IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 20 June 2022, accepted 13 July 2022, date of publication 1 August 2022, date of current version 25 August 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3195496

APPLIED RESEARCH

An End-to-End Named Entity Recognition Platform for Vietnamese Real Estate Advertisement Posts and Analytical Applications

BINH T. NGUYEN[1,2,3], TUNG TRAN NGUYEN DOAN[3], SON THANH HUYNH[1,2,3], KHANH QUOC TRAN[2,3,4], AN TRONG NGUYEN[2,3,4], AN TRAN-HOAI LE[2,3,4], ANH MINH TRAN[1,2,3], NHI HO[5], TRUNG T. NGUYEN[5], AND DANG T. HUYNH[1,2,3]

[1]Department of Computer Science, Faculty of Mathematics and Computer Science, Vietnam National University Ho Chi Minh City (VNUHCM)—University of Science, Ho Chi Minh City 700000, Vietnam
[2]Vietnam National University Ho Chi Minh City (VNUHCM), Ho Chi Minh City 700000, Vietnam
[3]AISIA Research Laboratory, Ho Chi Minh City 700000, Vietnam
[4]Vietnam National University Ho Chi Minh City (VNUHCM)—University of Information Technology, Ho Chi Minh City 700000, Vietnam
[5]Hung Thinh Corporation, Ho Chi Minh City 700000, Vietnam

Corresponding author: Binh T. Nguyen (ngtbinh@hcmus.edu.vn)

This work was supported by Hung Thinh Corporation under Grant HTHT2021-18-01.

ABSTRACT The volume and complexity of publicly available real estate data have been snowballing. As a result, information extraction and processing have become increasingly challenging and essential for many PropTech (Property Technology) companies worldwide. The challenges are even more pronounced with languages other than English, such as Vietnamese, where few studies in this field have taken place. This paper presents an end-to-end framework for automatically collecting real estate advertisement posts from different data sources, extracting useful information, and storing computed data into proper data warehouses and data marts for the Vietnamese advertisement posts in real estate. After that, one can serve aggregated data for other descriptive and predictive analytics. We combine two models for constructing the most appropriate extraction step: Noise Filtering and Named Entity Recognition (NER). These models can help process initial input data and extract all helpful information. The experiment results show that using PhoBERT$_{large}$ can achieve the best performance compared to other approaches. Furthermore, we can obtain the corresponding F1 scores of the Noise filtering module and the NER module as 0.8697 and 0.8996, respectively. Finally, we utilize Superset for implementing analytic dashboards to visualize the predicted results and serve for further analysis and management processes.

INDEX TERMS Information extraction, information retrieval and text mining, NLP applications.

I. INTRODUCTION

Nowadays, with the development of the internet and communication technologies, it has been much easier to advertise real estate postings compared to the last decades. People can quickly post necessary information about selling, buying, or renting their properties online and give more attention to others. Real-estate postings are currently available from a variety of sources, including real estate websites, as well as other news sources. Therefore, gathering meaningful information from these data sources becomes critical to comprehend the state of real-estate transactions, client interest levels, and rent and sale prices in other places with various sorts of real estate, especially for real estate companies. However, extracting meaningful information fields might be difficult because news feeds are structured and formatted differently, and pieces are written in multiple styles. As a result, large real-estate organizations in different countries have established a system to extract and standardize the source of listing data. They can store the normalized data in the appropriate data marts to analyze, create dashboards for data analytics, and do other predictive analytics.

The associate editor coordinating the review of this manuscript and approving it for publication was Arianna D'Ulizia.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

According to our knowledge, the literature regarding information extraction for Vietnamese real estate data is still not yet developed. In 2012, [1] proposed a rule-based information extraction system, which heavily depended on gazetteers and faced difficulties with diverse writing styles, very long entities, or improper capitalization, as was noted in their error analysis. In 2021, Huynh *et al.* [2] conducted experiments for named entity recognition for Vietnamese real estate data collected from three websites, but they did not correctly address noisy data. Additionally, in comparison with [2], our study includes more experiments with additional methods, a more extensive and diverse dataset, and a platform for analytical applications.

This work presents an end-to-end platform for collecting advertisement data from various sources, standardizing data, parsing useful information, storing extracted data into a data warehouse, and creating the necessary dashboards for further analytics. This project was kickstarted at one of Vietnam's largest real estate companies to provide a practical application for internal and external users. For this purpose, we have different modules in this framework. In the first module, we create relevant data pipelines that can help retrieve the advertisement data from various available sources on the internet. After that, we implement the data standardization for pre-processing collected data, transforming those data into suitable formats, and cleaning up messy data or attributes (including abbreviations, unsigned, misspellings, and duplications). One can later process the data through a noise filtering module developed with PhoBERT (one of the pre-trained language models for Vietnamese) to remove posts with missing information and no value, ensuring high-quality output. In this step, any advertisement post assigned a noisy label is automatically stored in the data warehouse for further action. Finally, the other posts are fed through another module to extract valuable attributes from each one. Finally, we implement a specific Named Entity Recognition (NER) model for collecting all valuable fields from the description text of each advertisement post and combining them with other attributes collected from the data collector for completing the final values of all entities contained. This step can be considered one of our platform's most complicated and essential modules due to the Vietnam languages' grammar complexity and challenges. All algorithms proposed can be integrated into our platform and then connected with analytical dashboards for further exploration from users.

The contribution of our paper can be summarized as follows: (1) First, we develop a Vietnamese dataset containing real estate information in Vietnam with many proper entities, which has a large enough scale and good quality for real estate-related challenges in general and extraction difficulties; (2) Next, we present a practical named entity recognition method for extracting information for real estate items in the Vietnamese language; (3) Finally, we provide several common examples in data that can cause ambiguity and overlap and the difficulty that comes from the descriptive nature of the data and mentions future proposals for extracting these fields of information.

The paper can be organized as follows

1) Section III Data collection and training dataset: In this section, we describe the collection process for our data as well as the necessary steps to create a dataset of sufficient quality. We also discussed the relevant statistics regarding the dataset at the end of this section.
2) Section IV Methodology: This section discusses our proposal to address the challenges mentioned above, which includes constructing two main modules: the noise filtering module and the named-entity module.
3) Section V Experiments: This is where we show the experimental settings as well as results of the proposed modules
4) Section VI A Practical Application Based On The Proposed Platform: This section shows a more practical side of this project, where we discuss the deployment of our system in real-world settings.

II. RELATED WORKS

One of the emerging research trends in natural language processing is NER, extracting information from textual data. The majority of research on this task has been conducted in English [3] due to the number of data sources and the various powerful pre-trained language models. We could specify a few case studies in this section, such as the CoNLL-2003 shared task [4] dataset. In addition to the CoNLL-2003 shared task published dataset, W-NUT [5] is another typical NER dataset we could find. Furthermore, various benchmark datasets for named entity recognition in other languages, including Arabic [6], Chinese [7], German [8] have been published in recent years. Despite significant progress in research on named entity recognition, the related works related to this task for the Vietnamese language, particularly the Vietnamese real estate information extraction, are still modest. For convenience, we will discuss the related works in the following subsections, and each concerns a specific aspect related to our work.

A. NAMED ENTITY RECOGNITION FOR THE VIETNAMESE LANGUAGE

Up to now, there have been only several research efforts related to this work. Tran *et al.* [9] developed a NER model using SVM for seven generic types of entities. Ba and colleagues [10] proposed a rule-based system for extracting several generic entity types. Pham *et al.* [11] presented a semi-supervised training method for conditional random field models. In VLSP 2016 and 2018 NER shared task [12], [13], the authors provided datasets collected from Vietnamese electronics newspapers and tested several NER methods on said datasets for the task of extracting generic information of locations, organizations, and persons. The work [14] also developed from the dataset in [12].

Regarding more specialized types of entities in the Vietnamese language, Truong and his team [15] investigated several methods, including PhoBERT-based [16] for extracting information regarding COVID-19 patients. In [2], Named Entity Recognition in the Vietnamese real estate domain was investigated, albeit on a smaller scale with fewer data in both quantity and variety and without addressing the challenges of low-quality data sources, which we deal with by noise filtering. It is worth noting that, among the mentioned studies, [10] gave a brief example illustrating the monosyllabic nature of the Vietnamese language, its impact on NER, and why one could employ word segmentation for such tasks. Meanwhile, Quan *et al.* [9] discussed the Vietnamese language's essential features and the associated challenges more thoroughly than [10].

B. INFORMATION RETRIEVAL

Pham and Pham [1] presented a rule-based approach for an information extraction system for Vietnamese Real-estate data in the broader context of information retrieval. The authors also gave more details on the data collection and normalization process. Finally, Hong and co-workers [17] were dedicated to the subject of text normalization for tweets in the Vietnamese language in the context of named entity recognition.

C. NOISE FILTERING

It is essential to note that what can be defined as ''noise'' may depend on a specific method and/or use case. For example, Huang *et al.* [18] determined noisy samples as records in the training dataset with mislabeled NER labels. On the other hand, in [19], the authors used distant supervision to automate the generation of NER labels, which would undoubtedly give records with mislabeled or insufficiently labeled (missing tags); such records can be classified as ''noise''. In our work, we asked the experts in the PropTech domain for correct labeling. They recommended labeling records as either ''noisy'' or ''not-noisy'' depending on certain conditions on quality (typos, ambiguous entities, etc.) and usefulness (having critical information such as an address, area, price, etc.).

D. DATA VISUALIZATION

While there are many well developed tools for data analytics and visualization tasks such as Superset,[1] Tableau,[2] Microsoft Power BI,[3] Looker,[4] SAP Analytics,[5] Qlik Sense,[6] Sisense,[7] Domo,[8] we chose Superset simply because it is open-source and it fits the technical expertise of our team.

III. DATA COLLECTION AND TRAINING DATASETS

A. DATA SOURCES

In this study, we obtained relevant datasets from publicly accessible sources on the Internet, namely from popular real estate listing websites. Such popular websites in Vietnam include www.batdongsan.com.vn,[9] nhadat247.com.vn,[10] www.prozy.vn,[11] homedy.com,[12] and muaban.net.[13] We stored all crawled data in appropriate databases. It is important to note that each website has a different and unique format. Therefore, a specific crawler should be specified for each website and require manual updates whenever the corresponding format changes to avoid potential bugs and missing data from the crawlers.

Each collected record of real estate post typically includes a post description and maybe other meaningful attributes, such as an address, area, and price. One can get post descriptions with decent reliability, including unwanted artifacts, invalid characters, or HTML markings. However, raw data sometimes miss some essential attributes from post descriptions. Therefore, smoothly combining all valuable factors extracted from the post descriptions and the existing ones from raw data can help build an efficient data collection process from different sources.

B. PRACTICAL CHALLENGES OF COLLECTED DATA

Due to datasets collected from different websites having various formats and qualities, it is essential to filter the data in terms of usability. Therefore, one of the main goals for building the platform is to centralize all advertisement data and create descriptive and predictive analytics platforms for further usage.

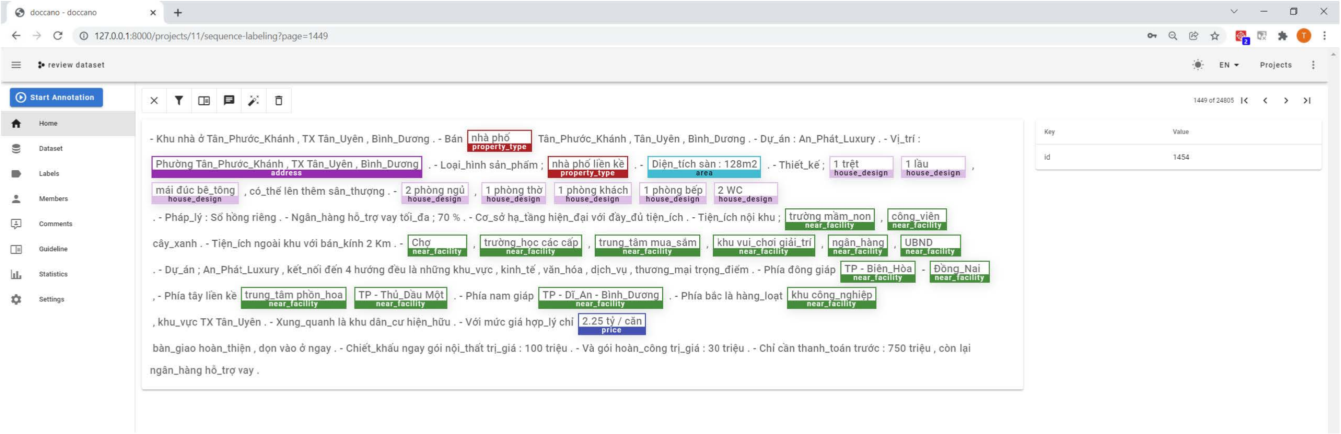
It is worth noting that most real estate websites in Vietnam usually require users to add important information or attributes when creating a new advertisement post for a given property. However, they sometimes forget to provide relevant information or even make short description text having fewer data. In addition, as they can freely write, it can easily create some grammar mistakes or typos. For instance, people might describe the property in some advertisements without information about how much it costs or how much area it has. As a result, these things create main challenges for the task we have to deal with.

