SEND: A Simple and Efficient Noise Detection Algorithm for Vietnamese Real Estate Posts

Khanh Quoc Tran[1,3,4], **An Tran-Hoai Le**[1,3,4], **An Trong Nguyen**[1,3,4], **Tung Tran Nguyen Doan**[4], **Son Thanh Huynh**[2,3,4], **Bao Le Hoang**[2,3,4], **Hoang Nguyen Minh**[2,3,4], **Triet Minh Thai**[1,3,4], **Hoang Le Huy**[2,3,4], **Dang T. Huynh**[2,3,4], **Binh T. Nguyen**[2,3,4]*, **Nhi Ho**[5], **Trung T. Nguyen**[5]

[1] *University of Information Technology, Ho Chi Minh City, Vietnam*
[2] *University of Science, Ho Chi Minh City, Vietnam*
[3] *Vietnam National University, Ho Chi Minh City, Vietnam*
[4] *AISIA Research Lab, Vietnam*
[5] *Hung Thinh Corporation, Ho Chi Minh City, Vietnam*

Abstract

One of the emerging research fields in Natural Language Processing is Noise Detection (ND), the process of identifying posts containing noise information on textual data. While numerous datasets and approaches are developed for ND research in other languages, equivalent resources for the Vietnamese are limited. To the best of our knowledge, no dataset or method has been investigated or proposed to address the noise Detection tasks in the Vietnamese language. In reality, noise data is constantly present in datasets and sometimes hurts relevant model performance. To overcome this limitation, we propose ViND, a first human-annotated dataset that is available to the scientific community as a benchmark for the task of **Vi**etnamese **N**oise **D**etection. The ViND dataset contains 12,862 posts collected from five major Vietnamese real estate news websites. This paper provides an overview of the Vietnamese Noise Detection task, the process of creating the ViND dataset, and the techniques for carrying out the baseline experiments. On the ViND dataset, the PhoBERT$_{large}$ model outperforms robust baseline models such as LSTM, Bi-LSTM, BERT, RoBERTa, XLM-R, and DistilBERT and achieves a macro F1-score of 0.9024. In addition, our proposed method also successfully improves the related task's performance, mainly Vietnamese Named Entity Recognition (NER) for real estate posts, about 0.0239 in terms of macro F1-score.

1 Introduction

In Natural Language Processing (NLP) and Machine Learning, data and data processing play a significant role, especially when working with user-generated or social network content formed in non-standardized text. Furthermore, data in this aspect generally contains meaningless or useless, and this kind of information often impacts negatively on purpose but is easily ignored. Therefore, to increase the quality of data and the performance of the NLP models, removing all of the noisy information from the dataset (Subramaniam et al., 2009) is necessary.

The explosion of data available from social networking and e-commerce platforms has opened the need and opportunities for noise information processing in NLP (Al Sharou et al., 2021). However, for real estate, the posts could be disturbed by the wrong input from creators, missing essential descriptions, and the mistakes made by real estate agents. Besides, there is a wide range of helpful values with complete information. Nevertheless, their definitions are still ambiguous and complicated, making them difficult to distinguish from the actual noise (Subramaniam et al., 2009). Moreover, Vietnamese real estate post data also includes many challenges such as abbreviations, spelling errors, or some unclear situations that lead to misunderstandings for readers.

In this research, we present a study on building a Vietnamese real estate post dataset and a proposed method for Noise Detection in Vietnamese real estate posts data to improve the efficiency of other vital tasks on this data. Firstly, we collected data from Vietnam real estate from websites. Next, according to annotation guidelines, we annotated the data to noise or non-noise. Finally, we conduct experiments on noise classification methods to compare, analyze and propose a suitable technique. Three main contributions of this paper are summarized as follows:

*Corresponding author: Binh T. Nguyen (e-mail: ngtbinh@hcmus.edu.vn).

1. We present ViND, the first manually-annotated dataset created to serve as a benchmark for the task of Vietnamese Noise Detection. There are 12,862 posts annotated using a strict and efficient process to assure the dataset's quality.

2. Based on the best-performing model PhoBERT$_{large}$, we proposed a simple and efficient method for the Vietnamese Noise Detection task. Various experiments were implemented and evaluated on the ViND dataset using state-of-the-art baseline models. Moreover, we experimented with a combination of Noise Detection and the NER task for Vietnamese real estate posts to verify the effectiveness and contribution of the proposed method.

3. We have analyzed the error cases, limitations, and specific case studies that need to be addressed to improve the models' performance and develop further studies.

The rest of this paper can be organized as follows. First, in Section 2, we survey and describe an overview of the fundamentals of the Noise Detection task and relevant studies. Next, Section 3 shows the process of building our ViND dataset, including three stages: data collection, data annotation, and validation of annotation. Then, Section 4 contains our experiment and analysis on the ViND dataset, which includes the performance of the baseline models and common error cases. Finally, Section 5 provides our main conclusion and future works.

2 Fundamental of Noise Detection

2.1 Task Definition

The starting step for the Noise Detection task is to determine a proper noise definition (Al Sharou et al., 2021). It is worth noting that the meaning of the noise can be different on diverse issues. In this section, we aim to recapitulate the Vietnamese Noise Detection task. The goal of this task is to classify the label **y** (noise or non-noise) corresponding to a provided real-estate advertisement post **X**.

