EXTRACTING RESIDENTIAL REAL ESTATE ATTRIBUTES IN THE DA NANG MARKET USING THE PHOBERT MODEL

Author: Nguyen Thanh Luan, Le Thi Lan Huong

1. Overview

The report focuses on researching valuable real estate information based on descriptions available on online trading platforms. Understanding the limitations in the completeness of information and the significant time costs for both buyers and sellers is critical. Traditional search methods relying on basic information make it difficult for buyers to find suitable properties, while sellers often face errors and time-consuming tasks when inputting complete property details.

To optimize user experience in terms of time efficiency, this report proposes leveraging existing information and property descriptions from popular online platforms in the Vietnamese market. The key contributions of the report include:

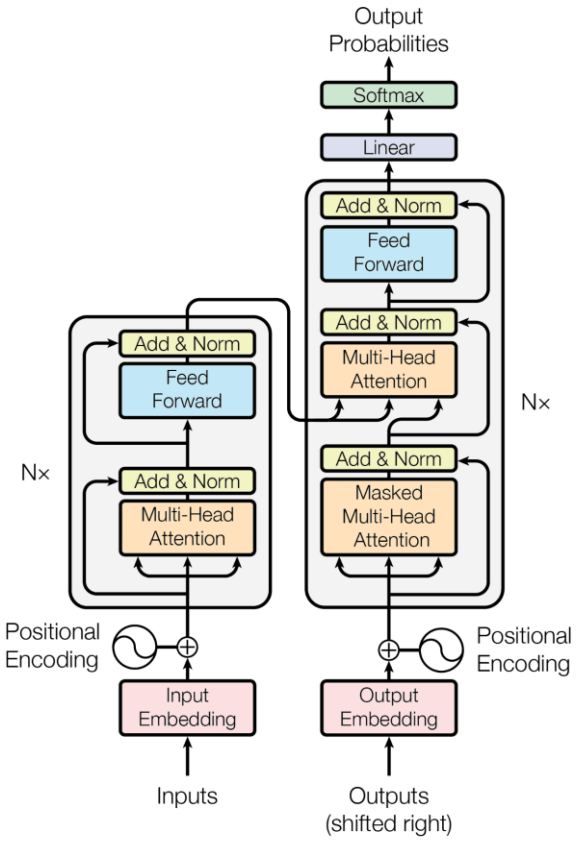
- Developing a pipeline to support the analysis of information that adds value to users.
- Proposing methods to address challenges in information extraction in the real estate domain.
- Providing a model training solution to overcome data limitations.
- Deploying research outcomes using real-world market data.

2. Methodology

Analysis of Behavior in the Real Estate Market:

- **Buyers:** Buyers are individuals who need real estate for personal use or investors looking to speculate through buying and reselling activities. Regardless of their ownership purpose, buyers require complete and transparent information. Besides tangible information such as price, area, etc., intangible factors like spiritual aspects, feng shui, or convenience are often undefined on search tools, making it challenging to find suitable properties.
- **For Sellers:** Sellers can be project developers, speculators, or individual homeowners looking to sell real estate for their needs. Their primary concern is the ability of their listings to reach potential customers. To achieve this, providing complete information is essential. However, this process can be time-consuming, lack expertise, and lead to errors.

Model Architecture: The study is implemented using the BERT (Bidirectional Encoder Representations from Transformers) model, an NLP model based on the Transformers architecture. Below is an illustration of the architecture.



Data normalization: plays a crucial role in ensuring consistency and comparability within textual real estate data. In this study, normalization is used to standardize spelling and values into a unified format, enabling more accurate analysis and comparison.