With those challenges in mind and discussion with our real estate partners, we defined the following criteria for clean data (''not noisy'') based on quality and usefulness:

1) Not having too many typos and ambiguous entities.
2) Possibility to identify where the property is, i.e., the address information of the property.
3) Enough information regarding the price or the property area, preferably both.

[1] https://superset.apache.org/
[2] https://www.tableau.com/
[3] https://powerbi.microsoft.com/en-au/
[4] https://www.looker.com/
[5] https://www.sap.com/products/cloud-analytics.html
[6] https://www.qlik.com/us/products/qlik-sense
[7] https://www.sisense.com/
[8] https://www.domo.com/

[9] www.batdongsan.com.vn
[10] nhadat247.com.vn
[11] www.prozy.vn
[12] homedy.com
[13] muaban.net



Key	Value
id	1454

- Khu nhà ở Tân_Phước_Khánh , TX Tân_Uyên , Bình_Dương . - Bán nhà phố Tân_Phước_Khánh , Tân_Uyên , Bình_Dương . - Dự_án : An_Phát_Luxury . - Vị_trí :
Phường Tân_Phước_Khánh , TX Tân_Uyên , Bình_Dương . - Loại_hình sản_phẩm ; nhà phố liền kề . - Diện_tích sàn : 128m2 . - Thiết_kế ; 1 trệt 1 lầu ,
mái đúc bê_tông , có_thể lên thêm sân_thượng . - 2 phòng ngủ , 1 phòng thờ 1 phòng khách 1 phòng bếp 2 WC
. - Pháp_lý : Sổ hồng riêng . - Ngân_hàng hỗ_trợ vay tối_đa ; 70 % . - Cơ_sở hạ_tầng hiện_đại với đầy_đủ tiện_ích . - Tiện_ích nội khu ; trường mầm_non , công_viên
cây_xanh . - Tiện_ích ngoài khu với bán_kính 2 Km . - Chợ , trường_học các cấp , trung_tâm mua_sắm , khu vui_chơi giải_trí , ngân_hàng , UBND
. - Dự_án ; An_Phát_Luxury , kết_nối đến 4 hướng đều là những khu_vực , kinh_tế , văn_hóa , dịch_vụ , thương_mại trọng_điểm . - Phía đông giáp TP - Biên_Hòa - Đồng_Nai
, - Phía tây liền kề trung_tâm phồn_hoa TP - Thủ_Dầu Một . - Phía nam giáp TP - Dĩ_An - Bình_Dương . - Phía bắc là hàng_loạt khu công_nghiệp
, khu_vực TX Tân_Uyên . - Xung_quanh là khu dân_cư hiện_hữu . - Với mức giá hợp_lý chỉ 2.25 tỷ / căn
bàn_giao hoàn_thiện , dọn vào ở ngay . - Chiết_khấu ngay gói nội_thất trị_giá : 100 triệu . - Và gói hoàn_công trị_giá : 30 triệu . - Chỉ cần thanh_toán trước : 750 triệu , còn lại
ngân_hàng hỗ_trợ vay .

FIGURE 1. We use Doccano as a tool to annotate named entities for our data in this work. The predicted entities from our proposed model are labeled with distinct colors and corrected by the annotators. The data annotation team also notes mistakes in the collaborative labeling spreadsheet in proper entries.

4) Each post should only describe one unique property, as having multiple properties in one post would cause significant difficulties in extracting relevant information, according to the experts in the PropTech domain (who advised us). For example, it is not very feasible to derive useful information from a post about multiple apartments and townhouses of different prices and areas (which price and area belong to a specific property ?), according to our knowledge.

The records that do not meet such criteria would be classified as noisy data in our study. Those criteria ensure the platform can collect enough information about different real estate properties posted in various cities in Vietnam and help filter out noisy data.

C. ANNOTATION PROCESS

For the data annotation, we have an annotation team consisting of several members who take responsibility for the data processing and labeling different entities in each advertisement post to create the relevant datasets. We employ a data annotation platform using Doccano,[14] of which we have a screenshot in Fig 1. It is worth noting that various processing steps were needed to ensure compatibility between annotated data and the model training pipeline.

D. ANNOTATION STANDARDS

To ensure consistent annotation throughout the whole dataset with multiple people working on data annotation, we prepare a guideline describing how the entities of each type should be labeled and resolving ambiguity. For instance, "split-level" (*lệch tầng*) should be labeled as `property_type` or `house_design`). As the actual guideline has a lot of details and is rather cumbersome, we only give the briefly summarized definitions for the entities to be addressed in our pipeline in Table 8. We defined this guideline with experts' knowledge of the Vietnamese real estate domain. The core annotators also had multiple training sessions on the subject.

E. DATASET CONSTRUCTION

We annotated a dataset having 24,695 post descriptions to prepare the training data for noise filtering and Named Entity Recognition (NER) models. We can summarize our data annotation process through the following steps:

1) DATA PREPARATION

Firstly, the data with empty descriptions are not considered for annotation, while the remaining samples have their descriptions preprocessed through the following procedures.

- Normalizing text to Unicode standard.
- Cleaning input formats (e.g., Html or Javascript from crawlers).
- Removing invalid characters (e.g., emojis, non-Vietnamese characters).
- Fixing non-standard placement of tonal marks and non-standard punctuations.
- Using the VnCoreNLP tool to do word segmentation [20]. We chose VnCoreNLP since it is one of the tools that achieve state-of-the-art results in the Vietnamese word splitting task with an F1 score of 97.90%.
- Other text-processing operations include removing unnecessary whitespace characters, removing invalid segmentation in some special cases where the segmentation "_" signs do not reflect the correct real estate terminology, and other words, characters, separation marks, and operation marks.

2) DATA ANNOTATION

The preprocessed descriptions then go through a pretrained NER model for entities predictions and pass to the annotation team to check and correct the predicted entities and classify the post as noise or not noise based on specific criteria defined

[14] https://doccano.github.io/doccano/

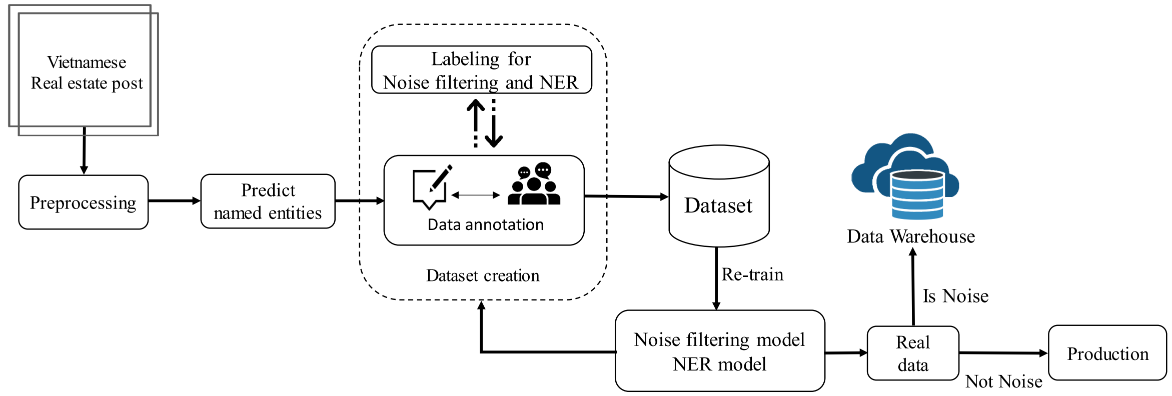


FIGURE 2. The data annotation–models retraining process plays an important role in our pipeline to ensure the quality of both our dataset and models. Therefore, this process is repeated until the performance metrics measured with the models and their respective datasets reach a plateau.

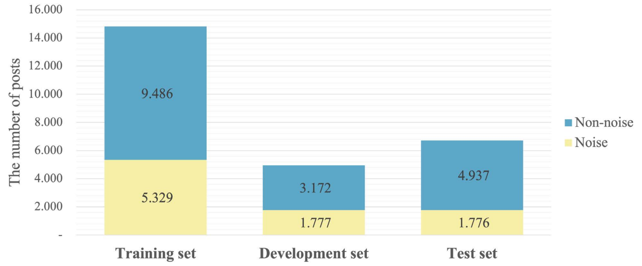


FIGURE 3. The proportions of noisy and not-noisy records in the training set, development set and test set.

in III-B. All cases that do not work reliably can be noted and inspected later. Finally, a benchmarking or golden set is independently evaluated by two specialists to be proportionally injected into each team member's assigned records to serve as a quality control measure, and as noted in Section III-D, these annotators had training sessions with our real estate partner.Other details on how we carried out this task were discussed in Section III-C.

3) MODEL RETRAINING

The newly annotated data is then used to retrain the models. Data augmentation might be employed for some corner cases noted during data annotation in this step. The retrained models are then evaluated and become the new models if old models are found. This data annotation - models retraining process is repeated until the performance no longer improves noticeably, and then the dataset and models are finalized. We illustrate the role of this step and its interaction with other parts in our pipelines in Fig. 2.

F. GENERAL STATISTICS FOR COLLECTED DATA

After data annotation step, we obtained a dataset with 15,813/24,695 ($\approx$ 64%) not-noisy records and 8,882/24,695 ($\approx$ 36%) noise records. This dataset was then used to train the noise filtering model. As for the NER model, we clean the not-noisy data by removing duplicate records and posts not strictly classified as either for-sale or for-rent. After this

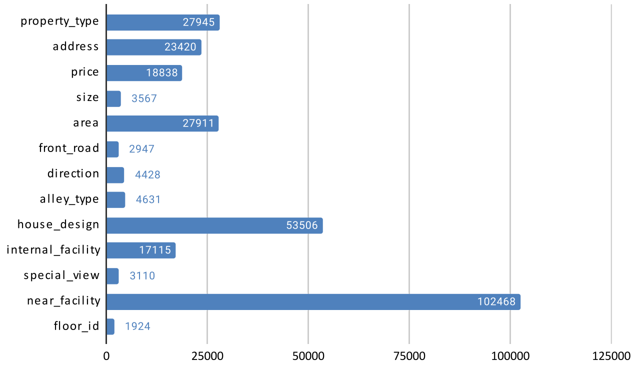


FIGURE 4. The occurrence statistics of named entities in the full dataset with 24,695 samples of which 14,400 samples are selected to train the NER model.

TABLE 1. The statistics for entity types of the dataset with 14,400 samples out of a total of 24,695 annotated samples, including the numbers of entities of each type for the train, development and test sets for the NER model.