Input: Given Vietnamese real estate posts on the real estate websites.

Output: One of the two labels described below.

1. **Noisy real estate posts (NOISE)** contains noisy data that are frequently intended to cause confusion and can harm the impression of information about a certain real estate. A post is identified as NOISE if it (1) mention many real estates in a single post; (2) does not provide critical information such as an address, price, or area; (3) refers to numerous pricing and places for one real estate.

2. **Non-noise real estate posts (Non-NOISE)** is a normal post. It can be an advertisement, brokerage, buying, or selling of real estate that contains transparent and necessary information without being confusing or restrictive.

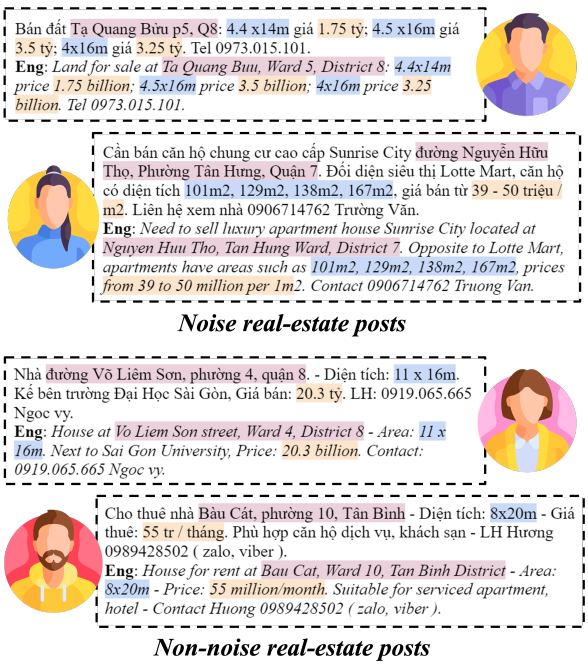

Figure 1: Several instances of the Noise Detection task in Vietnamese.

2.2 Existing Methods for Noise Detection

Subramaniam et al. (Subramaniam et al., 2009) presented a picture of text noise types for documents. The authors surveyed different research topics, including Information Retrieval, Text Classification, Text Summarization, and Information Extraction tasks. The review showed general text noises for different sources of documents from different task aspects. Therefore, there would be many methods for each text noise type of the noise-detection task. Jindal et al. (Jindal et al., 2019) proposed a framework to enable a DNN to learn better sentence representations in the presence of label noise for text classification tasks.

It helped noise models absorb most of the label noise. Bagla et al. (Kumar et al., 2020) conducted experiments by using SOTA methods such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), etc., on IMDB movie reviews datasets (Maas et al., 2011) and Stanford Sentiment Treebank (SST-2) (Socher et al., 2013) for the binary text classification task. The research also pointed out the importance of being mindful of any noise in text data when fine-tuning NLP models on noise text data.

2.3 Noise Detection in Vietnamese Real Estate

Text-noise is observable in most digital texts, including emails, SMS, blogs, and so on (Subramaniam et al., 2009). Furthermore, as real estate post data can be contributed by the community, being affected by noise is unavoidable. Noisy texts will degrade the performance of machine learning models, particularly transformer-based models (Kumar et al., 2020). Despite significant global development in this field (Jindal et al., 2019; Kumar et al., 2020), research in Vietnam is still limited. According to our survey, there is no research on this topic in Vietnam. As a result, we propose implementing Noise Detection into real estate posts. Our work contributes to the first dataset for Vietnamese Noise Detection in real estate posts. It also implements SOTA methods such as deep neural networks and transformer-based models to evaluate the findings.

3 Dataset Creation

Figure 2 depicts an overview of the process we performed to make the ViND dataset. Our dataset creation process goes through three phases: Dataset Collection, Data Annotation, and Validation of Annotation. These phases are described in detail as follows.

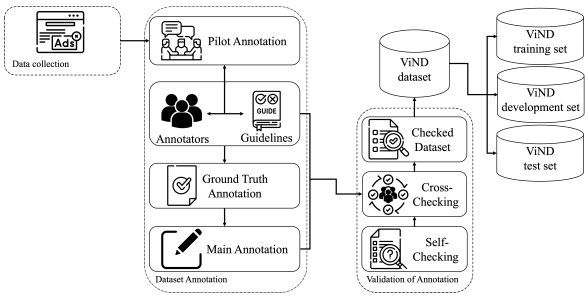

Figure 2: The procedure of creating ViND dataset.

3.1 Data Collection

Our data for this study comes from publicly available online sources, including major real estate websites in Vietnam, including https://nhadat247.com.vn, www.prozy.vn, https://homedy.com, https://muaban.net, and www.batdongsan.com.vn. For data collection, we use the robust Python tool - Beautiful Soup[1]. This library significantly assists us in extracting data from HTML or XML files by providing Pythonic techniques for iterating, searching, or editing directly on the parse tree. After the crawling is completed, an appropriate database is constructed to store the data.

