Expert Systems With Applications 213 (2023) 119147

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

# Automated real estate valuation with machine learning models using property descriptions

Katharina Baur *, Markus Rosenfelder, Bernhard Lutz

*Chair for Information Systems Research, University of Freiburg, Rempartstr. 16, 79098 Freiburg, Germany*

ARTICLE INFO

*Keywords:*
Real estate valuation
Text mining
Machine learning
Explainable artificial intelligence
Decision support

ABSTRACT

Automated and accurate real estate valuation benefits buyers and sellers in real estate markets. So far, the literature on expert systems for real estate valuation has primarily focused on structured features like the age of the building or the number of rooms. The description of the property presents another rich source of information, which received comparably less attention. In this study, we evaluate several machine learning models in predicting real estate prices using different numeric representations of the property descriptions. Our empirical evaluation, based on rental apartments offers in Berlin ($N$ = 30,218) and house purchase offers in Los Angeles ($N$ = 33,610), shows that the best approach achieves mean absolute errors (MAE) of 1.01€ monthly rent per square meter and 114.84$ per square foot, respectively. Including the property description into the best model reduces the MAE by up to 17.09 percent over the respective baseline models. In addition, we find that the benefit of including textual features of real estate descriptions only weakly depends on the description length. However, the benefit is comparatively less pronounced for rental apartment offers of low prices per square meter. We finally shed light on how the models arrive at decisions by visualizing description embeddings and presenting Shapley additive explanations.

## 1. Introduction

Real estate markets have been growing steadily over the past decade. Since 2010, the prices for rent and property in the USA and Germany have increased by 49.3 and 54.3 percent, respectively.[1] A fair and efficient real estate market requires accurate valuations of real estate (Kofner, 2014; Zhao et al., 2011). However, accurate real estate valuation, in turn, requires human expert knowledge and the process of manual real estate valuation is time consuming (Kok et al., 2017). Accordingly, real estate market participants could benefit from automated tools that allow an accurate estimation of real estate in a short time.

Prior studies have produced several models to automatically estimate the value of real estate property (Pavlov, 2000; Rosen, 1974; Sirmans et al., 2005). However, the proposed models mainly rely on numerical features, such as area, age, or the number of rooms of the property, while neglecting the textual description. As text provides a rich source of information (Gentzkow et al., 2019), the predictive models may benefit from additional features based on the description of the property. For instance, the description may contain important information that is not reflected in the structured attributes. This could include the neighborhood, past renovations, and the distance to cafes, restaurants, or public transport. The few existing studies (Goodwin et al., 2014; Liu et al., 2020; Nowak et al., 2019; Nowak & Smith, 2017) that have considered textual descriptions of real estate listings focus mainly on establishing causal relations between the textual content and the valuation of the property. For instance, Nowak et al. (2019) proposed an approach based on regularized regression models to extract words from English real estate descriptions with a pronounced positive or negative influence on the real estate price. However, none of the aforementioned studies have specifically focused on optimizing machine learning models in predicting real estate valuations by using the textual description.

In the last decade, the field of natural language processing (NLP) has been predominantly shaped by applications of word embeddings (Le & Mikolov, 2014; Mikolov et al., 2013). The basic idea of these methods follows the distributional hypothesis, which states that words that occur at similar positions have a similar meaning (Harris, 1954). Approaches for generating word embeddings are based on neural networks, which are trained to predict the next word given a window of context words. The internal hidden state of the neural network then presents the word

* Corresponding author.
*E-mail addresses:* katharina.baur@vwl.uni-freiburg.de (K. Baur), markus.rosenfelder@is.uni-freiburg.de (M. Rosenfelder), bernhard.lutz@is.uni-freiburg.de (B. Lutz).

[1] See the OECD statistic https://doi.org/10.1787/63008438-en.

https://doi.org/10.1016/j.eswa.2022.119147
Received 21 June 2022; Received in revised form 19 October 2022; Accepted 23 October 2022
Available online 28 October 2022
0957-4174/