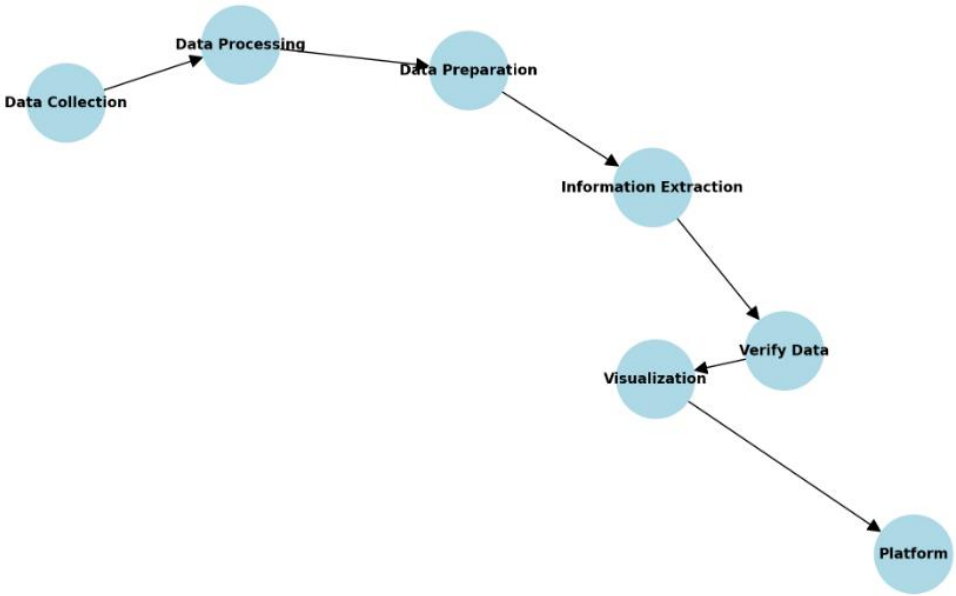
Named Entity Recognition: NER is an NLP task that identifies and categorizes specific entities within text into predefined categories such as names, locations, dates, or domain-specific entities.

Sequence classification: Sequence classification is an NLP task where an entire text sequence (e.g., a sentence, paragraph, or document) is assigned a single label or category.

Rule-based: Rule-based methods use manually defined linguistic patterns, regular expressions, or logical rules to extract, classify, or normalize information from text.

Data augmentation: Data augmentation in NLP involves generating new text samples from existing ones by modifying certain aspects of the text, such as synonyms, structure, or other contextual changes, while preserving the original meaning.

3. Framework



Data collection: The data was collected from three popular platforms in the real estate sector: nhatot.com.vn, batdongsan.com.vn, and alonhadat.com.vn. Some challenges encountered during the process included: session interruptions, being blocked, and restricted access to the platform, bypassing Cloudflare's user validation feature, and the loosely organized data structure, which made storage difficult.

Data processing: lowercasing, removing emojis, handling punctuation and redundant phrases.

Data Preparation: Use Doccano platform for the data labeling phase, apply pseudo-labeling to save time and reduce costs, and implement noisy classification to remove irrelevant or 'trash' data.

a. Noisy Definition

Descriptions containing information about multiple properties, or a single attribute with multiple values listed separately, or descriptions that do not include or clearly present information about the address, price, and area (the three most important details of a property) will be labeled as noisy data. Below are some examples of data considered to be noisy:

1) ks 5 sao diện tích 900m2 27 tầng 164p giá chuyển nhượng thoả thuận.
2) ks 5 sao diện tích 1.017m2 27 tầng 236p giá chuyển nhượng thoả thuận.
3) ks 5 sao diện tích 1.000m2 27 tầng 270p giá chuyển nhượng thoả thuận
4) resort 5 sao diện tích 70.000m2 (7ha) 300p và (quỹ đất xây 867 căn hộ) giá chuyển nhượng thoả thuận.
5) ks 4 sao diện tích 1.2000m2 16 tầng 200p giá 5xx tỷ.
Mọi thủ tục pháp lý tuỳ từng sản phẩm, bên thứ 2 hỗ trợ nhiệt tình về pháp lý, thẩm định, vay vốn.được hố trợ, thương thảo, đàm phán, cam kết làm việc chính chủ đầu tư , thông tin và pháp lý đưa ra chuẩn. và một số quỹ đất khách sạn khác được cập nhật mới nhất anh chị, quý nhà đầu tư, thiện chí khảo sát tại đà nẵng vui lòng liên hệ để được tư vấn.

