pages 319–333. Springer.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Nataliia Kholodna, Sahib Julka, Mohammad Khodadadi, Muhammed Nurullah Gumus, and Michael Granitzer. 2024. Llms in the loop: Leveraging large language model annotations for active learning in low-resource languages. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track*, pages 397–412, Cham. Springer Nature Switzerland.

Benjamin Philip King. 2015. *Practical Natural Language Processing for Low-Resource Languages*. Ph.D. thesis.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic data generation with large language models for text classification: Potential and limitations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461, Singapore. Association for Computational Linguistics.

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

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. *Advances in Neural Information Processing Systems*, 35:462–477.

Devon Myers, Rami Mohawesh, Venkata Ishwarya Chellaboina, Anantha Lakshmi Sathvik, Praveen Venkatesh, Yi-Hui Ho, Hanna Henshaw, Muna Alhawawreh, David Berdik, and Yaser Jararweh. 2024. Foundation and large language models: fundamentals, challenges, opportunities, and social impacts. *Cluster Computing*, 27(1):1–26.

Federico Neri, Carlo Aliprandi, Federico Capeci, and Montserrat Cuadros. 2012. Sentiment analysis on social media. In *2012 IEEE/ACM international conference on advances in social networks analysis and mining*, pages 919–926. IEEE.

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

Huyen TM Nguyen, Hung V Nguyen, Quyen T Ngo, Luong X Vu, Vu Mai Tran, Bach X Ngo, and Cuong A Le. 2018. Vlsp shared task: sentiment analysis. *Journal of Computer Science and Cybernetics*, 34(4):295–310.

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

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

Thanh-Nhi Nguyen, Thanh-Phong Le, and Kiet Nguyen. 2024a. ViLexNorm: A lexical normalization corpus for Vietnamese social media text. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1421–1437, St. Julian's, Malta. Association for Computational Linguistics.

Thanh-Nhi Nguyen, Thanh-Phong Le, and Kiet Nguyen. 2024b. ViLexNorm: A lexical normalization corpus for Vietnamese social media text. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1421–1437, St. Julian's, Malta. Association for Computational Linguistics.

Long Phan, Hieu Tran, Hieu Nguyen, and Trieu H. Trinh. 2022. ViT5: Pretrained text-to-text transformer for Vietnamese language generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 136–142, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

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

Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A Rothkopf, and Kristian Kersting. 2022. Large pre-trained language models contain humanlike biases of what is right and wrong to do. *Nature Machine Intelligence*, 4(3):258–268.

Stefan Stieglitz, Milad Mirbabaie, Björn Ross, and Christoph Neuberger. 2018. Social media analytics–challenges in topic discovery, data collection, and data preparation. *International journal of information management*, 39:156–168.

Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. 2017. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Luan Thanh Nguyen. 2024. ViHateT5: Enhancing hate speech detection in Vietnamese with a unified text-to-text transformer model. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5948–5961, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Nguyen Luong Tran, Duong Le, and Dat Quoc Nguyen. 2022. Bartpho: Pre-trained sequence-to-sequence models for vietnamese. In *Interspeech 2022*, pages 1751–1755.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. 2023. Twhin-bert: A socially-enriched pre-trained language model for multilingual tweet representations at twitter. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*, pages 5597–5607.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*, pages 745–760. Springer.

A ViSpoChek: Identifying Vietnamese Spoken-language Texts

A.1 Model Settings

For this evaluation, we select all available BERT-based pre-trained language models supporting the Vietnamese language, including multilingual and monolingual variants. The models were configured with a batch size of 16, a learning rate of 1e-6, four epochs, and a maximum sequence length of 128.

A.2 Results

The achieved results, illustrated in Table 8, show that TwHIN-BERT has the best performance for this task. Thus we choose it as the core model for the ViSpoChek component.