**Table 1**
Existing studies on real estate valuation using textual descriptions.

| Study | Real estate type(s) | Model(s) | Objective |
|---|---|---|---|
| Liu et al. (2020) | house purchases | LASSO | find influential words |
| Nowak and Smith (2017) | house purchases | LASSO | use words from listings for valuation |
| Nowak et al. (2019) | house purchases | LASSO | find influential words |
| Goodwin et al. (2014) | house purchases | Linear regression | examine effect of specific wording |
| This study | rental apartments, house purchases | Gradient Boosting, SVM, Elastic Net, Linear Regression | minimize predictive error |

embedding. Thereby, words can be represented in a dense representation that also captures semantic similarities (Young et al., 2018). Specifically, words with a more similar meaning are assigned geometrically closer vectors than unrelated words (Yao et al., 2018). Over the past few years, context-sensitive embeddings have been developed to additionally capture the polysemic nature of words, which refers to the fact that one word may reflect different meanings in different contexts (Vicente & Falkum, 2017). Word embeddings from BERT (Bidirectional Encoder Representations from Transformers, Devlin et al., 2019) present the current state-of-the-art method of how to represent text. Given these recent advances in NLP and the aforementioned considerations about potentially useful information in property descriptions, it appears promising to apply current NLP methods on the property description to build automated models for real estate valuation. In this study, we evaluate several machine learning models based on different text representations of property descriptions in predicting real estate prices. Thereby, we aim to find the best predictive model, while also quantifying the benefit of including textual descriptions in regard to how much the predictive error can be reduced.

Our numerical evaluation based on 30,218 rental offers from Berlin and 32,610 house offers from Los Angeles shows that a gradient boosting model with BERT embeddings achieves the lowest predictive errors. The relative reduction in the predictive error of including the textual descriptions reaches 17.09% on rental offers and 5.66% on house prices. We also analyze the benefit on including textual descriptions in regard to the description length and the price of the real estate object. While we found similar error reductions for small, medium, and long descriptions, our results indicate that rental apartment offers benefit most if they are of medium or high price. Our findings are relevant for researchers and practitioners in the field of real estate appraisal.

To the best of our knowledge, this study is the first to systematically analyze and quantify the benefits of including real estate descriptions in predictive models. Real estate buyers can use the presented approach as decision support for their purchases, while real estate sellers can evaluate the influence of different descriptions on the estimated price. In addition, websites for real estate listings can present more accurate estimates of real estate valuation.

The remainder of this paper is organized as follows. Section 2 provides an overview of related work. Section 3 details the data and methods used for our evaluation. Section 4 presents the results and Section 5 discusses our findings.

## 2. Related work

Prior studies have established hedonic pricing models as the standard method in valuating real estate since the 1970s (Rosen, 1974). The essential proposition of hedonic pricing theory is that real estate market participants value apartments and houses according to the total of their relevant attributes (Pavlov, 2000). These attributes comprise the internal and external attributes of the real estate property (Sirmans et al., 2005). The standard hedonic pricing function can be described as

$$P = P(X_1, X_2, \dots, X_n), \tag{1}$$

where $X_i$ represents the value of the $i$th attribute. In terms of properties, several categories of housing attributes may be distinguished. Features relating to the appearance of the housing itself, like the living area, the number of rooms, or the year of construction, are bundled into the group of structured characteristics and are most frequently investigated in traditional housing literature (Antipov & Pokryshevskaya, 2012). The standard modeling approach for estimating real estate prices within the hedonic pricing framework is a linear regression model based on specific housing characteristics (Rosen, 1974). However, this approach exhibits several shortcomings as it cannot account for non-linearity and unstructured data like text. Moreover, existing research has already demonstrated that non-linear machine learning models yield high-quality rental predictions and, in particular, more accurate valuations than hedonic regression models (Antipov & Pokryshevskaya, 2012; Park & Bae, 2015; Pérez-Rave et al., 2019; Rico-Juan & Taltavull de La Paz, 2021).

Only a few studies use textual descriptions of real estate to estimate real estate prices (see Table 1). Goodwin et al. (2014) retrieve the number of positive words and specific market signals within the listing. They conclude that real estate descriptions should be carefully worded to achieve the optimal price. Liu et al. (2020) incorporate textual information into hedonic pricing models. The authors find that the textual variables reveal an omitted variable bias that was not accounted for in prior studies, which led to incorrect conclusions. Another method for including text data in traditional hedonic pricing models is to combine a bag of words approach with principal component analysis (Nowak & Smith, 2017). Nowak and Smith (2017) find that adding text features in linear models can reduce the root mean squared error significantly for their dataset. Nowak et al. (2019) use a LASSO model to select the top-500 terms with the largest influences on the property valuation from the overall bag-of-words representation. The most important terms are then included as part of a linear regression model for real estate appraisal.

In this paper, we specifically focus on including the textual descriptions with the intention of minimizing the predictive error of machine learning models. For this purpose, we evaluate several machine learning models and text representations.