1.

Chính chủ cần bán lô đất thổ cư tại địa chỉ thôn trước đông, xã hòa nhơn, huyện hòa vang, đà nẵng. đất thổ cư. đất có thể tách được 2 thửa. đường bê tông rộng 4m, xe ô tô vào tận đất. nằm trong khu vực dân cư, an ninh đảm bảo, phù hợp xây nhà ở hay đầu tư đều rất tốt. đất cách ubnd xã hòa nhơn 2km. cách đường hoàng văn thái 1,5km. gần trường học hòa nhơn và nhiều tiện ích xung quanh...thuận tiện giao thông đi lại. giá bán có thương lượng liên hệ sđt chính chủ 0903.507.139

2.

Chào bán nhà mặt tiền cmt8, phường khuê trung, quận cẩm lệ, đà nẵng dt 4.5x19.8 89.2m2, 90,8 m2, 100m2, kết cấu nhà cấp 4 mới xây, 2 phòng ngủ, 1 wc. đường 30m. hướng cửa chính tây bắc vị trí nằm trên trục đường lớn, khu vực sầm uất, nhiều tiện ích xung quanh, bán kính 1km có trung tâm hội trợ triển lãm, công viên thanh niên, bệnh viện tâm trí, trường học,.. thích hợp mua đầu tư, cho thuê kinh doanh kiếm dòng tiền ổn định hàng tháng giá thương lượng liên hệ mr. lam 0901 178 83chíncty bđs vingrand center chi nhánh tp đà nẵng địa chỉ tt dvtm savico 66 võ văn tần, q.thanh khê, đn

3.

Cần bán căn nhà trung tâm thành phố nội thất 4 sao giá chỉ 4 tỷ 2 mặt đường thoáng , trung tâm hoàng diệu , gần nguyễn văn linh chính chủ cần bán căn nhà 4 tầng đường hoàng diệu quận hải châu . dt 4.5 x 11.5 50m dtsd 156m2 kết cấu 4 tầng tầng trệt 1 pk bếp và phòng ăn , tầng 1 2 pn 1 wc . tầng 2 1 pn 1 phòng thờ sân phơi wc . tầng 3 1 pn 1 wc. tầng thượng khu sân vườn phòng spa hình thật 100 và đúng giá chủ nhà , vui lòng liên hệ trực tiếp để thương lượng giá chính chủ nguyễn hiếu batdongsan43 . vn kênh thông tin bất động sản chính chủ đà nẵng

4.

Căn hộ fpt plaza 2 đà nẵng gồm gói quà tân gia đến 100 triệu tiền mặt hỗ trợ vay ngân hàng 700 triệu giá chỉ từ 1 tỷ 8 , miễn lãi 12 tháng vị trí khu đô thị fpt city , võ quý huân , p . hoà hải , q . ngũ hành sơn , tp đà nẵng quy mô 2 tầng hầm , 25 tầng nổi gồm 700 căn hộ loại căn hộ 2 phòng ngủ 56m2 75m2 sản phẩm căn hộ sở hữu lâu dài , sổ hồng trao tay . bàn giao nhà ở ngay . tiện ích nội khu tầng 1 với 23 căn shophouse tiện ích kinh doanh các mặt hàng , dịch vụ ăn uống và siêu thị . khu vui chơi , café , hồ bơi , phòng gym , yoga , tích hợp ngay trong tầng 2 của tòa nhà.

5.