Entity Type	Train	Dev.	Test	All
property_type	11,806	3,997	3,947	19,750
address	10,184	3,372	3,395	16,951
price	6,358	2,137	2,167	10,662
size	1,544	485	480	2,509
area	10,615	3,631	3,610	17,856
front_road	1,254	425	445	2,124
direction	1,854	615	638	3,107
alley_type	2,578	836	800	4,214
house_design	24,635	8,111	8,037	40,783
internal_facility	7,770	2,405	2,104	12,279
special_view	877	269	324	1,470
near_facility	45,994	15,655	14,936	76,585
floor_id	982	325	322	1,629
# Entities in total	137,648	45,882	44,907	22,8437
# Posts in total	8,640	2,880	2,880	14,400

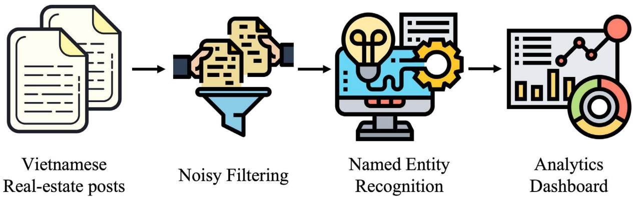


FIGURE 5. Our proposed approach for information extraction of Vietnamese real-estate posts.

step, only 14,400 samples among 24,695 (≈ 58.3%) initial records are then used to train the NER model. The distribution of noise posts in the working dataset can be depicted in Fig. 3. Besides, the statistics for entity types in the full dataset can be summarized by Fig. 4.

The detailed statistics for the sizes of chosen training, development, and test sets for the NER model are given in Table 1.

IV. METHODOLOGY

A. OUR PROPOSED SYSTEM

This section proposes an efficient and straightforward system for Vietnamese information extraction tasks. We focus on the transformer-based model PhoBERT to develop a best-performance model by fine-tuning techniques. Fig. 5 shows the overview of the system using two essential modules: the Noise filtering (Section IV-B), the primary task Named entity recognition (Section IV-C), and the Analytics dashboards using Superset (Section VI). We automate all relevant data pipelines using Apache Airflow[15] and store all extracted and aggregated data on PostgresDB[16] tables.

B. NOISE FILTERING

As discussed in Section III-B, ensuring data quality before feeding to the NER module is one of our top priorities. Fig. 3 indicates that the noisy records make up a significant amount of the sample in the working dataset. The ratio of noisy records on the three subsets, including training: development: test, is 35.97%, 35.91%, and 26.46%, respectively. Through experiments, we found that in addition to pre-processing the data, removing low-quality (noise) posts also plays an essential role in improving the performance of our proposed system (as mentioned in Section V-C).

We investigate several algorithms, including transformers-based pre-trained language models and deep neural network models to classify noisy records. In this study, state-of-the-art models such as Text-CNN [21], Bi-LSTM [22], BERT$_{cased}$ [23], BERT$_{uncased}$ [23], XLM-R$_{base}$ [24], XLM-R-Vietnamese$_{base}$ [25], PhoBERT base [16], and PhoBERT$_{large}$ [16] are deployed and fine-tuned to choose the most appropriate model for our noise filtering task.

We strive to improve the performance of the noise filtering module as it will directly impact our primary NER task. Therefore, we also evaluated the performance of the models we conducted using the precision, recall, and F1-macro metrics. Table 2 presents the experimental results obtained with the noise filtering module. Since the given dataset has a significantly imbalanced noise ratio, the average macro F1-score, the harmonic mean of precision and recall, is the most suitable measure for this task. The results show that PhoBERT$_{large}$ is the best performing model with a 0.8697 F1 score. Furthermore, PhoBERT$_{large}$ could execute parallel processing on words, minimizing vanishing gradients and assisting the model in learning more effectively.

TABLE 2. Evaluation results of noise filtering module on the test set.

Model	Precision	Recall	F1-score
Text-CNN + $fastText$	0.8197	0.8265	0.7878
Text-CNN + $PhoW2V_{word}$	0.8168	0.8265	0.7900
Text-CNN + $PhoW2V_{syllable}$	0.8174	0.8265	0.7896
Bi-LSTM + $fastText$	0.7848	0.8050	0.7899
Bi-LSTM + $PhoW2V_{word}$	0.7905	0.8105	0.7946
Bi-LSTM + $PhoW2V_{syllable}$	0.7864	0.8035	0.7919
BERT$_{cased}$	0.8648	0.8586	0.8615
BERT$_{uncased}$	0.8138	0.8069	0.8101
XLM-R$_{base}$	0.8458	0.8595	0.8526
XLM-R-Vietnamese$_{base}$	0.8706	0.8526	0.8603
PhoBERT$_{base}$	**0.8769**	0.8473	0.8591
PhoBERT$_{large}$	0.8694	**0.8701**	**0.8697**

Procedure 1 A Transformer-Based Method Using a Pre-Trained PhoBERT$_{large}$ Model for Noise Filtering

```
Input: Data of t real-estate posts $T_i$ (0 < i < t+1).
Output: Each record being flagged as either not-noisy
data or noise in the approriate database table.
Procedure: A Noise Filtering module based on a
pre-trained PhoBERT$_{large}$ model.
for each post i in T do
    if $T_i$ is predicted as noise then
        $T_i$ -> Flagged as noisy record
    else
        $T_i$ -> Flagged as not-noisy record
    end
end
end procedure
```

By employing PhoBERT$_{large}$, the best-performing model we found in this study, we construct a simple and efficient procedure for the noise filtering task, presented in Procedure 1.

C. NER-REAP: OUR PROPOSED NAMED ENTITY RECOGNITION MODEL FOR REAL ESTATE ADVERTISEMENT POSTS

This research develops a NER task for Vietnamese real estate advertisement posts with the IOB format (short for inside, outside, beginning). We investigate this task through various experiments on our annotated dataset: (i) the capabilities of traditional models including MishWindowEncoderW300 [2],

[15] https://airflow.apache.org/
[16] https://www.postgresql.org/

BiLSTM-CRF [26] (ii) the effectiveness of the pre-trained language models such as the variations of Transformers [27].

1) TRADITIONAL MODELS

In this section, we conduct experiments with two different models. the first model is MishWindowEncoderW300, this is the model is proposed by spacy.[17] In experiments, we set up the same hyperparameters proposed by Huynh *et al.* [2]. The second model is BiLSTM-CRF [26] which is the standard technique for the NER task, which was first released in 2018 by Rrubaa and Aravindh.

2) TRANSFORMERS MODEL

According to our knowledge, few studies on NER for the Vietnamese language use transformers structures. Therefore, in this work, we utilize the variations of transformers such as Bert [23], XLM-Roberta [24], and PhoBert [16] to enhance the performance effectively in terms of precision, recall, and F1-score.

- BERT model (Bidirectional Encoder Representations from Transformers) [23] was released in 2019 and has become one of the state-of-the-art models in NLP. This paper employs Bert$_{base}$ in two cases, cased and uncased.
- XLM-Roberta (XLM-R) [24] is the alternative for non-English NLP released in November 2019 by the Facebook AI team. Because of the large size of XLM-R$_{large}$, in this experiment, we only study with XLM-R$_{base}$ and XLM-R$_{base-Vietnamese}$ [25] which is the variant of XLM-R$_{base}$ trained only Vietnamese dataset in 2021.
- PhoBERT [16] is a monolingual variant of RoBERTa trained on a 20GB word-level Vietnamese dataset and has gained a lot of superior results in many Vietnamese natural language tasks. In our experiment, we use two variants of PhoBERT are PhoBERT$_{base}$ and PhoBERT$_{large}$.

In our experiment, we use the three models above as feature extractors to understand the context of real estate advertisements. After that, we use Softmax to classify the entity type of one word.

V. EXPERIMENTS

Before going through experimental results, we first go through the evaluation metrics used in this paper. For measuring the performance of different models, we use the macro-average precision, recall, and F1-score (%) to measure the performance of our models.

A. EXPERIMENTAL SETTINGS

1) NOISE FILTERING SETTINGS

In this approach, we use the various transformer-based pre-trained language models from HuggingFace.[18] The pre-trained language models are initialized with a max sequence length is 60. Furthermore, deep neural network models are implemented with several pre-trained word embeddings [28], [29] and a max sequence length of 40. These models have an Adam optimizer, the learning rate is 2e-5, epsilon is 1e-8, and dropout is 0.4.

TABLE 3. The experimental settings of our proposed the transformer-based models including BERT, XLM-RoBERTa, and PhoBERT.

Hyper-parameter	Value
beta$_1$	0.9
00 beta$_2$	0.999
L$_2$	0.01
epsilon	0.00000001

2) NAMED ENTITY RECOGNITION SETTINGS

In the NER model, to conduct the experiment and evaluate models, we divided our dataset into three subsets, including train: development: test, with the corresponding ratio of 6:2:2. We also use pre-trained word embedding FastText[19] with 300 dims to implement the BiLSMT-CRF model. All of the transformer-based models (such as RoBERTa, and XLM-RoBERTa) are fine-tuned using a batch size of 256, a learning rate of 5×10^{-4}, Adam optimizer, and trained with 500 epochs. One can see the Table 3 for more information about the hyper-parameters of Adam's optimization. Finally, we use cross-entropy as a loss function for updating the weights. One can see 3 for more information about the setting of the transformer-based models.

B. EXPERIMENTAL RESULTS

This experiment uses an NVIDIA Tesla P100 GPU to investigate the results using current state-of-the-art algorithms such as BiLSTM-CRF, Bert-base-uncased, Bert-Base-Cased, XLM-Robert-base. Especially two algorithms dedicated to the Vietnamese language are PhoBERT$_{base}$ and PhoBERT$_{large}$. In addition, to compare with the results from previous research, we also ran the best model in [2], MishWindowEncoder W300.

We can see that MishWindowEncoder W300 still gives over 85% results in the precision, recall, and F1-score metrics. This result shows that using MishWindowEncoder to combine TransitionParser in Spacy still provides a good result for the NER task in real estate.

The famous model for the NER task is Bi-LSTM CRF [26] of which the three metrics precision, recall, and F1-score are 83.94%, 78.87%, and 81.32% respectively.

In general, the results of transformer models are higher than MishWindowEncoder W300 and Bi-LSTM CRF, except for the two Bert-based models. BERT was trained on Wikipedia (2.5B words) and BookCorpus (800M words) instead of training on a particular Vietnamese dataset. Therefore, Bert's use as a pre-trained model might not perform

[17] https://spacy.io/

[18] HuggingFace Transformers - https://github.com/huggingface/transformers

[19] https://fasttext.cc/

TABLE 4. Evaluation results of the NER models on the test set.

Model	Precision	Recall	F1-score
Baseline Model [2]	0.8802	0.8534	0.8666
Bi-LSTM CRF	0.8394	0.7887	0.8132
BERT$_{cased}$	0.7639	0.8574	0.8079
BERT$_{uncased}$	0.8683	0.8838	0.8759
XLM-R$_{base}$	0.8907	0.9037	0.8971
XLM-R-Vietnamese$_{base}$	**0.8989**	0.8928	0.8959
PhoBERT$_{base}$	0.8935	**0.9057**	**0.8996**
PhoBERT$_{large}$	0.8982	0.9009	0.8995
LUKE$_{base}$	0.8200	0.8675	0.8431
vELECTRA$_{base}$	0.8703	0.8950	0.8825

TABLE 5. Evaluation results of the model using PhoBERT$_{base}$ on the test set of each entity type.

Model	Precision	Recall	F1-score
property_type	82.42	85.93	84.14
address	83.21	80.73	81.95
size	85.33	90.78	87.97
front_road	74.08	69.85	71.91
near_facility	89.90	90.74	90.32
price	96.01	95.74	95.87
realtor	94.66	97.19	95.91
phone	98.09	98.83	98.46
area	93.53	94.56	94.04
house_design	93.90	96.00	94.94
alley_type	80.77	85.69	83.16
direction	95.04	93.39	94.21
special_view	80.85	81.25	81.05
internal_facility	90.19	89.15	89.67

highly on non-English data compared to other language-specific pre-trained models. Next, XLM-R-Vietnamese$_{base}$ outperform XLM-R$_{base}$ in the precision metric. Because of specific training on the Vietnamese dataset instead of many languages, the pre-trained model XLM-R-Vietnamese$_{base}$ can perform better than XLM-R$_{base}$.