Each collected sample of real estate posts typically includes a post description and any additional information such as an address, area, and price. In some situations, such as when the post description contains undesirable artifacts, erroneous characters, or HTML markers, it is still possible to collect this information. Nonetheless, raw data could occasionally lack critical information from post descriptions. Therefore, combining all relevant features collected from the post description with those already available in raw data could be critical in building an effective data collection strategy from multiple sources. In the end, we collected 12,862 samples to create the ViND dataset.

3.2 Data Annotation Process

Metric For Inter-Annotator Agreement Cohen's Kappa is commonly used to evaluate inter-annotator agreement (IAA) in several tasks and is widely considered as the benchmark (McHugh, 2012). As a result, we employ Cohen's Kappa (Bhowmick et al., 2008) to compute inter-annotator agreements of annotators and quality assurance of human annotation. The Cohen's Kappa coefficient can be formulated as follows:

$$k = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}, \quad (1)$$

where k represents an inter-annotator agreement, $Pr(a)$ represents a relative observed agreement among raters, and $Pr(e)$ represents the hypothetical probability of chance agreement.

3.2.1 Stage 1: Pilot Annotation

We recruited six undergraduate students for our annotation tasks. Most of them had experience

[1] https://pypi.org/project/beautifulsoup4/

annotating several datasets in Vietnamese Natural Language Processing. The primary goal of this Pilot Annotation Stage was to acquaint our annotators with the task. After that, we created an initial set of annotation guidelines with examples and sent them to annotators. Then, before annotating the same 200 random samples from the collected data, all annotators were asked to proofread carefully and rigorously adhere to the annotation guidelines. They repeated these steps five times to compute IAA using Cohen's Kappa, which was obtained by averaging the results of pairwise comparisons among all annotators. Finally, annotators annotated the data independently until the inter-annotator agreement of the annotations achieved more than 0.80 (near perfect agreement) (McHugh, 2012), and they completed the annotation guidelines. Figure 3 presents the IAA of our staff on the tasks of Vietnamese Noise Detection during the Pilot Annotation stage.

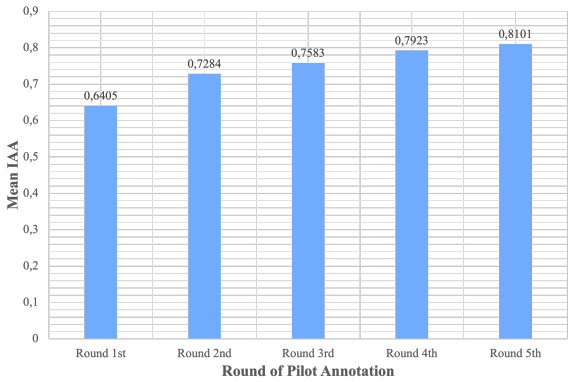

Figure 3: Inter-annotator agreements from five separate annotation training rounds.

Annotation Guidelines A detailed annotation guideline has been explicitly composed for the annotators to identify and label noises efficiently. The proposed guideline includes the following phases: (1) If the annotating post contains sufficient information in the description, it will be labeled as not noise; (2) Any post whose description mentions more than one real estate or contains many values for the same property, such as area or price, can be labeled as noise; (3) There are also posts with either selling or renting types or a combination of these two types that do not provide any information about the address and area or address and price of a specific real estate. Again, it implies that those posts can be labeled "noise" when no helpful information is found.

Sometimes the guidelines do not cover all the diversity of real estate posts and can create challenges that the annotation team may face during the annotation process. Some rare cases where the post contains ambiguous information that is too difficult to identify noise can be set aside for later discussion. After each discussion, the guideline will be modified in detail and complemented to generalize better to new cases in the future.

3.2.2 Stage 2: Ground Truth Annotation

We randomly selected a Ground Truth set of 1,200 posts from the collected data for this stage. Two guideline developers independently annotated the Ground Truth set using the well-developed guidelines from the previous step and reached an IAA of 0.87. These annotators have a deep awareness of the data and tutorials that ensure the reliability and efficacy of the annotation process. In addition, they discussed annotation concerns and solutions for further improvement.

3.2.3 Stage 3: Main Annotation

We divided the collected posts into six equal and non-duplicating subsets. Each well-trained annotation from Phase 1 will be assigned a subset to annotate. In addition, we add 200 samples from the ground truth set to each subset at random and separately. Annotators were asked to modify tasks until the IAA reached 0.80 or higher. Then, the IAA (Cohen's Kappa) was evaluated by comparing each annotator to the corresponding ground truth. This procedure was completed with a mean Cohen's Kappa of 0.83.

3.3 Validation of Annotations

We carefully validated the annotated data before publishing it for research purposes. We required our annotators to self-check the posts they had annotated and prepare short notes to report on their own mistakes after annotating every 500 samples to improve their annotation. This effort decreases the possibility of our annotators making the same mistake too often. To reduce the error rate, we have an additional step of cross-checking once we complete annotating every 3,000 samples. Our staff then investigates and validates any mistakes discovered by others.