Model	#archs	Acc	WF1	MF1
BERT (multilingual, cased) (Devlin et al., 2019)	base	85.55	85.53	85.53
BERT (multilingual, uncased) (Devlin et al., 2019)	base	82.49	82.40	82.40
DistilBERT (multilingual, cased) (Sanh et al., 2019)	base	78.33	78.32	78.32
XLM-RoBERTa (Conneau and Lample, 2019)	base	84.02	83.95	83.95
XLM-RoBERTa (Conneau and Lample, 2019)	large	74.98	74.96	74.96
DeBERTa_v3 (He et al., 2023)	base	84.98	84.94	84.94
TwHIN-BERT (Zhang et al., 2023)	base	90.38	90.38	90.38
TwHIN-BERT (Zhang et al., 2023)	large	**93.01**	**93.01**	**93.01**
PhoBERT (Nguyen et al., 2020)	base	84.21	84.21	84.21
PhoBERT (Nguyen et al., 2020)	large	82.68	82.63	82.63
PhoBERT_v2 (Nguyen et al., 2020)	base	88.52	88.51	88.51
ViSoBERT (Nguyen et al., 2023)	base	89.47	89.47	89.47
CafeBERT (Do et al., 2024)	base	91.82	91.82	91.82

Table 8: The experimental results of multiple fine-tuned BERT-based models on checking whether a Vietnamese text is written in spoken language form. All models are evaluated by Accuracy (Acc), Weighted F1-score (WF1), and Macro F1-score (MF1) (%).

B ViDenormalizer

To adjust the condition of data based on its textual form, we define ViDenormalizer for de-normalizing Vietnamese texts, respectively. We select multiple sequence-to-sequence pre-trained models and fine-tune them on the dataset ViLexNorm (Nguyen et al., 2024b) in the direction from normalized texts to original texts for ViDenormalizer.

B.1 Model Settings

The experiments are conducted over four epochs with a maximum sequence length of 128. We use the batch size of [16, 8] for BART-based models corresponding to their base and large versions. The learning rate is set at 2e-5. For T5-based models, the batch size is [8, 4] for the base and large models, respectively. We use the learning rate value of 2e-4.

B.2 Evaluation Metric

The task of ViDenormalizer to de-normalize texts is a one-to-many task, which may generate multiple correct outputs, and the BLEU score may not precisely reflect the model performance. Thus, we define the Agreement Rate Score (AR Score), which quantifies the degree of concordance between labels assigned to reference texts and their corresponding generated texts by a classification model. It is formally defined as:

$$\text{AR Score} = \frac{1}{n}\sum_{i=1}^{n} I(L(r_i), L(g_i)) \tag{3}$$

where n is the total number of text pairs, r_i represents the i-th reference text, g_i denotes the i-th generated text, and $L(\cdot)$ is the labeling function of the classification model. The function $I(\cdot, \cdot)$ is an indicator function defined as:

$$I(a,b) = \begin{cases} 1 & \text{if } a = b \\ 0 & \text{if } a \neq b \end{cases} \tag{4}$$

This indicator function yields 1 when its arguments are equal and 0 otherwise. In the context of AR Score, it evaluates to 1 when the labels of the reference and generated texts match and 0 when they differ. Consequently, the AR Score represents the proportion of text pairs for which the model assigns identical labels, providing a measure of label preservation across the reference and generated text sets.

In this study, the classification is the ViSpoChek component, which checks whether a text is written in spoken language.

B.3 Results

Table 9 shows the results in two tasks. It is obvious that ViT5-large is the most effective model and, thus, has been chosen for further experiments in this work.

Models	#archs	ViDenormalizer (AR Score)
mBART-50 (Tang et al., 2020)	large	74.16
mT5 (Xue et al., 2021)	small	62.87
mT5 (Xue et al., 2021)	base	73.68
mT5 (Xue et al., 2021)	large	76.75
BARTpho-syllable (Tran et al., 2022)	base	66.41
BARTpho-word (Tran et al., 2022)	base	63.35
BARTpho-syllable (Tran et al., 2022)	large	56.75
BARTpho-word (Tran et al., 2022)	large	72.25
ViHateT5 (Thanh Nguyen, 2024)	base	77.22
ViT5 (Phan et al., 2022)	base	76.84
ViT5 (Phan et al., 2022)	large	**79.90**

Table 9: The experimental results of multiple fine-tuned sequence-to-sequence models on de-normalizing Vietnamese texts (%).