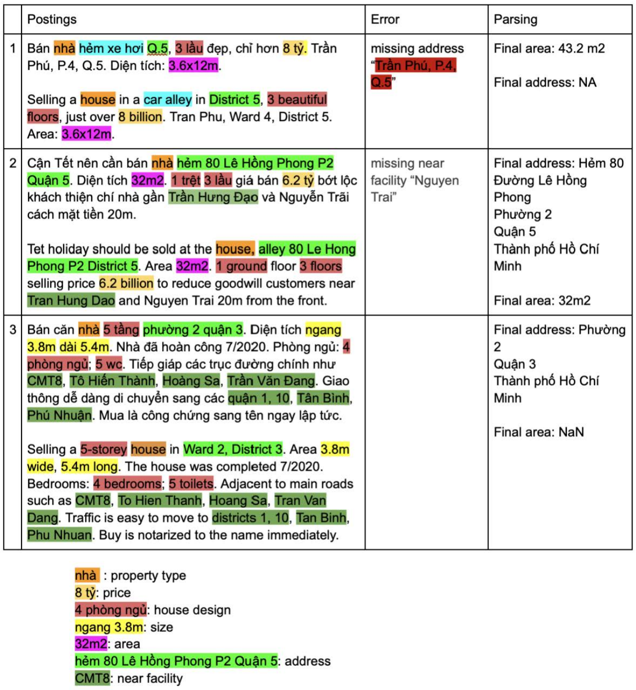
Interestingly, the results for the two models, PhoBERT$_{base}$ and PhoBERT$_{large}$, are almost identical. However, at the same time, the latter is significantly heavier than the former, suggesting that one may have reached some limit using PhoBERT. From this, we conclude that using PhoBERT$_{base}$ is the most optimal for performance and inference speed. One can see more detail regarding the overall metrics for the models in this study in Table 4 while Table 5 shows the metrics of PhoBERT$_{base}$ for each type of entity.

C. ABLATION ANALYSIS

We perform an ablation analysis on the proposed system to demonstrate the efficacy and alignment of the modules.

TABLE 6. Ablation analysis of our proposed system: demonstrating the impact of the noise filtering module within our proposed system.

Method	F1-score
NER-REAP	0.7614
Noise Filtering + NER-REAP	**0.8995**



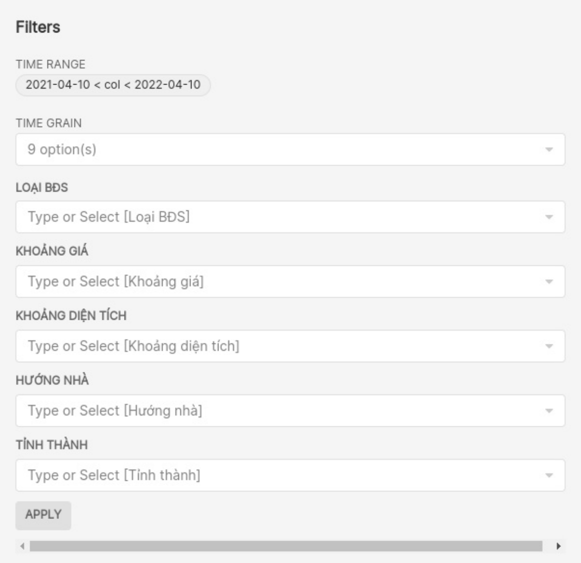
	Postings	Error	Parsing
1	Bán nhà hẻm xe hơi Q.5, 3 lầu đẹp, chỉ hơn 8 tỷ. Trần Phú, P.4, Q.5. Diện tích: 3.6x12m. Selling a house in a car alley in District 5, 3 beautiful floors, just over 8 billion. Tran Phu, Ward 4, District 5. Area: 3.6x12m.	missing address "Trần Phú, P.4, Q.5"	Final area: 43.2 m2 Final address: NA
2	Cận Tết nên cần bán nhà hẻm 80 Lê Hồng Phong P2 Quận 5. Diện tích 32m2. 1 trệt 3 lầu giá bán 6.2 tỷ bớt lộc khách thiện chí nhà gần Trần Hưng Đạo và Nguyễn Trãi cách mặt tiền 20m. Tet holiday should be sold at the house, alley 80 Le Hong Phong P2 District 5. Area 32m2. 1 ground floor 3 floors selling price 6.2 billion to reduce goodwill customers near Tran Hung Dao and Nguyen Trai 20m from the front.	missing near facility "Nguyen Trai"	Final address: Hẻm 80 Đường Lê Hồng Phong Phường 2 Quận 5 Thành phố Hồ Chí Minh Final area: 32m2
3	Bán căn nhà 5 tầng phường 2 quận 3. Diện tích ngang 3.8m dài 5.4m. Nhà đã hoàn công 7/2020. Phòng ngủ: 4 phòng ngủ; 5 wc. Tiếp giáp các trục đường chính như CMT8, Tô Hiến Thành, Hoàng Sa, Trần Văn Đang. Giao thông dễ dàng di chuyển sang các quận 1, 10, Tân Bình, Phú Nhuận. Mua là công chứng sang tên ngay lập tức. Selling a 5-storey house in Ward 2, District 3. Area 3.8m wide, 5.4m long. The house was completed 7/2020. Bedrooms: 4 bedrooms; 5 toilets. Adjacent to main roads such as CMT8, To Hien Thanh, Hoang Sa, Tran Van Dang. Traffic is easy to move to districts 1, 10, Tan Binh, Phu Nhuan. Buy is notarized to the name immediately.		Final address: Phường 2 Quận 3 Thành phố Hồ Chí Minh Final area: NaN

FIGURE 6. Error examples. The texts with colored background are the detected entities. The first column show the descriptions of several real estate posts, the second column show their problem while the third column show some vital information extracted from the descriptions.

In particular, we want to find out if arranging the noise filtering module before the NER module positively impacts our main task. Table 6 shows the experiment results of our system on the test set with and without the noise filtering module. One can observe that including the noise filtering module improves system performance by up to 13.81%. As a result, both modules are essential in the named entity recognition approach to extracting information from real estate postings.

D. ERROR ANALYSIS

There are still cases of misidentification by our pipeline due to the ambiguity in detecting named entities. Even with entity types with many occurrences like `near_facility`, we still have incorrect or missed recognitions due to highly varied and complex contexts in the Vietnamese language like in some examples as shown in Table 6. The following reasons could potentially explain these detection errors: (a) From specific points of view, the quantity and diversity of our data are still lacking, especially in the cases of entity types with relatively modest occurrence rates such as



Filters

TIME RANGE

2021-04-10 < col < 2022-04-10

TIME GRAIN

9 option(s)

LOẠI BĐS

Type or Select [Loại BĐS]

KHOẢNG GIÁ

Type or Select [Khoảng giá]

KHOẢNG DIỆN TÍCH

Type or Select [Khoảng diện tích]

HƯỚNG NHÀ

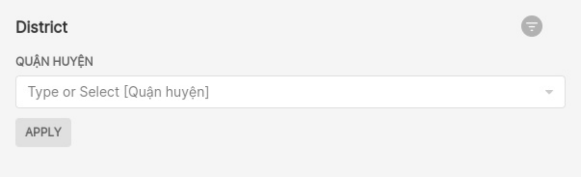
Type or Select [Hướng nhà]

TỈNH THÀNH

Type or Select [Tỉnh thành]

APPLY

FIGURE 7. General filter settings. Translation notes: *Loại Bất Động Sản*: property type, *Khoảng giá*: price range, *Khoảng diện tích*: area range, *Hướng nhà*: direction, *Tỉnh thành*: Province-level jurisdictional area.



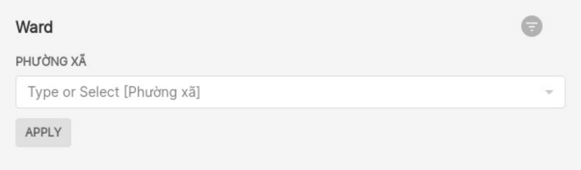
District

QUẬN HUYỆN

Type or Select [Quận huyện]

APPLY

FIGURE 8. Filter for district-level jurisdictional areas. Translation notes: *Quận huyện* refers to a district-level jurisdictional area, *Quận*: Urban district, *Huyện*: Rural district.



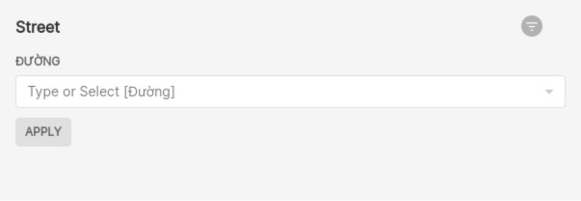
Ward

PHƯỜNG XÃ

Type or Select [Phường xã]

APPLY

FIGURE 9. Filter forward-level jurisdictional areas. Translation notes: *Phường xã* refers to a ward-level jurisdictional area, *Phường*: Ward, *Xã*: Commune.



Street

ĐƯỜNG

Type or Select [Đường]

APPLY

FIGURE 10. Filter for streets. Translation notes: *Đường*: Street.

`floor_id` or `special_view`, as can be seen in Figure 4; (b) Some entity types, such as address or area, are usually more complicated (e.g., substantial use of abbreviations because of free text entry, many different expressions for the same thing) and require polished post-processing for reliable outputs.

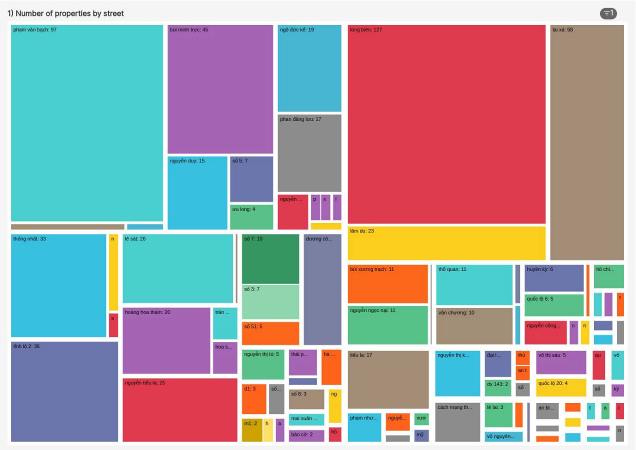


FIGURE 11. Number of properties by street: A proportional area chart where the relative size of each square illustrates the number of properties in a particular street in the chosen settings of the filters.

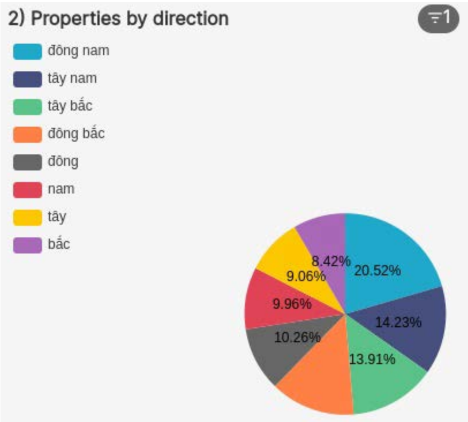


FIGURE 12. Properties by direction: The proportion of properties facing each direction. Translation notes: *Đông*: East, *Tây*: West, *Nam*: South, *Bắc*: North, *Đông Bắc*: North-East, *Tây Bắc*: North-West, *Đông Nam*: South-East, *Tây Nam*: South-West.

E. OTHER LIMITS OF THE CURRENT PLATFORMS

Besides the shortcomings discussed in Section V-D, our current platforms are also being faced with other limitations, which we now discuss.

- In Table 8, we have 13 types of entities addressed in this study. While entities of types encompass a wide range of interests to our real estate partner, many potentially useful entities are absent here, such as the real estate project that a property belongs to and the property's legal status.
- While the performance of our proposed system is of great priority, there are still some cases where they do not work as well as expected. Noteworthy mentions include the accuracy of detecting the correct address of each property, which can be a very complicated task for free text entry data with much ambiguity and plenty of abbreviations.

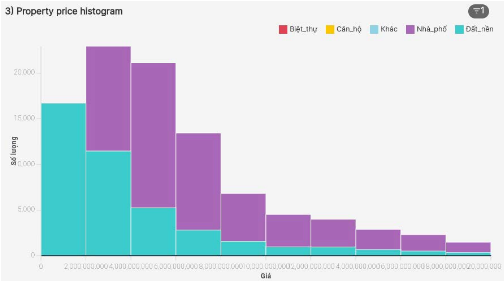


FIGURE 13. Property price histogram: The number of properties at each price point, subjected to the price range filter. Translation notes: *Biệt thự′*: Mansion, *Căn hộ*: Apartment, *Khác*: Other, *Nhà phố*: Townhouse, *Đất nền*: Land plot, *Giá*: price.