3.4 Dataset Statistics

The ViND dataset contains 12,862 posts divided into three subsets: training, development, and test,

with a ratio of 6:2:2. Basic statistics of the three ViND subsets are shown in Table 1[2]. We can see that the length of posts ranges from 8 to 1,038 words, with an average of 110 words utilized in each post. Besides, having a large quantity of training vocabulary allows us to train and fine-tune the models more effectively. Interestingly, posts labeled with noise are, on average, 12.2±1.5 words less in length than non-noise posts. It is because non-noise data contains posts that do not provide essential information, making it shorter.

Table 1: Basic statistics of proposed ViND dataset.

		Training set	Development set	Test set
Full Data	Number of posts	7,717	2,571	2,574
	Average posts length	109.5	111.2	108.3
	Total Vocabulary size	845,632	283,404	278,808
	Maximum posts length	1038	622	622
	Minimum posts length	8	11	8
Noise	Number of posts	1,524	417	420
	Average posts length	98.1	101.2	98.4
	Total Vocabulary size	123,117	42,229	41,337
	Maximum posts length	585	579	505
	Minimum posts length	8	13	8
Non-noise	Number of posts	6,463	2,154	2,154
	Average posts length	111.8	111.9	110.2
	Total Vocabulary size	722,515	241,175	237,471
	Maximum posts length	1038	622	622
	Minimum posts length	8	11	10

Figure 4 depicts the distribution statistics of the number of posts with and without noise labels in each ViND subset. We can observe from the statistical results that the dataset is unbalanced since posts with noise labels account for a considerable proportion of the total. It can be explained by the fact that, in real life, the real estate websites we choose to collect have a team of administrators who filter and eliminate noise posts. However, for objective reasons, a small proportion of such posts still exist and should be deleted from the system.

Figure 4: The labels distribution on each ViND subset.

[2]Note that vocabulary size and comment length are computed at the word level.

4 Experiments

Figure 5 presents an overview of the experimental procedure for our task in this paper.

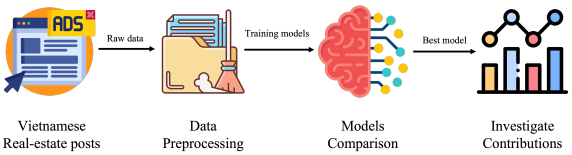

Figure 5: Overview of the experimental procedure for Vietnamese Noise Detection.

4.1 Data Preprocessing

The dataset is processed using the following techniques before it is used in experiments: (1) Normalizing text to Unicode standard; (2) Cleaning input formats (e.g., HTML or Javascript from crawlers); (3) Removing invalid characters (e.g., emojis, non-Vietnamese characters). **Example:** *Nhà đẹp lung linh** $_{*beautiful\ house*}$ → Nhà đẹp lung linh$_{beautiful\ house}$; (4) Fixing non-standard diacritical marks and non-standard punctuations. **Example:** Thanh Hoá → Thanh Hóa, uỷ ban$_{committee}$ → ủy ban$_{committee}$; (5) Using the VnCoreNLP tool (Vu et al., 2018) to do word segmentation.

4.2 Baseline models for Noise Detection performance comparison

We experiment with various approaches to classify noise data, including transformer-based pre-trained language models and deep neural network models. In this study, state-of-the-art models, including LSTM, Bi-LSTM, BERT, RoBERTa, XLM-R, DistilBERT, and PhoBERT, are implemented and fine-tuned to find the best model for the task of Vietnamese Noise Detection.

Long Short Term Memory (LSTM) is a particular type of RNN, which was introduced by Hochreiter et al. (Hochreiter and Schmidhuber, 1997) in 1997. This method has an additional memory cell compared to traditional RNN to capture long-term dependencies information. Moreover, LSTM also presents new gates like input and forget gates to control gradient flow, helping us avoid vanishing gradients when training.

Bidirectional-LSTM (Bi-LSTM) is an extended version of LSTM (Hochreiter and Schmidhuber, 1997) that was proposed by Grave et al. (Schuster and Paliwal, 1997). Unlike standard LSTM, this approach utilizes the information flow from both directions, thus, enhancing the understandability

of the model. BiLSTM can be used in various NLP tasks like machine translation, named entity recognition, and our text classification task.

Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) is a transformer-based approach for pre-trained in various NLP task which was developed by Delvin et al. in 2018. This technique was pre-trained on two tasks: masked language modeling and next sentence prediction. This work uses the training set for fine-tuning the pre-trained BERT model before classifying posts.

Robustly optimized BERT approach (RoBERTa) (Liu et al., 2019) is trained with the help of dynamic masking. This technique forces the system to predict the hidden sections of text truly while unannotated language examples. RoBERTa, implemented in PyTorch, is an extension of the BERT approach (Devlin et al., 2019) with some adjustments in terms of key hyper-parameters like mini-batches size and learning rates. This method also discards the next-sentence pre-trained objective from the original BERT.

XLM-RoBERTa (XLM-R) (Conneau et al., 2020) is a multilingual language model introduced Conneau el at. in 2019. It is a variant of RoBERTa (Liu et al., 2019) which was pre-trained on 2.5T of data across 100 languages containing 137GB of Vietnamese texts. On several cross-lingual benchmarks, this technique outperforms mBERT (Conneau et al., 2020).