## 3. Dataset and methods

The expert system for real estate valuation is illustrated in Fig. 1. The raw data from real estate listings consists of structured features and the textual description. The property description is first transformed into a numerical representation using different methods for text representation. All features are then processed by a predictive model, which outputs the estimated price of the real estate object. The output can subsequently be used by real estate market participants in their purchase decisions. In addition, real estate agents can use the output as writing guidelines to analyze how different descriptions influence the estimated price.

### 3.1. Datasets

We consider two datasets of real estate listings as they present an established data source for research on real estate pricing (Antipov & Pokryshevskaya, 2012; Ho et al., 2015; Nowak et al., 2019). Specifically, we consider rental apartment offers from Berlin and the

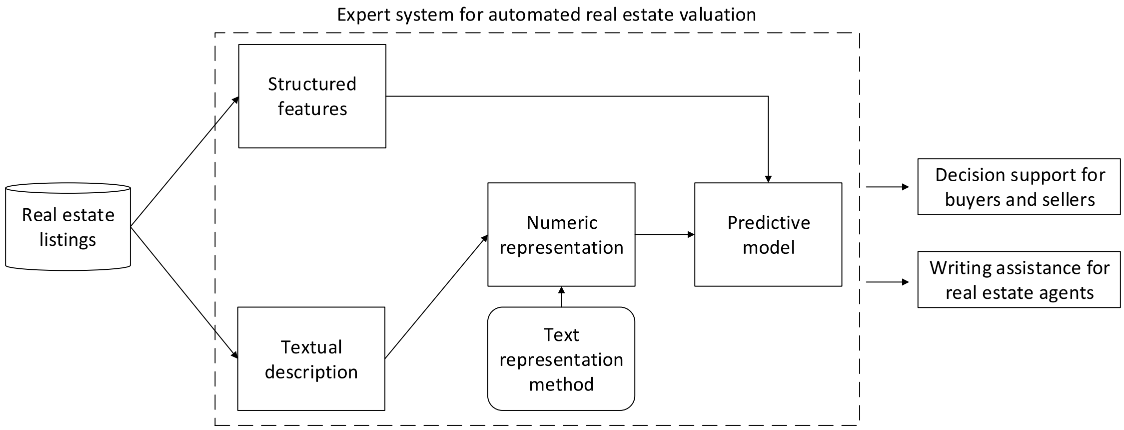



**Fig. 1.** Illustration of expert system for real estate valuation.

house purchase offers from Los Angeles. Accordingly, the dependent variables of interest are the monthly rent in € per square meter (sqm) in the Berlin dataset and the house price in $ per square foot (sqft) in the Los Angeles dataset. Note that rent and sales are positively correlated, so that higher rent prices entail greater valuation of the property (Bourassa et al., 2019). The Berlin dataset spans the time period of April 2019 to May 2019, while the Los Angeles spans June 2020 to July 2020. The Berlin dataset is retrieved from the real estate platform Immobilienscout24 and the Los Angeles dataset is retrieved from Zillow. Both data sources are common choices in related studies (e.g., Baldauf et al., 2020; Scholz et al., 2017). Every observation in our data represents a real estate property on offer, which is advertised on the respective real estate platform. Since both platforms use fraud detection and human moderation, fraudulent data should not be an issue for our analyses.

**Table 2**
Numerical and categorical housing attributes.

| Feature | Rental apartment offers (Berlin) | House purchase offers (Los Angeles) |
|---|---|---|
| Numerical | | |
| Living area | ✓ | ✓ |
| Number of rooms | ✓ | ✗ |
| Number of bedrooms | ✗ | ✓ |
| Number of bathrooms | ✗ | ✓ |
| Age of building (years) | ✓ | ✓ |
| Lot size | ✗ | ✓ |
| Floor of apartment | ✓ | ✗ |
| Longitude, Latitude | ✓ | ✓ |
| Categorical | | |
| Apartment type | ✓ | ✗ |
| Condition | ✓ | ✗ |
| Ground plan | ✓ | ✗ |
| Heating type | ✓ | ✗ |
| Certificate type | ✓ | ✗ |
| Textual | | |
| Property description | ✓ | ✓ |
| Description length | ✓ | ✓ |

Table 2 provides an overview of all features. The Berlin dataset has 13 features, while the Los Angeles dataset has 9 features. Both datasets contain features about the living area, age of the building, longitude and latitude, the property description and the description length. The Berlin dataset has the monthly rent per sqm, number of rooms, floor of the apartment, apartment type, apartment condition, ground plan, heating type, and certificate type as unique features. The variable values of the categorical variables are reported in Appendix A of the supplementary material. The price per sqft, number of bed- and bathrooms, as well as the lot size only occur in the Los Angeles dataset.