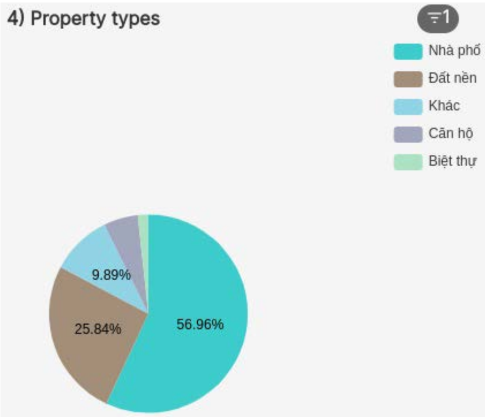


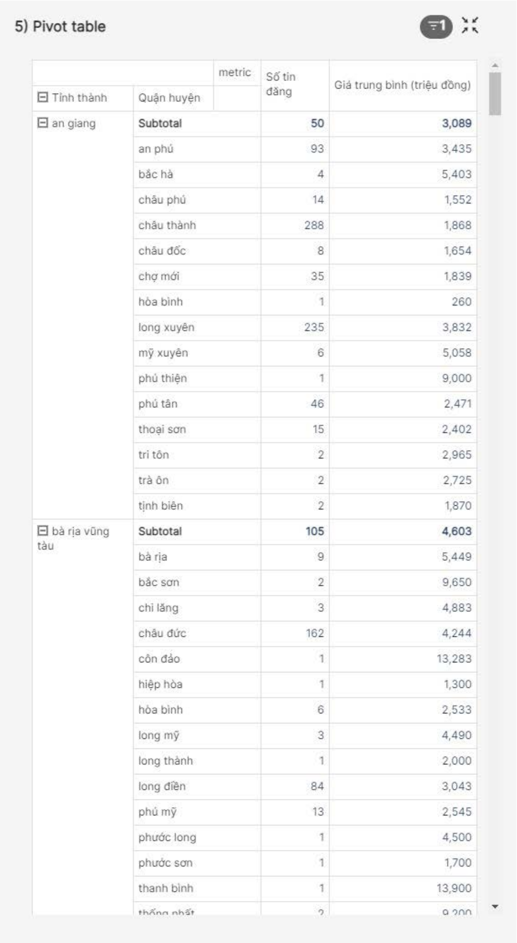
FIGURE 14. Property types: The proportion of properties of each type. Translation notes: *Nhà phố*: Townhouse, *Đất nền*: Land plot, *Khác*: Other, *Căn hộ*: Apartment, *Biệt thự′*: Mansion.

- While being deployed in real production, it is inevitable that faulty predictions made by our system stack up and cause significant consequences. One should enforce inspection protocols to scan the databases and correct wrongly predicted records periodically.

VI. A PRACTICAL APPLICATION BASED ON THE PROPOSED PLATFORM

On the more practical side, the proposed pipeline is deployed to process real estate data for the real estate partner. In particular, our pipeline has processed more than 400,000 records. The system can be summarized as follows:

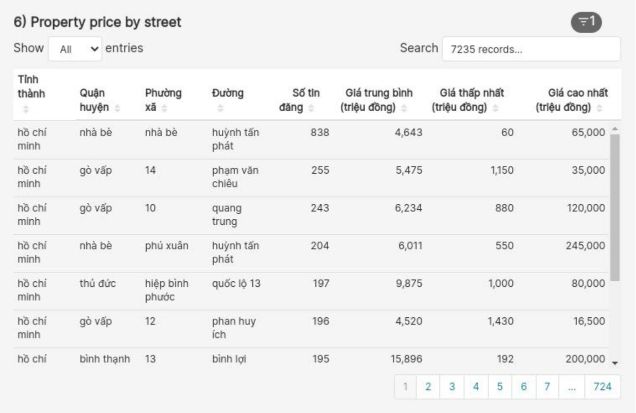
1) Data from various real estate sources are periodically collected through a dedicated pipeline into several PostgreSQL databases, each with its source-specific format.
2) Data from those databases are mapped into a new table (let's say, table *real_estate_news*) in a unified format in a Redshift database. At the same time, the original records still stay in their source databases as backups.
3) The records in *real_estate_news* go through the noise filtering module and the prediction results for "noisy"/"not noisy" are stored in another table (called table *real_estate_noises*).



5) Pivot table

	metric	Số tin đăng	Giá trung bình (triệu đồng)
⊟ Tỉnh thành	Quận huyện		
⊟ an giang	**Subtotal**	**50**	**3,089**
	an phú	93	3,435
	bắc hà	4	5,403
	châu phú	14	1,552
	châu thành	288	1,868
	châu đốc	8	1,654
	chợ mới	35	1,839
	hòa bình	1	260
	long xuyên	235	3,832
	mỹ xuyên	6	5,058
	phú thiện	1	9,000
	phú tân	46	2,471
	thoại sơn	15	2,402
	tri tôn	2	2,965
	trà ôn	2	2,725
	tịnh biên	2	1,870
⊟ bà rịa vũng tàu	**Subtotal**	**105**	**4,603**
	bà rịa	9	5,449
	bắc sơn	2	9,650
	chi lăng	3	4,883
	châu đức	162	4,244
	côn đảo	1	13,283
	hiệp hòa	1	1,300
	hòa bình	6	2,533
	long mỹ	3	4,490
	long thành	1	2,000
	long điền	84	3,043
	phú mỹ	13	2,545
	phước long	1	4,500
	phước sơn	1	1,700
	thanh bình	1	13,900
	thống nhất	2	9,200

FIGURE 15. Pivot table: A pivot table showing every province/city and its districts along with the number of properties (*Số tin đăng*) and average price (*Giá trung bình*, in millions VND) in the said district. Translation notes: *Tỉnh thành*: Province-level jurisdictional area, *Quận huyện*: District-level jurisdictional area, *Số tin đăng*: number of posts, *Giá trung bình*: average price.



6) Property price by street

Show All entries

Search 7235 records...

Tỉnh thành	Quận huyện	Phường xã	Đường	Số tin đăng	Giá trung bình (triệu đồng)	Giá thấp nhất (triệu đồng)	Giá cao nhất (triệu đồng)
hồ chí minh	nhà bè	nhà bè	huỳnh tấn phát	838	4,643	60	65,000
hồ chí minh	gò vấp	14	phạm văn chiêu	255	5,475	1,150	35,000
hồ chí minh	gò vấp	10	quang trung	243	6,234	880	120,000
hồ chí minh	nhà bè	phú xuân	huỳnh tấn phát	204	6,011	550	245,000
hồ chí minh	thủ đức	hiệp bình phước	quốc lộ 13	197	9,875	1,000	80,000
hồ chí minh	gò vấp	12	phan huy ích	196	4,520	1,430	16,500
hồ chí	bình thạnh	13	bình lợi	195	15,896	192	200,000

FIGURE 16. *Tỉnh thành*: Province-level jurisdictional area, *Quận huyện*: District-level jurisdictional area, *Phường xã*: Ward-level jurisdictional area, *Đường*: Street, *Số tin đăng*: number of posts, *Giá trung bình*: average price, *Giá thấp nhất*: minimum price, *Giá cao nhất*: maximum price, *triệu đồng*: Millions of VNDs. (VND is the currency of Vietnam.)

4) The records classified as not noise then goes through the NER module and some post-processing, where the useful information is extracted.

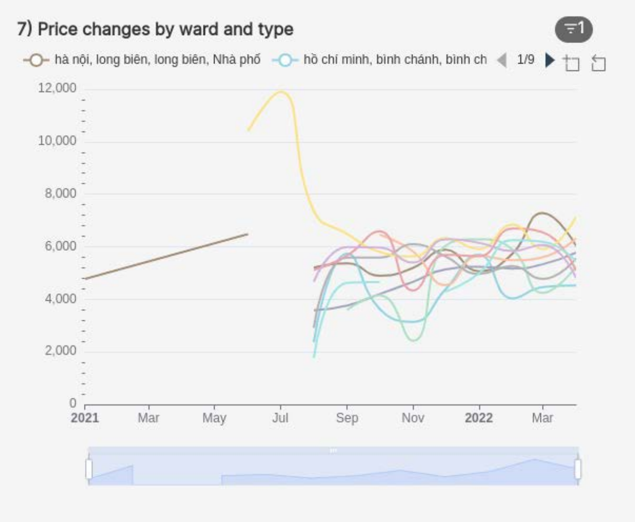


FIGURE 17. Price changes by ward and type: Graphical visualization where each point in a line represents each property type's average price in each ward.

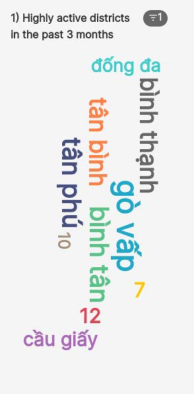
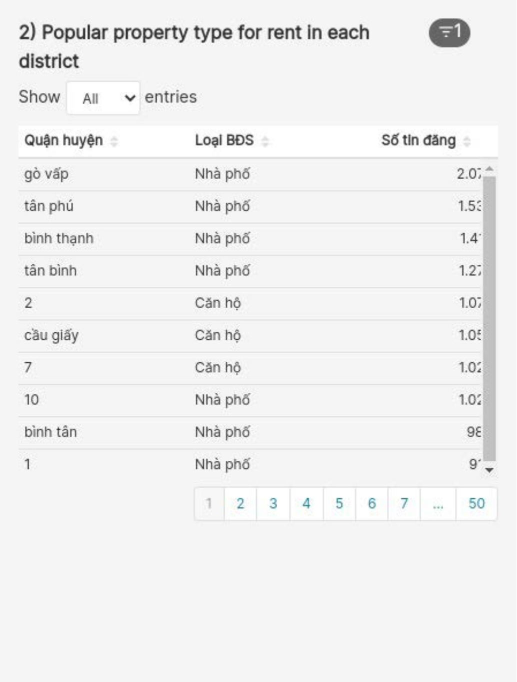


FIGURE 18. Highly active districts in the past three months: The name of the districts with the largest numbers of posts in the past three months. Translation notes: *Tỉnh thành*: Province-level jurisdictional area, *Quận huyện*: District-level jurisdictional area, *Loại Bất Động Sản*: property type, *Số tin đăng*: number of posts.

5) Those not-noisy records and their extracted information are stored in a new table (*real_estate_recognitions*). One can then copy these records to another PostgresSQL database for data analytics.

There are several important things to note in this pipeline.

- The pipeline orchestration is performed using Apache Airflow.
- We automated the pipeline to separate the aggregated data for analytics with the extracted data from the NER model by storing them in different databases. This work can help us make the NER extraction step and the implementation of the analytics dashboards work separately and independently to avoid dependency issues when problems or bugs happen during the NER extraction process.



2) Popular property type for rent in each district

Quận huyện	Loại BĐS	Số tin đăng
gò vấp	Nhà phố	2.07
tân phú	Nhà phố	1.53
bình thạnh	Nhà phố	1.4
tân bình	Nhà phố	1.27
2	Căn hộ	1.07
cầu giấy	Căn hộ	1.05
7	Căn hộ	1.02
10	Nhà phố	1.02
bình tân	Nhà phố	98
1	Nhà phố	9

FIGURE 19. Popular property type for rent in each district: Show each jurisdictional district and the property type with most posts in said district. Translation notes: *Loại Bất Động Sản*: property type, *Nhà phố*: Townhouse, *Căn hộ*: Apartment.

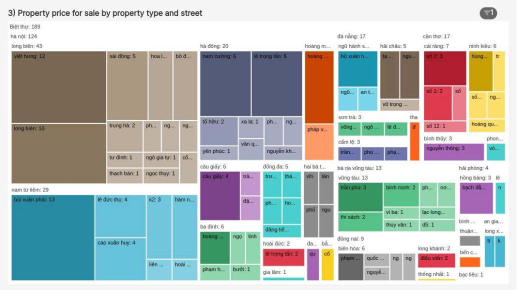


FIGURE 20. Property price for sale by property type and street: A proportional area chart where the relative size of each square illustrates the number of properties of the most popular type in a particular street in the chosen settings of the filters.