DistilBERT (Sanh et al., 2019) is a smaller version of BERT approach (Devlin et al., 2019) which was introduced by Sanh et al. in 2019. Although it just contains 40% fewer parameters than BERT, this method enables it to preserve over 95% of BERT's performance and execute 60% faster. Taking advantage of this efficient method, we utilize DistillBERT in our experiment with the belief of achieving a sustainable result.

PhoBERT (Nguyen and Tuan Nguyen, 2020) is a monolingual pre-trained language model that was trained on a 20GB Vietnamese dataset using the same architecture and approach as RoBERTa (Liu et al., 2019). PhoBERT enables to outperform many state-of-the-art approaches in Vietnamese-specific NLP tasks, including text classification (Nguyen and Tuan Nguyen, 2020; Tran et al., 2022, 2021).

4.3 Experimental Settings

We use the training set to fit experimental parameters and the development set to fine-tune classifier hyper-parameters. We utilize the test set to evaluate our baselines and implement an Adam optimizer (Kingma and Ba, 2015) with a Dropout of 0.2 and a fixed learning rate of 1e-5, and num_train_epochs equal to 4 to fine-tune the hyper-parameters of baseline models. Deep neural network models, including LSTM and Bi-LSTM, are implemented with a 300 embedding size and a Dense two output layer with a Sigmoid activation function. We also pass our input through several well-known word embedding (Tuan Nguyen et al., 2020; Vu Xuan et al., 2019) before feeding it to LSTM and Bi-LSTM model.

In this work, we use simpletransformers[3] to implement all pre-trained language models in transformer-based models. These pre-trained models have a max sequence length equal to 100 and a learning rate decay of 0.01. In addition, we also apply both variants of BERT, RoBERTa, XLM-R, and PhoBERT, including base and large versions.

4.4 Noise Detection Performance Evaluation Metrics

This section presents the evaluation metrics employed in this study. The commonly used metrics for text classification tasks in general (Sokolova and Lapalme, 2009), and detecting noise posts in particular, are Precision, Recall, and F1-score. However, because the proposed datasets have notably imbalanced classes, the average macro F1-score, the harmonic mean of Precision and Recall, is the optimal metric for this task. As a result, we used the average macro F1-score as the critical measure, with the Precision and Recall providing additional information.

4.5 Experimental Results

Table 2 shows our results from the experiments. Compared to deep learning models, the combination between Bi-LSTM and FastText has the highest F1 score of 0.6594 for the ViND test Full Data. Furthermore, the model mentioned above has an F1 score of 0.4119, the highest in the Noises Data category. Besides, combining Bi-LSTM with PhoW2V$_{word}$ achieves the greatest F1-score of 0.9077 in the Non-noises Data category.

[3] https://simpletransformers.ai/

Table 2: Noise Detection results on the ViND test set using various methods. The best outcomes in each category are highlighted.

Model	Full Data			Noises Data			Non-noises Data		
	Pre.	**Rec.**	**F1-score**	**Pre.**	**Rec.**	**F1-score**	**Pre.**	**Rec.**	**F1-score**
LSTM + fastText	0,8489	0,5972	0,6248	0,8333	0,2024	0,3257	0,8645	0,9921	0,9239
LSTM + Word2vec	0,8133	0,6116	0,6429	0,7576	0,2381	0,3623	0,8690	0,9851	0,9234
LSTM + $PhoW2V_{word}$	0,8449	0,6015	0,6307	0,8241	0,2119	0,3371	0,8658	0,9912	0,9242
LSTM + $PhoW2V_{syllable}$	0,8634	0,5946	0,6213	0,8632	0,1952	0,3184	0,8637	**0,9940**	0,9242
Bi-LSTM + fastText	0,6952	0,6404	0,6594	0,5106	0,3452	0,4119	0,8799	0,9355	0,9068
Bi-LSTM + Word2vec	0,6839	0,6383	0,6549	0,4883	0,3476	0,4061	0,8796	0,9290	0,9036
Bi-LSTM + $PhoW2V_{word}$	0,6976	0,6372	0,6573	0,5165	0,3357	0,4069	0,8787	0,9387	0,9077
Bi-LSTM + $PhoW2V_{syllable}$	0,6926	0,6406	0,6590	0,5052	0,3476	0,4118	0,8801	0,9336	0,9061
$BERT_{base}$	0,9172	0,8326	0,8674	0,8938	0,6810	0,7730	0,9406	0,9842	0,9619
$BERT_{large}$	0,8973	0,8701	0,8829	0,8390	0,7690	0,8025	0,9557	0,9712	0,9634
$RoBERTa_{base}$	0,9219	0,8209	0,8608	0,9076	0,6548	0,7607	0,9362	0,9870	0,9609
$RoBERTa_{large}$	0,9016	0,8820	0,8914	0,8430	0,7929	0,8172	0,9601	0,9712	0,9656
$XLM\text{-}R_{base}$	0,9307	0,8313	0,8711	0,9218	0,6738	0,7785	0,9396	0,9889	0,9636
$XLM\text{-}R_{large}$	0,9324	0,8270	0,8686	0,9269	0,6643	0,7739	0,9380	0,9898	0,9632
DistilBERT	0,9153	**0,8894**	0,9016	0,8686	**0,8024**	0,8342	**0,9620**	0,9763	0,9691
$PhoBERT_{base}$	**0,9416**	0,8661	0,8983	**0,9313**	0,7429	0,8265	0,9518	0,9893	0,9702
$PhoBERT_{large}$	0,9312	0,8790	**0,9024**	0,9053	0,7738	**0,8344**	0,9571	0,9842	**0,9705**