C BERT-based Model on Social Media Classification Tasks

We use a single BERT-based pre-trained model to evaluate the effectiveness of multiple data types through all experiments. To choose the most optimal, we fine-tune all available BERT-based models on three benchmark tasks in the social media domain. These models include the ones pre-trained on formal texts and the ones on informal texts.

C.1 Model Settings

To fine-tune these BERT-based language models, we configured the experiments with the following settings: 4 epochs, a batch size of 16, a learning rate of 2e-5, and a maximum sequence length of 128.

C.2 Results

Table 10 below shows us the performance of multiple models on three selected tasks. The results show that ViSoBERT outperforms other models in these tasks in terms of the average macro F1 score (AF1).

Model		Offensive Language Identification			Sentiment Analysis			Emotion Recognition			AF1
		Acc	**WF1**	**MF1**	**Acc**	**WF1**	**MF1**	**Acc**	**WF1**	**MF1**	
Formal Text-based SLMs	BERT (multilingual, cased)	86.21	84.23	57.14	62.29	61.81	61.81	49.35	45.72	33.53	50.83
	BERT (multilingual, uncased)	86.24	85.10	59.38	60.57	60.42	60.42	49.06	44.43	31.18	50.33
	DistilBERT (multilingual, cased)	85.96	85.22	60.49	53.05	52.79	52.79	45.45	40.60	27.30	46.86
	XLM-R (base)	86.24	85.42	59.92	71.14	70.99	70.99	53.97	48.10	32.67	54.53
	DeBERTa_v3	85.54	84.31	56.80	62.76	62.62	62.62	41.85	36.18	23.91	47.78
	PhoBERT	86.14	85.47	61.08	68.38	68.25	68.25	51.08	45.52	31.27	53.53
	PhoBERT_v2	87.14	86.63	64.37	73.62	73.47	73.47	54.69	49.15	33.46	57.10
	CafeBERT	88.07	87.24	65.45	76.38	76.13	76.13	66.67	66.55	62.41	68.00
Informal Text-based SLMs	TwHIN-BERT	86.77	85.83	61.81	66.57	66.72	66.72	57.14	52.99	40.08	56.20
	ViSoBERT	88.82	88.47	69.59	74.10	74.07	74.07	67.39	66.87	61.75	**68.47**

Table 10: The comparison of multiple SLMs on three benchmark social media classification tasks (%).