- Most computationally intensive operations, mainly the noise filtering and NER modules, are performed using AWS stacks (in particular, EKS cluster and EC2 instances) on data stored on the Redshift databases. This arrangement allows us to scale up our operations through AWS cloud computing services in case we have more data than initially planned.

Our analytics system includes several data marts for various purposes, each with a dashboard built in Superset to display vital analytics of interest for our real estate partner. For illustration, we will show the dashboard samples and give a brief Vietnamese-English translation if necessary. In the following subsections, we will provide the individual charts in each dashboard, while the full dashboards themselves are given in the Appendix .

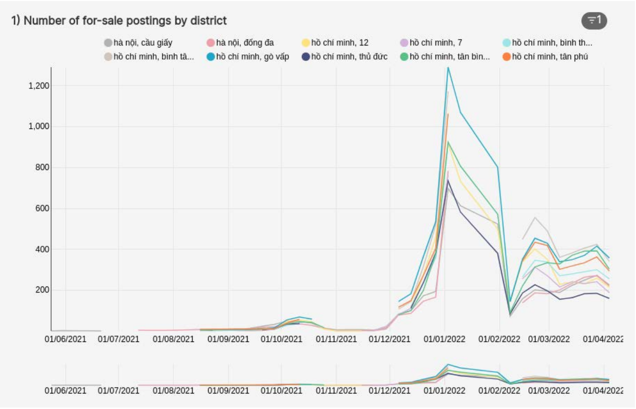


FIGURE 21. Number of for-sale posts by district: Each line illustrates the number of for-sale posts in each district over a period of time.

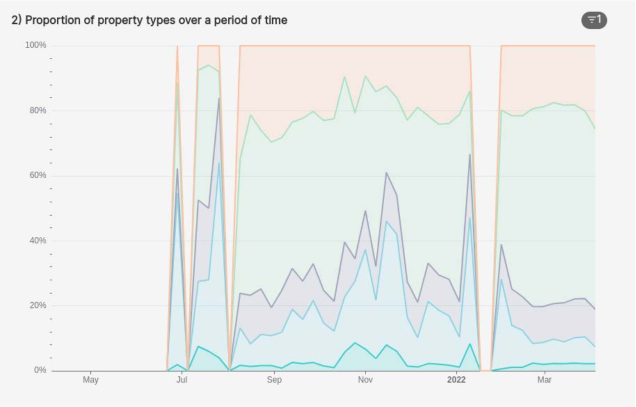


FIGURE 22. Proportion of property types over a period of time: Each time point on this graph shows the proportions of each property type at that particular time. Note that the empty area means that there are no posts.

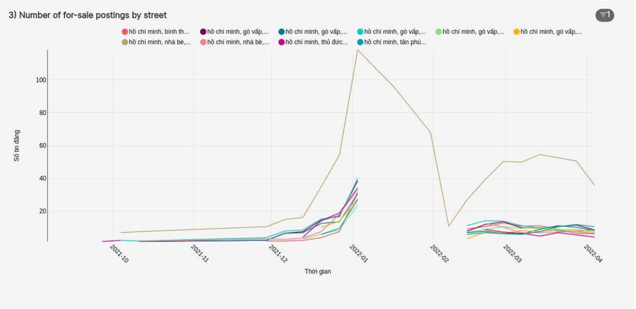


FIGURE 23. Number of for-sale posts by street: Each line illustrates the number of for-sale posts in each street over a period of time.

A. DATA FILTERS FOR THE DASHBOARDS

First, a common but essential piece of our dashboards is the filters, which allow one the option to display only the information of interest. The settings of these filters can be explained as follows.

1) Time range: limiting the shown information to a specific given period of time, e.g. 2022-01-01 00:00:00 to 2022-04-14 00:00:00

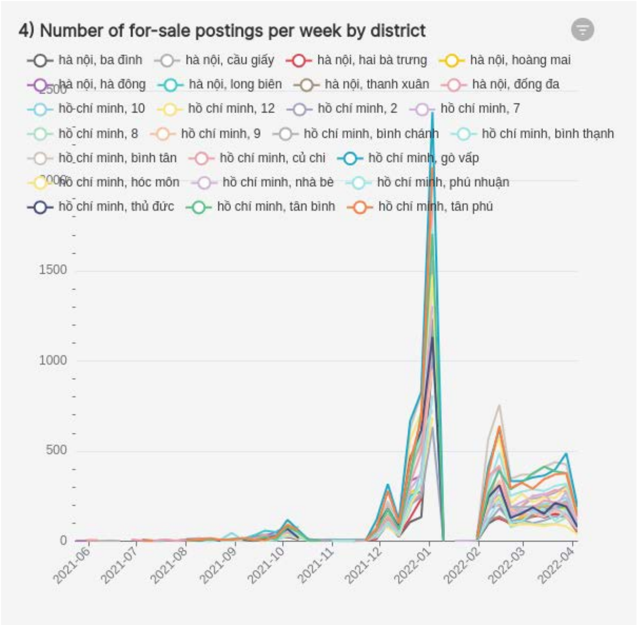


FIGURE 24. Number of for-sale posts per week by district: Each line illustrates the number of for-sale posts in each district over a period of time. This has more districts than the first chart (this chart has 25 districts while the first chart has ten districts).

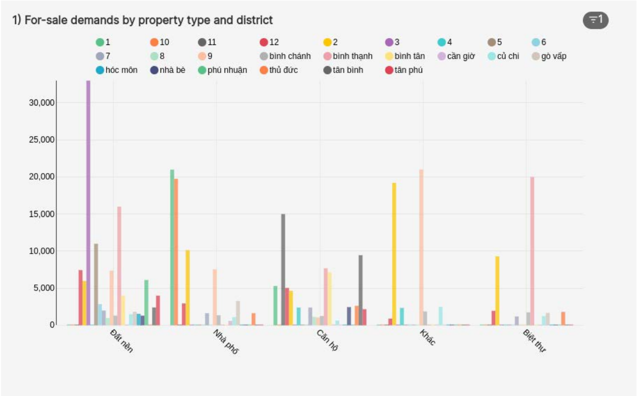


FIGURE 25. For-sale demands by property type and district: This chart shows each district's the number of for-sale demands by property type. Translation notes:***Đất nền***: Land plot, ***Nhà phố***: Townhouse, ***Căn hộ***: Apartment, ***Khác***: Other, ***Biệt thụ'***: Mansion.

2) Time grain: time granularity for the visualization, can be second, minute, hour, day, week and so on
3) Property types (*Loại Bất Động Sản* / *Loại BĐS*): specific types such as Townhouse (*Nhà phố*), Apartment (*Căn hộ*), Mansion (*Biệt thụ'*), Land plot (*Đất*).
4) Price range (*Khoảng giá*): Limit the price inside a specific range
5) Area range (*Khoảng diện tích*): Limit the area of the property inside a specific range
6) Direction (*Hướng nhà*): The direction of the property, e.g. East (*Đông*), West (*Tây*), South (*Nam*), North (*Bắc*), North-East (*Đông Bắc*), North-West (*Tây Bắc*), South-East (*Đông Nam*), South-West (*Tây Nam*)

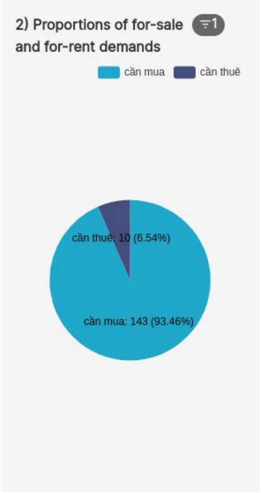


FIGURE 26. Proportions of for-sale and for-rent demands: Proportions of for-sale demand and for-rent demand posts. Translation notes: *Cần mua*: for-sale demands, *Cần thuê*: for-rent demands.

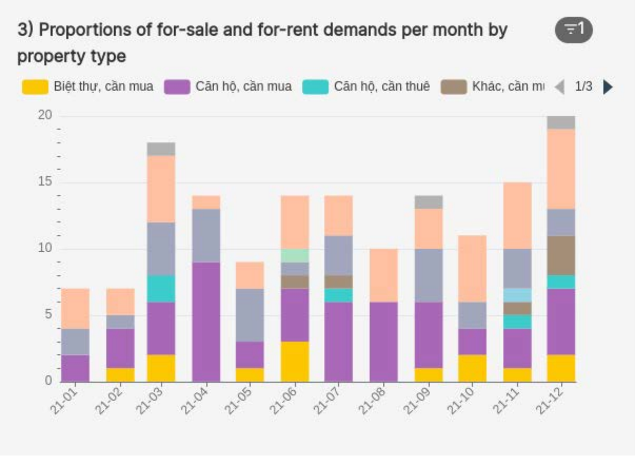


FIGURE 27. Proportions of for-sale and for-rent demands per month by property type: This stacked bar chart illustrates each property type's relative numbers of demand posts in each month. Translation notes: *Biệt thự′*: Mansion, *Căn hộ*: Apartment, *Khác*: Other, *Cần mua*: for-sale demands, *Cần thuê*: for-rent demands.

7) Jurisdictional area: Limit where the properties are located

- *Tỉnh thành*: Province-level jurisdictional area.
- *Quận huyện*: District-level jurisdictional area.
- *Phường xã*: Ward-level jurisdictional area.
- *Đường*: Street

The explanations for the charts included in this dashboard are as follows.

B. DASHBOARD 1: PROPERTY PRICE STATISTICS

(Fig. 30) In this dashboard, we mainly show the statistics regarding the number of property posts in each jurisdictional area with the following charts.

C. DASHBOARD 2: HIGHLY ACTIVE AREAS

This dashboard includes two tabs which can be presented as follows.

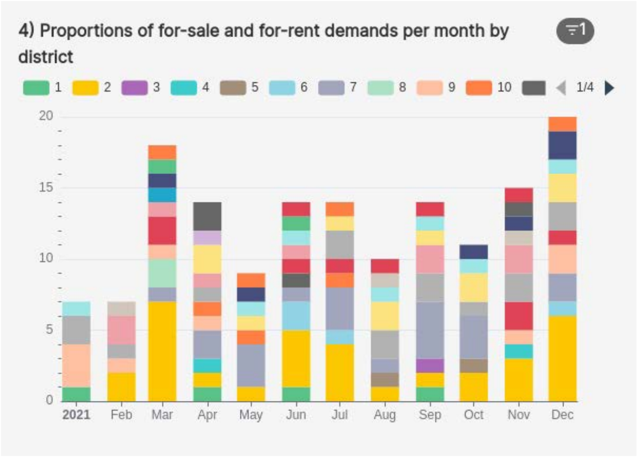
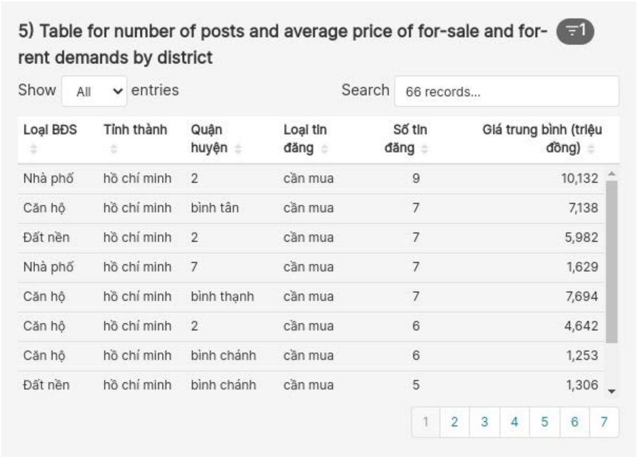


FIGURE 28. Proportions of for-sale and for-rent demands per month by district: This stacked bar chart illustrates each district's relative numbers of demand posts in each month.


5) Table for number of posts and average price of for-sale and for-rent demands by district

Show All entries Search 66 records...