Bán nhà gỗ liêm giang châu 1 100m2 khuê mỹ, đà nẵng diện tích 5x20 100m2. đường 7m5. vỉa hè 4m. nhà 3 tầng 3 pn 3wc để lại nội thất cơ bản xây dựng hoàn thanh năm 2018 gần sông mát mẻ chủ ở sài gòn cần bán gấp. giá 6 tỷ còn bớt liên hệ xem nhà làm việc chính chủ

6.

b. Entity Definition

STT	Entity Name	Description
1	LOAI_BDS	Type of real estate
2	HUONG_NHA	Direction of property/house
3	MOI_TRUONG_SONG	Population density around the property, convenience for business
4	VIA_HE	Presence of front yard
5	LOAI_HEM	Type of alley of the property
6	THOAT_NUOC	Water drainage capability of the property
7	HINH_DANG_DAT	Shape of the house
8	NOI_THAT	Condition of interior furniture
9	KV_DE_XE	Parking area
10	STHUONG_BCONG	Roof or balcony
11	PHAP_LY	Legal status
12	AN_NINH	Security in the area
13	DAN_TRI	Educational level of surrounding residents
14	PHONG_THO	Presence of worship room
15	SAN_VUON	Presence of garden

1. The definition of entities is extracted using predefined rules

STT	Entity Name	Description
1	GIA	Price of the real estate
2	DIA_CHI	Address of the real estate
3	DIEN_TICH	Area
4	VI_TRI	Location of the real estate
5	DUONG_TRUOC_NHA	Road in front of the house
6	VI_TRI_SO_VOI_BIEN	Proximity of the real estate to the sea
7	VI_TRI_SO_VOI_SONG	Proximity of the real estate to the river
8	SO_PHONG_NGU	Number of bedrooms
9	SO_TOILET	Number of toilets
10	SO_TANG	Number of floors
11	DICH_VU	Nearby services
12	KICH_THUOC	Dimensions of the real estate

2. The definition of entities is extracted using model

4. Experiments

4.1. Experimental Method

After the data preparation and statistical analysis, it was observed that the data distribution among classes was imbalanced. Additionally, certain ambiguities were identified in real-world cases concerning the definitions. Therefore, the research project was conducted and evaluated using various methods but remained focused on addressing three key issues: imbalance, semantic challenges, and definitional ambiguities.

a. Focal loss

Focal loss was introduced as a variant of cross-entropy to mitigate the issue of data imbalance in the distribution when training classification models. The concept behind the method is to assign higher weights to observations belonging to the minority class.

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma} log(p_t)$$

b. Dice loss

The goal of this method is similar to that of focal loss; however, instead of relying on weights, Dice loss is considered an improved version of the F1-score.

$$DSC(A,B) = \frac{2|A \cap B|}{|A| + |B|}$$

c. Intrust loss

Unlike Dice loss or Focal loss, Instruct loss is used to reduce the risk of overfitting to noisy data in the model. However, this method is also built upon the cross-entropy loss function.

$$L_{in-trust} = \alpha L_{CE} + \beta L_{DCE}$$

d. Label smoothing

Label Smoothing is a regularization technique commonly used to address the overfitting issue. The method works by adding noise to the labels during the training process, helping the model to be 'less confident' in its predictions. As a result, the model focuses more on learning the key features of the data and avoids memorizing irrelevant details.

$$y_k^{LS} = y_k(1-\alpha) + \alpha/K$$

e. Data Augmentation

My approach in this method will be based on noise-inducing techniques: truncating the sentence length, deleting/adding certain elements within the sentence, slightly changing the position of some elements, and so on.

4.2. **Convention:** GPU: GeForce RTX 3090, CUDA 11.4 and phobert-base-v2.

4.3. Noisy classification - ***This step is carried out with only one approach based on the model.***

	Precision	Recall	F1
Base	0.93	0.85	0.88
Dice loss	0.93	0.87	0.90
Focal loss	0.94	0.89	0.91
Intrust loss	0.95	0.87	0.90
Label smoothing	0.95	0.84	0.88
DA	0.86	0.86	0.86

➔ Although each experimental method has its own strengths and weaknesses, the project prioritizes f1-macro, and thus ***Focal loss*** was chosen for implementing the noisy classification.