D Data Samples

Task	Generated Text (from LoSo)	Label	Label Characteristics
Sentiment Analysis	Công ty này làm việc từ thứ Hai đến thứ Sáu hay cả tuần vậy nhỉ? (Translated: Is this company working from Monday to Friday or all week?)	NEUTRAL	- Factual statements or observations. - Questions or requests for information. - General comments without emotional bias. - Mild or balanced opinions.
	Wow! Sự hỗ trợ của bạn thật là tuyệt vời, mình cảm thấy vui vẻ quá đi mà 😊🤲 (Translated: Wow! Your support is really amazing, I feel so happy 😊🤲)	POSITIVE	- Expressions of joy, excitement, or gratitude. - Compliments or praise for a person, product, or experience. - Hopeful or optimistic statements. - Encouragement or support.
	Ăn mày à, dịch vụ kém cỏi như thế này thì tao chả bao giờ quay lại đâu 😠 (Translated: You scoundrel, with such poor service like this, I'll never come back 😠)	NEGATIVE	- Expressions of frustration, anger, or sadness. - Complaints or criticism about a product, service, or situation. - Pessimistic or hopeless statements. - Expressions of regret or disappointment.
Emotion Recognition	Ôi dồi ôi, sao đồ ăn trong cái video này trông như cục phân thế kia? Nấu ăn kiểu đó thì ớn quá đi! (Translated: Oh my, why does the food in this video look like that shit Cooking like that is disgusting!)	DISGUST	- Expressions of revulsion, repugnance, or aversion. - Comments about things that are gross, unpleasant, morally reprehensible, or other negative qualities. - Reactions to offensive behaviour, ideas, or substances.
	Zồi ơi, hôm nay được ăn bánh mì thịt nướng ngon tuyệt vời! ☕Ai bảo cuộc sống không có niềm vui, hihi (Translated: Oh my, today I got to eat a delicious grilled pork sandwich! ☕Who says life has no joy, hehe)	ENJOYMENT	- Expressions of pleasure, delight, or satisfaction. - Comments about fun experiences, tasty food, great entertainment, or other enjoyable things. - Reactions to achieving goals or receiving good news.
	Đm, làm ơn đi chỗ khác mà chơi! 😠 Đã gọi giao từ sáng sớm, giờ muốn trưa rồi vẫn chưa thấy nổi một con nhỏ, chán thật! (Translated: Damn, please go somewhere else to play! 😠Called for delivery since early morning, now it's almost noon and still no sign, so frustrating!)	ANGER	- Expressions of rage, fury, or irritation. - Comments about unfair situations, betrayals, disrespect, or other negative experiences. - Reactions to mistakes, delays, or poor service.
	Có phải là tớ đã đủ ngu ngốc để mất cả người mình yêu thương không? 😭Cảm giác lạc lõng và cô đơn quá, không biết phải làm sao... (Translated: Have I been stupid enough to lose the person I love? 😭Feeling lost and lonely, don't know what to do...)	SADNESS	- Expressions of sorrow, grief, or melancholy. - Comments about loss, disappointment, or loneliness. - Reactions to bad news, failures, or missing someone.
	Ôi troi ơi, đi ngang qua khu rừng hoang này thấy tối om, khóe mắt nhìn cứ như có ma vậy, sợ quá đi mất (Translated: Oh my goodness, passing by this deserted forest feels so eerie, corners of my eyes feel like there are ghosts, it's so scary)	FEAR	- Expressions of terror, anxiety, or worry. - Comments about dangerous situations, threats, uncertainties, or other scary things. - Reactions to phobias, dark places, or scary stories.
	Ôi chết, mở hộp quà sinh nhật từ crush ra toàn hàng hiệu, shock quá trời lun á 🎁 (Translated: Oh my god, opened the birthday gift box from my crush and it's all branded stuff, I'm totally shocked 🎁)	SURPRISE	- Expressions of astonishment, shock, or amazement. - Comments about unexpected events, gifts, revelations, or other surprising things. - Reactions to plot twists, magic tricks, or sudden changes.
	Đợt này mình thấy thời tiết Hà Nội ổn hơn hẳn, không nóng quá không lạnh quá. (Translated: This time I find the weather in Hanoi much better, not too hot, not too cold.)	OTHER	- Neutral statements or questions. - Comments without clear emotional content. - General observations or mild opinions.
Hate Speech Detection	Chủ đề này quan trọng lắm, mình muốn biết thêm thông tin về nó nữa! (Translated: This topic is very important, I want to know more information about it!)	CLEAN	- Opinions or emotions expressed respectfully. - Informal language, slang, or internet abbreviations without profanity. - Respectful comments, even in disagreement.
	Mẹ kiếp, cái thời tiết này nóng như con cặc, đéo chịu được! (Translated: Damn, this weather is as hot as hell, can't stand it!)	OFFENSIVE	- General profanity not directed at anyone. - Crude expressions of frustration. - Offensive descriptions of situations. - Vulgar language about non-personal things.
	Mấy thằng lẻn vào quê người ta rồi lại đòi đất, tao cho mày biết đường về trại giam luôn đấy, đập chết mày con đĩ lồn 😠😠🖕 (Translated: Those bastards sneaking into other people's villages and demanding land, I'll show you the way to prison, punch you to death you fucking asshole 😠😠🖕)	HATE	- Harassment and abuse aimed at an individual or group based on characteristics such as religion, nationality, ethnicity, gender, sexuality, or race. - Offensive words attacking a specific target. - Racist, harassing, or hateful content, even if figurative.

Table 11: Some samples generated from our proposed LoSo system.