The textual description is presented differently in the datasets. While the Los Angeles dataset has one description text per observation, the Berlin dataset has three textual descriptions serving different purposes. The first description in the Berlin dataset contains the general information about the apartment. The second description presents information about the equipment of the apartment, and the third describes the location.

### 3.2. Descriptive statistics

Table 3 shows the descriptive statistics of both datasets. The average monthly rent in Berlin is 11.56€ per sqm, the average living area is 37 sqm with 2.37 rooms. The average apartment is on the second floor and 64 years old. The average price of a house is Los Angeles is 583.04$ per sqft. The average house has 3.38 bedrooms and 2.63 bathrooms, an area of 2,030 sqft, and is 56 years old. The descriptions of Berlin rent offers are on average 42 words longer than the descriptions of houses in Los Angeles.

Fig. 2 presents histograms of the description lengths for both datasets. For the Los Angeles dataset, we observe that descriptions between 100 and 150 words occur with the highest relative frequency. Moreover, 98.29% of all descriptions contain less than 250 words. For the Berlin dataset, we also find that property descriptions between 100 to 150 words are most frequent. However, 16.38% of all German descriptions have more than 250 words. We further observe a notable difference in the total vocabulary used in the descriptions. In the Berlin property descriptions, we have 45,620 unique words and in the Los Angeles dataset we have 26,671 unique words.

### 3.3. Preprocessing

We perform the following preprocessing steps. First, we remove all observations with implausible features, like a number of rooms smaller than 1 or a negative living area. Appendix B of the supplementary material provides more information on which ranges are set. Second, we remove all observations with more than 25% missing data. For all remaining features, we impute the mean to fill missing values. The filtered Berlin dataset contains 30,218 observations and the filtered Los Angeles dataset contains 33,610 observations.

In addition, we transform all descriptions to lower-case and we remove all non-alphabetical characters, including numbers and punctuation. We finally generate an additional feature as the length of the description in words.

**Table 3**
Descriptive statistics.

| | Rental apartment offers (Berlin, N = 30,218) target: monthly rent in € per sqm | | | | House purchase offers (Los Angeles, N = 33,610) target: price in $ per sqft | | | |
|---|---|---|---|---|---|---|---|---|
| | Mean | SD | Min | Max | Mean | SD | Min | Max |
| Target variable | 11.56 | 3.74 | 4.00 | 42.31 | 583.04 | 301.23 | 101.63 | 3777.42 |
| Living area[a] | 0.37 | 0.34 | 0.12 | 4.82 | 20.30 | 10.72 | 2.00 | 100.00 |
| Number of rooms | 2.37 | 1.01 | 1.00 | 15.00 | – | – | – | – |
| Number of bedrooms | – | – | – | – | 3.38 | 0.98 | 1 | 13.00 |
| Number of bathrooms | – | – | – | – | 2.63 | 1.27 | 0.50 | 9.00 |
| Age of building (years) | 64.19 | 42.90 | 0.00 | 191.00 | 56.34 | 25.98 | 0.00 | 196.00 |
| Lot size[a] | – | – | – | – | 80.75 | 40.74 | 6.01 | 249.99 |
| Floor of apartment | 2.55 | 1.91 | −1.00 | 23.00 | – | – | – | – |
| Description length[b] | 1.72 | 0.96 | 0.02 | 10.62 | 1.30 | 0.59 | 0.07 | 6.35 |

[a]In 100 square meters or square feet.
[b]In 100 words.

### 3.4. Text representations

A main challenge in text mining is to transform text into a numerical representation that can be processed by predictive models. We make use of multiple text representation methods to assess the predictive performance of real estate texts on the price. Specifically, we consider the bag-of-words method and a dictionary approach according to Nowak et al. (2019). Both extract single words from the descriptions and include these words as features in the prediction. Additionally, we derive advanced text representations with doc2vec (Le & Mikolov, 2014) and BERT that represent each description in their entirety instead of as single words. To the best of our knowledge, previous research has not considered such representations on document level for the purpose of predicting housing prices but rather has focused on the extraction of single words from real estate texts.