With transformer-based models, we get better outcomes in every area when utilizing DistilBERT and PhoBERT. For example, whereas DistilBERT has the two highest Recall scores for Full Data and Noises Data categories (0.8894 and 0.8024, respectively) and the highest value of 0.9620 for Precision of Non-noise Data category, $PhoBERT_{base}$ has the two highest Precision of Full Data and Noise Data with 0.9416 and 0.9313 orderly.

Experimental results indicate that the $PhoBERT_{large}$ model outperforms transformer-based models on Full Data, Noise Data, and Non-noise Data, respectively, by an F1 score of 0.9024, 0.8344, and 0.9705. The PhoBERT model, particularly the $PhoBERT_{large}$ model, has the benefit of being trained on a substantial Vietnamese data domain gathered from various subjects and news websites (Nguyen and Tuan Nguyen, 2020). Due to the diversity of language in the pre-trained and training data, domain knowledge and terminology are better represented, improving model performance.

Furthermore, the traditional technique using LSTM and PhoW2Vec provides an outstanding result in 0.9940 of Recall. Besides, one can see that almost the metric values of transformer-based models are higher than the older ones. On the other hand, the monolingual pre-trained language model for Vietnamese (PhoBERT) beats the multilingual models on the task of Vietnamese Noise Detection. It turns out that existing solutions can solve Noise Detection tasks and generate positive results.

4.6 Error Analysis and Discussion

Based on the macro F1 scores, we select the best baseline models, DistilBERT, $PhoBERT_{base}$, and $PhoBERT_{large}$, to perform error analysis. Then, as shown in Figure 6, we report the statistics of the ratio of various types of error cases[4] of 200 random samples in the ViND development set.

Table 3: Case studies in ViND development set. We evaluate DistilBERT, $PhoBERT_{base}$, and $PhoBERT_{large}$ on 3 sampled posts, with their gold labels and model predictions.

Post	Model			Gold
	DistilBERT	$PhoBERT_{base}$	$PhoBERT_{large}$	
Diện tích từ 100 - 200 - 300m2, có xe đưa đón đi xem Miễn Phí từ TP. HCM. View đồi chè, khí hậu mát mẻ, trong lành. Đường nhựa 10m, hạ tầng điện nước sẵn có. Gần thác Đam ri, tu viện Bát Nhã, Hồ Bảo Lâm. # datnen # datnenbaoloc # baoloc # damri # bds **Eng:** *Area from 100 to 200 and 300m2. Free bus from Ho Chi Minh City. View of tea hill, cool and fresh air. 10m asphalt road, electricity is available. Near Dambri Waterfall, Bat Nha Monastery, Bao Lam Lake. # datnen # datnenbaoloc # baoloc # damri # bds*	Non-noise	Non-noise	**Noise**	**Noise**
Dự án: Bcons Sala. Thông tin chi tiết: Căn hộ Bcons Sala -Trung Tâm Thành Phố Dĩ An 2 PN - 2 WC - 51m2 -Số 1995 Phan Bội Châu, Thành Phố Dĩ An, Bình Dương. Cty CP BDS Phú Mỹ Hiệp (thành viên của BCONS): 3.909 m2: cao 29 tầng : 513 Căn hộ + 11 căn shophouse: 2 PN - 2 WC từ 51m2 - 56m2 **Eng**: *Project: Bcons Sala. Detailed information: Bcons Sala Apartment - Center Di An City. 2 bedrooms - 2 WC - 51m2 – 1995 Phan Boi Chau Street, Di An City, Binh Duong Province. Phu My Hiep - a real estate joint stock company (a member of BCONS): 3909m2: 29 floors: 513 apartments + 11 shophouses: 2 bedrooms - 2 WC with area 51m2 - 56m2*	Non-noise	Non-noise	Non-noise	**Noise**
Hai phòng ngủ thoáng mát, an ninh đảm bảo và tiện ích **Eng:** Two spacious bedrooms, as well as assured security and convenience	**Noise**	**Noise**	Non-noise	**Noise**

We notice that *Information overlap*[5], *Needing syntactic knowledge*[6], and *Short sequence*[7] are

[4]Definition of errors are explained in the Appendix A.

[5]The presence of several overlapping items of information on a given property in the case.

[6]The case contains complicated syntactic structure, and the model is unable to recognize the exact meaning.

[7]The input case is very short (< 20 words).