Loại BĐS	Tỉnh thành	Quận huyện	Loại tin đăng	Số tin đăng	Giá trung bình (triệu đồng)
Nhà phố	hồ chí minh	2	cần mua	9	10,132
Căn hộ	hồ chí minh	bình tân	cần mua	7	7,138
Đất nền	hồ chí minh	2	cần mua	7	5,982
Nhà phố	hồ chí minh	7	cần mua	7	1,629
Căn hộ	hồ chí minh	bình thạnh	cần mua	7	7,694
Căn hộ	hồ chí minh	2	cần mua	6	4,642
Căn hộ	hồ chí minh	bình chánh	cần mua	6	1,253
Đất nền	hồ chí minh	bình chánh	cần mua	5	1,306

FIGURE 29. Table for the number of posts and the average price of for-sale and for-rent demands by district: The number of posts and average price for for-sale demands and for-rent demands in each district. Translation notes: *Loại Bất Động Sản*: property type, *Nhà phố*: Townhouse, *Căn hộ*: Apartment, *Đất nền*: Land plot, *Tỉnh thành*: Province-level jurisdictional area, *Quận huyện*: District-level jurisdictional area, *Loại tin đăng*: post type, *Cần mua*: for-sale demands, *Cần thuê*: for-rent demands, *Số tin đăng*: number of posts, *Giá trung bình*: average price, *triệu đồng*: Millions of VNDs (VND is the currency of Vietnam).

1) TAB 1: OVERVIEW

(Fig. 31) Tab 1 presents the statistics on the jurisdictional areas with the highest amount of posts for each type with the filters mentioned in VI-B as well as the hottest areas (having the largest number of posts) in recent time. The charts in this tab are as follows

2) TAB 2: CHANGES BY TIME

(Fig. 32) Tab 2 presents more detailed statistics (compared to Tab 1) with information over a chosen period. The charts in this tab are as follows

D. DASHBOARD 3: PROPERTY DEMANDS

(Fig. 33) This dashboard shows the statistics on the number of the demands of for-sale and for-rent properties with filters

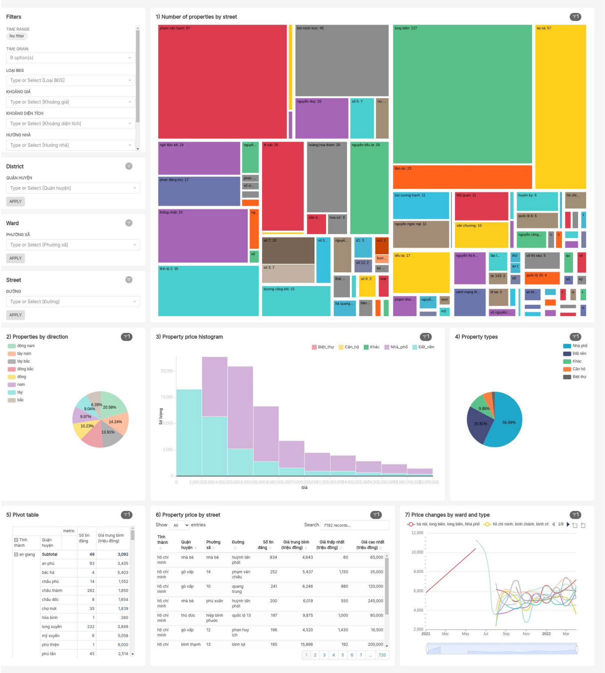

Filters
District
Ward
Street
1) Number of properties by street
2) Properties by direction
3) Property price histogram
4) Property types
5) Pivot table
6) Property price by street
7) Price changes by ward and type

FIGURE 30. The 1st dashboard in our analytics platform: showing various information regarding the number of property posts and average price within each jurisdictional area as well as the proportions of each property type or direction. The changes in price through time by street and property type are also presented as well. Translation notes: *Loại Bất Động Sản*: property type (*Nhà phố*: Townhouse, *Căn hộ*: Apartment,*Biệt thự*: Mansion, *Đất*: Land plot, *Khác*: Other), *Khoảng giá*: price range, *Khoảng diện tích*: area range, *Hướng nhà*: direction (*Đông*: East, *Tây*: West, *Nam*: South, *Bắc*: North, *Đông Bắc*: North-East, *Tây Bắc*: North-West, *Đông Nam*: South-East, *Tây Nam*: South-West), *Tỉnh thành*: Province, City, Municipality, *Quận huyện*: District, *Phường xã*: Ward, Town, Commune, *Đường*: Street, *Số tin đăng*: number of posts, *Giá trung bình*: average price, *Giá thấp nhất*: minimum price, *Giá cao nhất*: maximum price.

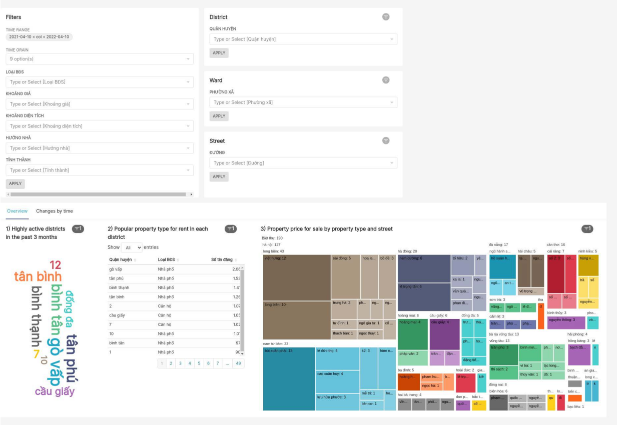

Filters
District
Ward
Street
Overview
Changes by time
1) Highly active districts in the past 3 months
2) Popular property type for rent in each district
3) Property price for sale by property type and street

FIGURE 31. Tab 1 of the 2nd dashboard in our analytics platform: showing the areas with high activity in a chosen period of time and their overall statistics. Translation notes: *Loại Bất Động Sản*: property type (*Nhà phố*: Townhouse, *Căn hộ*: Apartment,*Biệt thự*: Mansion, *Đất*: Land plot, *Khác*: Other), *Khoảng giá*: price range, *Khoảng diện tích*: area range, *Hướng nhà*: direction, *Tỉnh thành*: Province, City, Municipality, *Quận huyện*: District, *Phường xã*: Ward, Town, Commune, *Đường*: Street, *Số tin đăng*: number of posts.

mentioned in VI-B. The associated charts are explained as follows.

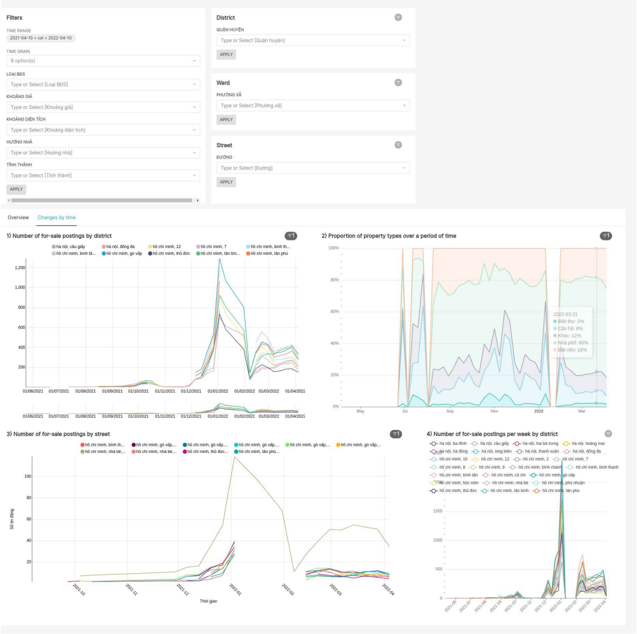

Filters
District
Ward
Street
Overview
Changes by time
1) Number of for-sale postings by district
2) Proportion of property types over a period of time
3) Number of for-sale postings by street
4) Number of for-sale postings per week by district

FIGURE 32. Tab 2 of the 2nd dashboard in our analytics platform: showing statistics over a period of time for the areas with high activity. Translation notes: *Loại Bất Động Sản*: property type, *Khoảng giá*: price range, *Khoảng diện tích*: area range, *Hướng nhà*: direction, *Tỉnh thành*: Province, City, Municipality, *Quận huyện*: District, *Phường xã*: Ward, Town, Commune, *Đường*: Street.

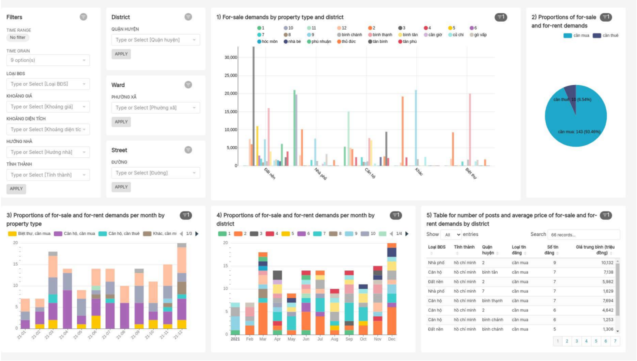

Filters
District
Ward
Street
1) For-sale demands by property type and district
2) Proportions of for-sale and for-rent demands
3) Proportions of for-sale and for-rent demands per month by property type
4) Proportions of for-sale and for-rent demands per month by district
5) Table for number of posts and average price of for-sale and for-rent demands by district

FIGURE 33. The 3rd dashboard in our analytics platform: showing statistics for demands of for-sale and for-rent properties. Translation notes: *Loại Bất Động Sản*: property type (*Nhà phố*: Townhouse, *Căn hộ*: Apartment,*Biệt thự*: Mansion, *Đất*: Land plot, *Khác*: Other), *Khoảng giá*: price range, *Khoảng diện tích*: area range, *Hướng nhà*: direction, *Tỉnh thành*: Province, City, Municipality, *Quận huyện*: District, *Phường xã*: Ward, Town, Commune, *Đường*: Street, *Số tin đăng*: number of posts, *Giá trung bình*: average price, *Loại tin đăng*: post type (*cần mua*: looking for for-sale properties, *cần mua*: looking for for-rent properties).

VII. CONCLUSION AND FUTURE WORKS

We have presented a Vietnamese dataset in the real estate domain, focusing on the named entity recognition task. Next, we have constructed a system that can give real-estate listing posts as input and eliminate noisy records while keeping those that are not noisy. Those posts are then passed through pre-processing steps, and then the NER module can extract these inputs into 13 selected entity types. Finally, the parsed information from the predicted named entity will be visualized

TABLE 7. Third party software and licenses.

No.	Software	License
1	Apache Airflow	Apache License https://airflow.apache.org/docs/apache-airflow/stable/license.html
2	PostgresDB	PostgreSQL License https://www.postgresql.org/about/licence
3	FastAPI	MIT license https://github.com/tiangolo/fastapi/blob/master/LICENSE
4	Doccano	MIT license https://github.com/doccano/doccano/blob/master/LICENSE
5	Superset	Apache License https://github.com/apache/superset/blob/master/LICENSE.txt
6	Spacy	MIT License https://github.com/explosion/spaCy/blob/master/LICENSE
7	PhoBERT	MIT License https://github.com/VinAIResearch/PhoBERT/blob/master/LICENSE

TABLE 8. Summarized definitions of 13 entity types that we use for our study.

No.	Entity Type	Definition
1	property_type	The type of the specific property, such as apartment, house, land as well as their subtype, e.g. penthouse apartment, front road house
2	address	The information to identify the location of the property, e.g. house number, street, ward, district, city.
3	price	The price related to real estate mentioned in the real estate.
4	size	The information regarding the dimensions of the mentioned property, e.g. width and depth of a land lot.
5	area	The area of the real estate, e.g. 100 square meters
6	front_road	The information regarding if the mentioned property is on a front road.
7	direction	The house direction of the real estate, e.g. East or West.
8	alley_type	The information regarding if the mentioned property is in an alley.
9	house_design	The architectural design of the house, e.g. number of bedrooms, bathrooms
10	internal_facility	The available facilities and amenities inside the property, e.g. television set, fully equiped kitchen, refrigerator.
11	special_view	The information regarding if the property have a special view, e.g. cityscape view, river view.
12	near_facility	The near by facilities, areas that might be of interest, e.g. parks, supermarkets, shopping malls.
13	floor_id	The floor where the mentioned property (usually apartment) belongs.

comprehensively on the data analytics dashboards for tasks such as extracting house prices by area or type of property.