*Bag-of-words representations*

The basic idea behind the bag-of-words approach is to count the number of incidences of a word within a text. For this approach, we limit the total word corpus to the 1,000 most frequent words occurring over all descriptions. These words are then represented as a sparse matrix indicating the number of term occurrences within each property description. We include the word counts as additional features in our prediction models. A major drawback of this approach is that it does not account for the word ordering and thus may not accurately capture the relationship between words (Gentzkow et al., 2019). For instance, the sentences “this apartment has a balcony but no basement” and “this apartment has a basement but no balcony” would result in the same matrix representation despite their opposite meaning.

*LASSO approach*

Nowak et al. (2019) extend the classical bag-of-words approach by performing variable selection over the sparse matrix to identify a real estate dictionary from property descriptions. Essentially, this approach first erects a bag-of-words matrix from the $N$ most frequent words in the overall real estate text corpus. Then, a cross-validated and heteroscedastic LASSO regression is performed on the price where the LASSO shrinkage parameter is imposed on the regression coefficients of the words. The LASSO penalty term is estimated using 5-fold cross-validation. LASSO enables variable selection since some regression coefficients of the words in the matrix that are less associated with the price are set to 0 and further removed from the matrix (Nowak et al., 2019). We follow this approach to generate a word list with the most relevant words which we then include in our predictive models. We use cross-validated LASSO as this achieves slightly better results.

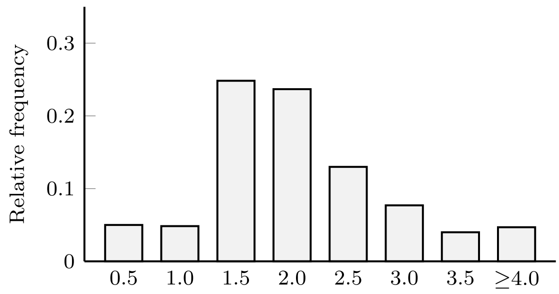
*Static word embeddings*

The purpose of word embeddings is to learn a high-dimensional vector representation for each word, so that a word vector can be put in relation to others. The original idea behind word embeddings is to train a neural network to predict the missing word given its context words. The internal state of the neural network then presents the word embedding (Mikolov et al., 2013). Thereby, the resulting word embeddings are able to carry semantic meaning (Yao et al., 2018). Vector representations of words with a similar meaning are geometrically closer to each other than vectors of different words. Depending on the embedding model, the resulting word embeddings can be either context-free or context-sensitive. Static methods produce a global context-free embedding for each word, meaning that different contexts of a word are not reflected. The representation of a given word encapsulates the meaning of each instance of the word in the whole text corpus. For instance, consider the word “property” in the phrases “common property” and “real estate property”. Static or context-free embedding models like word2vec (Mikolov et al., 2013) output the same embedding for “property”. We employ doc2vec, which is built on word2vec, to generate fixed-length vector representations for each property description. Doc2vec (Le & Mikolov, 2014) extends the network architecture of word2vec by training a paragraph vector in the input layer, which stores the missing information of the current context not represented by the sampled input words. This vector then presents the embedding of the entire document (Le & Mikolov, 2014).

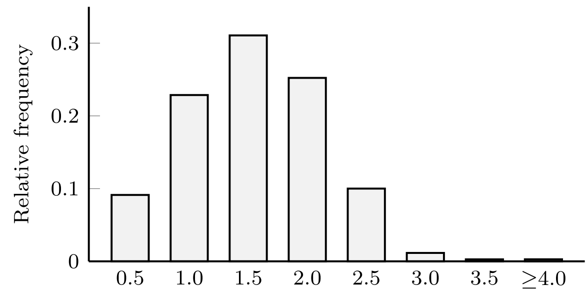
*Contextualized word embeddings*

We also make use of the BERT model (Devlin et al., 2019), which presents the current state-of-the-art method for text representation. The BERT model generates contextualized text representations, so that the embedding of a word depends on the context of the word in its sentence. In the above example, context-based models will generate a different, contextualized embedding for the word “property”. Since many words are polysemic (Vicente & Falkum, 2017), contextualized text representations provide more accurate language representations.