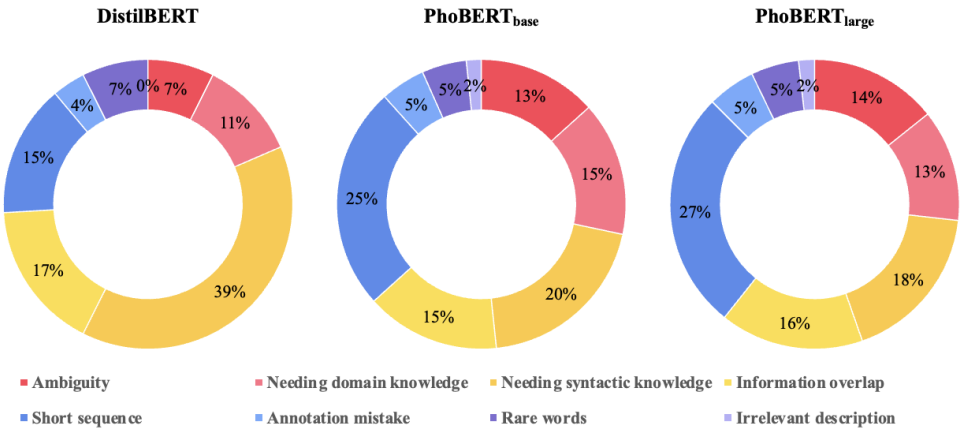

Figure 6: We conduct error analysis on ViND datasets with three best-performing models. We divide error cases into eight categories: Ambiguity, Needing domain knowledge, Needing syntactic knowledge, Information overlap, Short sequence, Annotation mistake, Rare words, and Irrelevant description.

the most common types of failure in DistilBERT, PhoBERT$_{base}$, and PhoBERT$_{large}$. DistilBERT model, in particular, has a high proportion of type Needing domain knowledge errors because it was pre-trained on a limited Vietnamese data domain.

We further show some cases study on ViND development set in Table 3. In the first case, we notice that both DistilBERT and PhoBERT$_{base}$ fail to predict the post's label, while PhoBERT$_{large}$ can obtain the correct prediction. Since this post includes real estate terms and PhoBERT$_{large}$ pre-trained language models that have the advantage of being trained on a larger Vietnamese data domain, one could use domain knowledge to comprehend those terms. In the second case, all three models fail to detect the noise since the post is highly complicated and contains information overlap. In the last point, PhoBERT$_{large}$ incorrectly predicts a short post because the model requires a significant amount of information to identify.

As depicted in Figure 6 and Table 3, overlapping information, lack of information due to limited input data, and complexity in posting structure are challenging issues that need to be addressed in the future. In conclusion, we conclude that the task of Vietnamese Noise Detection is challenging due to the unique peculiarities of the Vietnamese language, and more state-of-the-art methods should be investigated and proposed.

4.7 Contribution Verification For The Proposed Noise Detection Model

We have investigated the proposed method's effectiveness in reducing noise posts in the NER task for Vietnamese real estate posts. In experiments, we utilize our best model, PhoBERT$_{large}$, as a classifier, and the system can eliminate the advertisement posts predicted by the model to be noise. In contrast, the remaining posts can be used to train and evaluate NER models.

To ensure reliability and transparency in the comparison, we conduct the experiment using the same data partitioning, experimental settings, and metrics as Son et al. (Huynh et al., 2021). Furthermore, we decide to compare our approach with Son et al. (2021) as it is the most recent research in the Vietnamese NER task and is in the same field as the real estate news we are addressing.

The experimental results are presented in Tale 4. The results show that our proposed Noise Detection method significantly improves the NER performance (increase up to 0.0239 F1-score). As a result, using PhoBERT$_{large}$ to reduce noise data is efficient and generates state-of-the-art results on the task of NER for Vietnamese real estate.

5 Conclusion and Future Works

This paper presents ViND, a new Vietnamese real estate posts dataset for Noise Detection, including 12,862 samples. In addition, we implement LSTM and BERT models to find a better model that achieves better performance in the Noise Detection task. Finally, from the obtained results, we analyze and record typical error cases that need to be handled to help improve model performance.

As discussed in Section 4.6, we will investigate and test approaches for dealing with imbalanced, overlapping data to improve the solution's performance in the future. Moreover, our research lays

Table 4: The results compare NER models performance without and with using Noise Detection.

Model	Previous study (Huynh et al., 2021)			Our Noise Detection + (Huynh et al., 2021)		
	Precision	Recall	F1-score	Precision	Recall	F1-score
MaxoutWindowEncoder W64	0.8623	0.8933	0.8775	0.8739 (↑0.0116)	0.9224 (↑0.0291)	0.8975 (↑0.0200)
LSTM W64	0.8486	0.8628	0.8556	0.8581 (↑0.0095)	0.8591 (↓0.0037)	0.8586 (↑0.0030)
MishtWindowEncoder W64	0.8677	0.8669	0.8673	0.8753 (↑0.0076)	0.8874 (↑0.0205)	0.8813 (↑0.0140)
BiLSTM W64	0.8573	0.8331	0.8450	0.8547 (↓0.0026)	0.8525 (↑0.0194)	0.8536 (↑0.0086)
MaxoutWindowEncoder W300	0.8739	0.8871	0.8805	0.8783 (↑0.0044)	0.8968 (↑0.0097)	0.8875 (↑0.0070)
LSTM W300	0.8649	0.8869	0.8758	0.8675 (↑0.0026)	0.8946 (↑0.0077)	0.8808 (↑0.0050)
MishWindowEncoder W300	0.8914	0.9237	0.9072	0.9174 (↑0.0260)	0.9452 (↑0.0215)	0.9311 (↑0.0239)
BILSTM W300	0.8524	0.8549	0.8535	0.8725 (↑0.0201)	0.8782 (↑0.0233)	0.8753 (↑0.0218)

the groundwork for various emerging research in Natural Language Processing, such as: (1) Named entity recognition; (2) Text classification; (3) Natural Language Inference.