4.4. NER - ***This step is carried out with parallel rules and models.***

a. Model

1. Base

	Precision	Recall	F1	Support
BIEN	0.94	0.94	0.94	308
DIA_CHI	0.94	0.99	0.97	787
DICH_VU	0.90	0.90	0.90	426
DIEN_TICH	0.95	0.92	0.94	1918
DUONG_TRUOC_NHA	0.96	0.98	0.97	383
GIA	0.97	0.99	0.98	555
KICH_THUOC	0.89	0.98	0.94	362
PHONG_NGU	0.90	0.97	0.93	348
SONG	0.84	0.90	0.87	467
SO_TANG	0.82	0.94	0.88	472
TOILET	0.94	0.91	0.93	325
VI_TRI	0.88	0.97	0.92	466
micro avg	0.92	0.94	0.93	6817
macro avg	0.91	0.95	0.93	6817
weighted avg	0.92	0.94	0.93	6817

2. Focal loss

	Precision	Recall	F1-score	Support
BIEN	0.95	0.83	0.89	308
DIA_CHI	0.97	0.93	0.95	787
DICH_VU	0.92	0.89	0.9	426
DIEN_TICH	0.93	0.92	0.92	1918
DUONG_TRUOC_NHA	0.97	0.96	0.97	383
GIA	0.99	0.89	0.94	555
KICH_THUOC	0.94	0.96	0.95	362
PHONG_NGU	0.91	0.95	0.93	348
SONG	0.88	0.88	0.88	467
SO_TANG	0.88	0.89	0.88	472
TOILET	0.83	0.92	0.87	325
VI_TRI	0.88	0.93	0.90	466
micro avg	0.92	0.91	0.92	6817
macro avg	0.92	0.91	0.92	6817
weighted avg	0.93	0.91	0.92	6817

3. Label smoothing

	Precision	Recall	F1-score	Support
BIEN	0.95	0.93	0.94	308
DIA_CHI	0.95	0.99	0.97	787
DICH_VU	0.88	0.93	0.90	426
DIEN_TICH	0.90	0.96	0.93	1918
DUONG_TRUOC_NHA	0.96	0.99	0.97	383
GIA	0.98	0.98	0.98	555
KICH_THUOC	0.92	0.96	0.94	362
PHONG_NGU	0.88	0.95	0.91	348
SONG	0.91	0.83	0.87	467
SO_TANG	0.88	0.90	0.89	472
TOILET	0.93	0.88	0.91	325
VI_TRI	0.88	0.94	0.91	466
micro avg	0.92	0.95	0.93	6817
macro avg	0.92	0.94	0.93	6817
weighted avg	0.92	0.95	0.93	6817

➔ Although each experimental method has its own strengths and weaknesses, the project prioritizes F1-macro and thus selects ***Focal loss*** as the main method for implementing the extraction model.

5. Data Normalization

This is also the final module implemented by the project, aiming to standardize the extracted data to a single value in order to build input for future development tasks. For example, expressions like "500m to the beach" and "walk 500m to the beach," or "3ty4" and "3 billion 400 million" will be standardized into a single representation. This module is applied to entities extracted by the model, especially for the two entities **KICH_THUOC** and **DIA_CHI**, where the values are further split to optimize information retrieval as follows:

- **KICH_THUOC:** This is split into two entities, CHIEU_DAI (length) and CHIEU_RONG (width), in the normalization module. For example, "the size of the house is 5 x 20 meters" → CHIEU_DAI = 20 and CHIEU_RONG = 5.
- **DIA_CHI:** This is also divided into three sub-entities: QUAN/HUYEN (district), PHUONG (ward), and DUONG (street). For example, "hai thuong lang ong street, phuoc ning ward, HaiChau district, Da Nang" → QUAN/HUYEN: Hai Chau, PHUONG: Phuoc Ninh, and DUONG: Hai Thuong Lang Ong.