The BERT model is based on bidirectional transformers. Transformer architectures are attention-based sequence-to-sequence models (Vaswani et al., 2017), which means that input and output are given in the form of sequences. The structure of general transformers consists of an encoder and decoder. The encoder maps the input sequence into a fixed-size vector representation, while the decoder converts this representation into an output sequence. This architecture is, for instance, useful in tasks like language translation. A major advantage of transformer architectures over previous language models is that they can also handle long-term dependencies between words in a sequence (Cho et al., 2014; Trinh et al., 2018). “Bidirectional” refers to the fact that an element of the input sequence is processed in regard to its predecessor and successor. The BERT architecture is trained based on two objectives: next sentence prediction (NSP), and




(a) Rental apartment offers (Berlin)




(b) House purchase offers (Los Angeles)

**Fig. 2.** Histograms of description lengths (in 100 words) for both datasets.

masked language modeling (MLM). For MLM, BERT randomly masks 15% of the words in each input sequence and uses the context words to predict the masked word. This enables a bidirectional representation by simultaneously looking at both sides of the input word. Similarly, the NSP task focuses on learning relationships between two pairs of sentences. Given two input sentences, it aims to predict whether one sentence follows another (Devlin et al., 2019).

For the purpose of our study, we use two different pretrained versions of the BERT model as our datasets differ in language. Both the German and English models are pretrained on large open-domain text sources in either German or English.[2] To derive a fixed embedding for each description, we apply a mean-pooling layer to the output of the BERT model. This yields fix-sized vector representations of 768 dimensions.

**Table 4**
Comparison of text representation methods.

| Method | Interpretability | Computational effort | Performance |
|---|---|---|---|
| Bag-of-words | +++ | +++ | – |
| LASSO | ++ | ++ | + |
| Static embeddings | – | – | ++ |
| Contextualized embeddings | –– | –– | +++ |

The main differences of the considered text representation methods are summarized in Table 4. While embedded text representations generally achieve the best predictive performance when being used as features of machine learning models, they are much harder to interpret than bag-of-words representations.

### 3.5. Predictive models

We evaluate several predictive models; namely, linear regression, elastic net, support vector regression, random forest, and a gradient boosting algorithm (LightGBM) (Ke et al., 2017). We choose these models due to their strong performance in prior studies about real estate appraisal (Kok et al., 2017; Rico-Juan & Taltavull de La Paz, 2021). Furthermore, these models allows us to shed light on explainability in terms of Shapley additive explanations (Lundberg & Lee, 2017) due to their low complexity.

In addition, we evaluate several simple baselines, including the overall mean, mean per real estate type, and the mean per quarter. These baselines do not use any numerical or textual features and are solely based on different averages. As the real estate type is unique in the Los Angeles data, this baseline only exists for the Berlin dataset.

The data is randomly split into 70% training and 30% test set. The hyperparameters are selected on the training dataset with 5-fold cross validation and grid search. The grid ranges as well as the optimal hyperparameter choices for both datasets are provided in Appendix C of the supplementary material. Each model is trained with and without the numeric representation of the descriptions so that we can assess the influence of including the text representations in the predictive models. The out-of-sample performance is then measured based on the test set.

### 3.6. Performance metrics

In concordance with previous work (Nowak et al., 2019), we use root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) for evaluating the predictive performance of all considered approaches.

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}. \quad (2)$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_i\right|. \quad (3)$$

$$\mathrm{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right|. \quad (4)$$

For each performance metric (2)–(4), $y_i$ denotes the true price of real estate property $i$ and $\hat{y}_i$ the estimated price of for apartment $i$. $N$ denotes the total number of observations in the dataset. All evaluation metrics are reported for the test data. We choose MAE as our primary performance metric as it is easy to interpret.

We test the differences in predictive performance between the model with solely tabular data and each model including text data with the Diebold–Mariano test (Diebold & Mariano, 2002). Moreover, we make use of the Model Confidence Set by Hansen et al. (MCS, 2011). The MCS approach extends classical statistical testing as it tests for the superiority of several models by taking each model at a time as the benchmark model. This procedure then erects a set of superior models where the hypothesis of equal predictive ability cannot be rejected within this set. Hence, this procedure allows equally superior models to be identified. All models within the superior set are marked with * in the results tables. Detailed information on the MCS is provided in Appendix D of the supplementary material.

## 4. Results

### 4.1. Main analyses

We start by presenting the prediction results for rental apartment offers in Berlin and house purchase offers in Los Angeles. We report the estimation errors of rent in € per sqm and price in $ per sqft. For each model, we present the results when using structured features only, and the results when additionally including the real estate description using different approaches to generate the numerical text representations.

[2] See https://huggingface.co/bert-base-german-cased and https://huggingface.co/bert-base-uncased for details.