Limitations and Ethics

We recognize the risk of releasing a dataset for detecting noise texts. For example, because of the subjectivity of manual annotation, our dataset may contain mislabeled data. In addition, due to limits in data coverage and training approaches, our benchmarks cannot detect all types of noise. However, we believe that our proposed benchmark provides more advantages than risks.

All comments in ViND are collected from real estate news websites. This study has ensured user anonymity by eliminating all relevant information when constructing the dataset and rigorously adhering to data source protocol. As a result, the items in our dataset **DO NOT** reflect our opinions or thoughts. ViND is made available to the public for research purposes only.

Acknowledgments

We thank the University of Science, Vietnam National University Ho Chi Minh City, Hung Thinh Corp., and AISIA Research Lab for supporting us throughout this paper. This research is funded by Hung Thinh Corp. under grant number HTHT2021-18-01.

References

Khetam Al Sharou, Zhenhao Li, and Lucia Specia. 2021. Towards a better understanding of noise in natural language processing. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 53–62, Held Online. INCOMA Ltd.

Plaban Kumar Bhowmick, Anupam Basu, and Pabitra Mitra. 2008. An agreement measure for determining inter-annotator reliability of human judgements on affective text. In *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, pages 58–65.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

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.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Son Huynh, Khiem Le, Nhi Dang, Bao Le, Dang Huynh, Binh T. Nguyen, Trung T. Nguyen, and Nhi Y. T. Ho. 2021. Named entity recognition for vietnamese real estate advertisements. In *2021 8th NAFOSTED Conference on Information and Computer Science (NICS)*, pages 23–28.

Ishan Jindal, Daniel Pressel, Brian Lester, and Matthew Nokleby. 2019. An effective label noise model for DNN text classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Ankit Kumar, Piyush Makhija, and Anuj Gupta. 2020. Noisy text data: Achilles' heel of BERT. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 16–21, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Mary L McHugh. 2012. Interrater reliability: The kappa statistic. *Biochemia medica : časopis Hrvatskoga društva medicinskih biokemičara / HDMB*, 22:276–82.

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

M. Schuster and K.K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, 45(4):427–437.

L. Venkata Subramaniam, Shourya Roy, Tanveer A. Faruquie, and Sumit Negi. 2009. A survey of types of text noise and techniques to handle noisy text. In *Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data*, AND '09, page 115–122, New York, NY, USA. Association for Computing Machinery.

Khanh Q. Tran, An T. Nguyen, Phu Gia Hoang, Canh Duc Luu, Trong-Hop Do, and Kiet Van Nguyen. 2022. Vietnamese hate and offensive detection using phobert-cnn and social media streaming data.

Khanh Quoc Tran, Phap Ngoc Trinh, Khoa Nguyen-Anh Tran, An Tran-Hoai Le, Luan Van Ha, and Kiet Van Nguyen. 2021. An empirical investigation of online news classification on an open-domain, large-scale and high-quality dataset in vietnamese. In *New Trends in Intelligent Software Methodologies, Tools and Techniques*, pages 367–379. IOS Press.

Anh Tuan Nguyen, Mai Hoang Dao, and Dat Quoc Nguyen. 2020. A pilot study of text-to-SQL semantic parsing for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4079–4085, Online. Association for Computational Linguistics.

Thanh Vu, Dat Quoc Nguyen, Dai Quoc Nguyen, Mark Dras, and Mark Johnson. 2018. VnCoreNLP: A Vietnamese natural language processing toolkit. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 56–60.

Son Vu Xuan, Thanh Vu, Son Tran, and Lili Jiang. 2019. ETNLP: A visual-aided systematic approach to select pre-trained embeddings for a downstream task. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1285–1294.

A Definition of Error Cases for Error Analysis

We introduce the error definition as follows and illustrate some error cases for Vietnamese Noise Detection tasks in Figure 6:

- **Ambiguity:** the case has the same context but a distinct meaning, which causes the prediction to be incorrect.
- **Needing domain knowledge:** there is real-estate terminology in the case that requires domain knowledge to comprehend.
- **Needing syntactic knowledge:** the case contains complicated syntactic structure, and the model fails to recognize the exact meaning.
- **Information overlap:** the presence of several overlapping items of information on a given property in the case.
- **Short sequence:** the input case is very short (< 20 words).
- **Annotation mistake:** the annotated label is incorrect.
- **Rare words:** the case contains low-frequency terms.
- **Irrelevant description:** the instance has a large amount of irrelevant information, which causes the prediction to be incorrect.