In the future, we aim to enhance the performance of each module in the system and extend the scope of information that can be processed.

APPENDIX. THIRD PARTY SOFTWARE AND LICENSES

This section provides licensing information for the third-party software used by us in this study. One can check the list of sofware used and the corresponding licenses at Table 7.

APPENDIX. ANNOTATION GUIDELINE

See Table 8.

APPENDIX. DASHBOARDS - FULL VIEW

A. DASHBOARD 1: PROPERTY PRICE STATISTICS

See Figure 30.

B. DASHBOARD 2: HIGHLY ACTIVE AREAS

See Figures 31 and 32.

C. DASHBOARD 3: PROPERTY DEMANDS

See Figure 33.

REFERENCES

[1] L. V. Pham and S. B. Pham, "Information extraction for Vietnamese real estate advertisements," in *Proc. 4th Int. Conf. Knowl. Syst. Eng.*, Aug. 2012, pp. 181–186.

[2] S. Huynh, K. Le, N. Dang, B. Le, D. Huynh, B. T. Nguyen, T. T. Nguyen, and N. Y. T. Ho, "Named entity recognition for Vietnamese real estate advertisements," in *Proc. Int. Conf. NICS*, Dec. 2021, pp. 23–28.

[3] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticæ Investigationes*, vol. 30, no. 1, pp. 3–26, Jan. 2007.

[4] E. F. Tjong Kim Sang and F. De Meulder, "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition," in *Proc. 7th Conf. Natural Lang. Learn. (HLT-NAACL)*, 2003, pp. 142–147.

[5] W. Xu, A. Ritter, T. Baldwin, and A. Rahimi, "Proceedings of the seventh workshop on noisy user-generated text (W-NUT 2021)," in *Proc. 7th Workshop Noisy User-Generated Text (W-NUT)*, 2021, pp. 1–18.

[6] K. Shaalan and H. Raza, "Arabic named entity recognition from diverse text types," in *Advances in Natural Language Processing* (Lecture Notes in Computer Science), vol. 5221, B. Nordström and A. Ranta, Eds. Berlin, Germany: Springer, 2008, doi: 10.1007/978-3-540-85287-2_42.

[7] Y. Wu, J. Xu, M. Jiang, Y. Zhang, and H. Xu, "A study of neural word embeddings for named entity recognition in clinical text," in *Proc. AMIA Annu. Symp.* Bethesda, MD, USA: American Medical Informatics Association, 2015, p. 1326.

[8] D. Benikova, C. Biemann, and M. Reznicek, "Nosta-D named entity annotation for German: Guidelines and dataset," in *Proc. LREC*, 2014, pp. 2524–2531.

[9] Q. T. Tran, T. T. Pham, Q. H. Ngo, D. Dinh, and N. Collier, "Named entity recognition in Vietnamese documents," *Prog. Informat.*, p. 5, Mar. 2007. [Online]. Available: https://www.nii.ac.jp/pi/n4/4_5.pdf

[10] D. B. Nguyen, S. H. Hoang, S. B. Pham, and T. P. Nguyen, "Named entity recognition for Vietnamese," in *Intelligent Information and Database Systems*, N. T. Nguyen, M. T. Le, and J. Świątek, Eds. Berlin, Germany: Springer, 2010, pp. 205–214.

[11] T.-N. Pham, L. M. Nguyen, and Q.-T. Ha, "Named entity recognition for Vietnamese documents using semi-supervised learning method of CRFs with generalized expectation criteria," in *Proc. Int. Conf. Asian Lang. Process.*, Nov. 2012, pp. 85–88.

[12] H. T. M. Nguyen, Q. T. Ngo, L. X. Vu, V. M. Tran, and H. T. T. Nguyen, "VLSP shared task: Named entity recognition," *J. Comput. Sci. Cybern.*, vol. 34, no. 4, pp. 283–294, Jan. 2019. [Online]. Available: https://vjs.ac.vn/index.php/jcc/article/view/13161

[13] H. T. M. Nguyen, Q. T. Ngo, L. X. Vu, V. M. Tran, and H. T. T. Nguyen, "VLSP shared task: Named entity recognition," *J. Comput. Sci. Cybern.*, vol. 34, no. 4, pp. 283–294, Jan. 2019.

[14] N. C. Lê, N.-Y. Nguyen, A.-D. Trinh, and H. Vu, "On the Vietnamese name entity recognition: A deep learning method approach," in *Proc. RIVF Int. Conf. Comput. Commun. Technol. (RIVF)*, Oct. 2020, pp. 1–5.

[15] T. H. Truong, M. H. Dao, and D. Q. Nguyen, "COVID-19 named entity recognition for Vietnamese," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2021, pp. 2146–2153.

[16] D. Q. Nguyen and A. T. Nguyen, "PhoBERT: Pre-trained language models for Vietnamese," in *Proc. Findings Assoc. Comput. Linguistics (EMNLP)*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 1037–1042.

[17] V. H. Nguyen, H. T. Nguyen, and V. Snasel, "Text normalization for named entity recognition in Vietnamese tweets," *Comput. Social Netw.*, vol. 3, no. 1, pp. 1–16, Dec. 2016.

[18] X. Huang, Y. Chen, S. Wu, J. Zhao, Y. Xie, and W. Sun, "Named entity recognition via noise aware training mechanism with data filter," in *Proc. Findings Assoc. Comput. Linguistics (ACL-IJCNLP)*. Stroudsburg, PA, USA: Association for Computational Linguistics, Aug. 2021, pp. 4791–4803.

[19] Y. Meng, Y. Zhang, J. Huang, X. Wang, Y. Zhang, H. Ji, and J. Han, "Distantly-supervised named entity recognition with noise-robust learning and language model augmented self-training," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2021, pp. 10367–10378.

[20] T. Vu, D. Q. Nguyen, D. Q. Nguyen, M. Dras, and M. Johnson, "VnCoreNLP: A Vietnamese natural language processing toolkit," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Demonstrations*. New Orleans, LA, USA: Association for Computational Linguistics, 2018, pp. 56–60.

[21] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1746–1751.

[22] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1. Minneapolis, MN, USA: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[24] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, Jul. 2020, pp. 8440–8451.

[25] H. A. Pandya, B. Ardeshna, and B. S. Bhatt, "Cascading adaptors to leverage English data to improve performance of question answering for low-resource languages," 2021, *arXiv:2112.09866*.

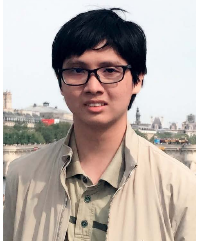
[26] R. Panchendrarajan and A. Amaresan, "Bidirectional LSTM-CRF for named entity recognition," in *Proc. 32nd Pacific Asia Conf. Lang., Inf. Comput.*, 2018, pp. 1–10.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NIPS)*. Red Hook, NY, USA: Curran Associates, 2017, pp. 6000–6010.

[28] A. T. Nguyen, M. H. Dao, and D. Q. Nguyen, "A pilot study of text-to-SQL semantic parsing for Vietnamese," in *Proc. Findings Assoc. Comput. Linguistics (EMNLP)*, 2020, pp. 4079–4085.

[29] S. V. Xuan, T. Vu, S. Tran, and L. Jiang, "ETNLP: A visual-aided systematic approach to select pre-trained embeddings for a downstream task," in *Proc. Int. Conf. Recent Adv. Natural Lang. Process. (RANLP)*. Varna, Bulgaria: INCOMA, Sep. 2019, pp. 1285–1294.

BINH T. NGUYEN is the Head of the Department of Computer Science, Faculty of Mathematics and Computer Science, Vietnam National University Ho Chi Minh City (VNUHCM)—University of Science. He has had over ten years of experience in AI and data science. He defended his Ph.D. thesis with the highest honors at the École Polytechnique, Paris, France, in 2012. He has had over 60 publications and four patents filed in the USA and Canada. He also had substantial experience in building research and development teams for helping the company or the startup deliver AI products.



TUNG TRAN NGUYEN DOAN received the master's degree from the University of Rennes 1. He is currently working with the AISIA Research Laboratory, Ho Chi Minh City, Vietnam. During his master's study, he was an intern with the Laboratory of Analysis, Geometry, and Applications (LAGA), Sorbonne Paris Nord University. He also worked as a Research Assistant with the Max Planck Institute for Informatics, Saarbrücken, Germany. He is working on problems in natural language processing and data science.

SON THANH HUYNH received the Bachelor of Science degree from the Vietnam National University Ho Chi Minh City (VNUHCM)—University of Science, in 2020, where he is currently pursuing the master's degree with the Faculty of Mathematics and Computer Science. He is working as a Senior Research Engineer with the AISIA Research Laboratory, Ho Chi Minh City, Vietnam. His major research interests include applied machine learning and recommendation systems.

KHANH QUOC TRAN is currently an undergraduate research student with the Faculty of Information Science and Engineering, Vietnam National University Ho Chi Minh City (VNUHCM)—University of Information Technology. He is working as a Research Engineer with the AISIA Research Laboratory, Ho Chi Minh City, Vietnam. His publications are scientific papers for natural language processing and data science.

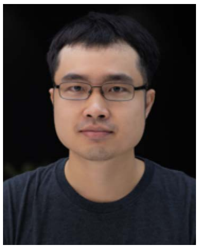
AN TRONG NGUYEN is currently an undergraduate research student in data science major with the Faculty of Information Science and Engineering, Vietnam National University Ho Chi Minh City (VNUHCM)—University of Information Technology. He is working as a Research Engineer with the AISIA Research Laboratory, Ho Chi Minh City, Vietnam. His research interests include natural language processing, big data, and data analytics. His latest publications are about visual question answering and credit scoring.

AN TRAN-HOAI LE is currently an undergraduate research student with the Faculty of Information Science and Engineering, University of Information Technology, VNU HCM. He is working as a Research Engineer with the AISIA Research Laboratory, Ho Chi Minh City, Vietnam. His publications are scientific papers for natural language processing and data science. His latest publications are about natural language processing and data science.

ANH MINH TRAN received the bachelor's degree from the Faculty of Mathematics and Computer Science, Vietnam National University Ho Chi Minh City (VNUHCM)—University of Science. Currently, she is an AI Engineer with JobhopIn, which is the first artificial intelligence recruitment platform in Vietnam. She spends most of her time and effort evenly blood and tears to gain more knowledge in the natural language processing field as well as in her career.

NHI HO received the bachelor's degree in mathematics from the Vietnam National University Ho Chi Minh City (VNUHCM)—University of Science, in 2020. She is currently working as a Data Engineer with DataFirst, Hung Thinh Corporation.

TRUNG T. NGUYEN received the Ph.D. degree in applied mathematics from Aix-Marseille University, France. After his Ph.D. degree, he has worked for several institutes including Centrale Supélec, École Centrale Paris, Paris, France; and the University of Bath, U.K., as a Postdoctoral Researcher, a Visiting Researcher, and Research Associate. He is currently working at Hung Thinh Corporation, Ho Chi Minh City, Vietnam. Along with the academic career, he had more than three year experience working at CEA Saclay (the French Alternative Energies and Atomic Energy Commission) and IRSN Cadarache (the French Institute of Radiation Protection and Nuclear Safety), France. He has had over ten years of experience in data science and applied mathematics.



DANG T. HUYNH received the Ph.D. degree in computer science from Sorbonne University, France. He is currently working with the AISIA Research Laboratory, Ho Chi Minh City, Vietnam. He is also an Invited Lecturer at the Department of Mathematics and Computer Science, University of Science (HCMUS), along with Vietnam-Franco programs organized by Sorbonne University and Bordeaux University in collaboration with Vietnam National University, Ho Chi Minh City (VNU-HCMC). He has over ten years of experience in AI and data science. He has held various AI/data science-related positions in research labs and companies across the U.S. and Europe, such as Bell Labs, INRIA (French Institute for Research in Computer Science and Automation), and Axon Enterprise. He has owned patents and published papers in leading AI conferences, including CVPR, ECCV, and WACV. His research interests include computer vision, natural language processing, and human–robot interaction.