**Table 5**
Evaluation results for rental apartment offers (Berlin, target: monthly rent in € per sqm).

| Model | Predictive error | | | Error reduction in % when including description | | |
|---|---|---|---|---|---|---|
| | MAE | MAPE | RMSE | MAE | MAPE | RMSE |
| Gradient boosting | | | | | | |
| Data without description | 1.21 | 10.94 | 1.83 | | | |
| BERT | **1.01** | **9.17** | **1.57** | 17.09 * | 16.14 * | 14.10 * |
| Doc2vec | 1.15 | 10.43 | 1.68 | 5.09 | 4.63 | 8.04 |
| Bag-of-words (unigrams) | 1.10 | 9.53 | 1.67 | 9.07 | 12.90 | 8.41 |
| LASSO Dictionary | 1.15 | 10.36 | 1.70 | 5.52 | 5.25 | 6.73 |
| Random forest | | | | | | |
| Data without description | 1.22 | 10.96 | 1.83 | | | |
| BERT | **1.04** | **9.25** | **1.61** | 14.78 * | 15.52 * | 11.72 * |
| Doc2vec | 1.16 | 10.44 | 1.71 | 4.34 | 4.72 | 6.37 |
| Bag-of-words (unigrams) | 1.14 | 10.37 | 1.70 | 6.13 | 5.36 | 6.87 |
| LASSO Dictionary | 1.17 | 10.47 | 1.75 | 3.75 | 4.42 | 4.39 |
| Support vector regression | | | | | | |
| Data without description | 1.54 | 13.49 | 2.37 | | | |
| BERT | **1.09** | **9.95** | **1.72** | 28.95 * | 26.19 * | 27.59 * |
| Doc2vec | 1.21 | 11.00 | 1.77 | 21.26 | 18.44 | 25.60 |
| Bag-of-words (unigrams) | 1.45 | 13.18 | 2.13 | 5.86 | 2.29 | 10.40 |
| LASSO Dictionary | 1.39 | 12.69 | 2.21 | 9.67 | 5.92 | 6.84 |
| Elastic net | | | | | | |
| Data without description | 2.23 | 20.48 | 3.10 | | | |
| BERT | **1.62** | **14.76** | **2.26** | 27.30 * | 27.94 * | 27.24 * |
| Doc2vec | 1.81 | 16.52 | 2.52 | 18.89 | 19.34 | 18.79 |
| Bag-of-words (unigrams) | 1.92 | 17.59 | 2.41 | 13.65 | 14.09 | 22.28 |
| LASSO Dictionary | 1.94 | 17.94 | 2.67 | 12.79 | 12.39 | 13.86 |
| Linear regression | | | | | | |
| Data without description | 2.18 | 19.78 | 3.03 | | | |
| BERT | **1.57** | **14.47** | **2.16** | 27.77 * | 26.84 * | 28.77 * |
| Doc2vec | 1.75 | 15.97 | 2.36 | 19.75 | 19.30 | 22.14 |
| Bag-of-words (unigrams) | 1.74 | 15.44 | 2.46 | 20.10 | 21.96 | 18.76 |
| LASSO Dictionary | 1.93 | 17.70 | 2.62 | 11.44 | 10.53 | 13.58 |
| Mean baselines | | | | | | |
| Mean overall | 2.82 | 26.43 | 3.80 | | | |
| Mean per quarter | 2.23 | 20.12 | 3.18 | | | |
| Mean per type | 2.64 | 24.52 | 3.49 | | | |

$N$ = 9,066.
*Denotes statistical superiority according to the model confidence set.

*Rental apartment offers*

The results for the rental apartment offers from Berlin are shown in Table 5. The best performance is achieved by gradient boosting using the contextualized embeddings from BERT to represent the property descriptions. The predictive error in terms of MAE of gradient boosting combined with BERT is 1.01, which means that the estimated monthly rent per sqm differs on average by 1.01 € from the true value. This corresponds to an error reduction of 17.09% relative to the gradient boosting model without any textual features. The benefit of including BERT embeddings instead of bag-of-words features in the gradient boosting model amounts to a reduction of 8.91% in MAE. The gradient boosting model also achieves the lowest predictive errors over the random forest and support vector regression as well as linear models for each individual text representation. The combination of random forest and BERT embeddings results in an error of 1.04, which is the second best result after the gradient boosting model. This translates to an error reduction of 14.78% compared to the random forest model without any textual data. The support vector regression achieves a MAE of 1.09, which is 8 cents above the MAE of the gradient boosting model. All non-linear models achieve a MAPE of below 10% when using BERT embeddings.

Representing the textual descriptions using BERT embeddings yields the most accurate rent valuations for all predictive models. The finding that BERT outperforms all other text representation methods in terms of more accurate rent predictions is further supported by the results of the MCS implementation. We observe that for each prediction model the unique superior model is the model including BERT embeddings. Considering the most influential words using the LASSO-based dictionary by Nowak et al. (2019) also reduces the predictive errors, however, to a smaller extent than including bag-of-words features of the complete description. We present an overview on the top-30 most influential words from the LASSO-dictionary in Appendix F.

The linear models Elastic Net and linear regression clearly perform worse than the non-linear models as the MAE of the linear models is more than 50 cents higher. In addition, their MAPE values are approximately five percentage points above those of the non-linear models. Given that non-linear effects between different features seem very likely, it is not surprising that the linear models achieve a lower performance. At the same time, it is remarkable that they can also benefit from textual descriptions.

We find that the improvement induced by including textual information is statistically significant, with $p < 0.001$ according to a Diebold–Mariano test for all text representations, as well as, all linear and non-linear prediction models.

The mean baselines perform poorly, with MAE scores of between two and three euros. Predicting the mean price of a quarter still provides the best simple baseline with an MAE of 2.23 € per rented sqm.

**Table 6**
Evaluation results for house purchase offers (Los Angeles, target: price in $ per sqft).

| Model | Predictive error | | | Error reduction in % when including description | | |
|---|---|---|---|---|---|---|
| | MAE | MAPE | RMSE | MAE | MAPE | RMSE |
| Gradient boosting | | | | | | |
| Data without description | 121.68 | 21.82 | 197.97 | | | |
| BERT | 114.84 | 20.73 | **186.79** | 5.62 * | 5.03 * | 5.65 * |
| Doc2vec | 116.04 | 20.91 | 189.10 | 4.64 | 4.16 | 4.48 |
| Bag-of-words (unigrams) | 121.76 | 21.81 | 198.15 | −0.06 | 0.06 | −0.09 |
| LASSO Dictionary | **114.79** | **20.57** | 189.06 | 5.66 * | 5.75 * | 4.50 |
| Random forest | | | | | | |
| Data without description | 124.49 | 23.39 | 205.32 | | | |
| BERT | 116.78 | 21.84 | 191.13 | 6.20 * | 6.62 * | 6.91 * |
| Doc2vec | 119.02 | 22.40 | 196.10 | 4.40 | 4.25 | 4.49 |
| Bag-of-words (unigrams) | 124.54 | 23.37 | 205.41 | −0.03 | 0.09 | −0.04 |
| LASSO Dictionary | **116.79** | **21.80** | **191.10** | 6.18 * | 6.81 * | 6.93 * |
| Support vector regression | | | | | | |
| Data without description | 173.06 | 29.90 | 274.78 | | | |
| BERT | **147.11** | **25.47** | **236.77** | 15.00 * | 14.81 * | 13.83 * |
| Doc2vec | 152.02 | 26.46 | 242.85 | 12.16 | 11.51 | 11.62 |
| Bag-of-words (unigrams) | 173.07 | 29.91 | 274.72 | 0.00 | −0.04 | 0.02 |
| LASSO Dictionary | 150.81 | 25.62 | 246.44 | 12.86 | 14.30 * | 10.31 |
| Elastic net | | | | | | |
| Data without description | 180.00 | 33.95 | 268.93 | | | |
| BERT | **155.28** | **29.17** | **232.18** | 13.73 * | 14.09 * | 13.67 * |
| Doc2vec | 160.57 | 30.43 | 237.98 | 10.80 | 10.36 | 11.51 |
| Bag-of-words (unigrams) | 180.11 | 33.97 | 268.99 | −0.06 | −0.06 | −0.02 |
| LASSO Dictionary | 159.09 | 29.81 | 239.23 | 11.62 | 12.20 | 11.05 |
| Linear regression | | | | | | |
| Data without description | 179.17 | 33.55 | 268.93 | | | |
| BERT | **154.02** | **28.72** | **231.74** | 14.04 * | 14.40 * | 13.83 * |
| Doc2vec | 158.38 | 32.76 | 255.92 | 12.01 | 3.49 | 4.84 |
| Bag-of-words (unigrams) | 180.12 | 33.96 | 268.99 | −0.06 | −0.05 | −0.02 |
| LASSO Dictionary | 157.57 | 29.26 | 238.98 | 12.06 | 12.79 | 11.14 |
| Mean baselines | | | | | | |
| Mean overall | 217.01 | 42.06 | 301.19 | | | |
| Mean per quarter | 205.38 | 40.49 | 292.40 | | | |
| Mean per type | – | – | – | | | |

$N$ = 10,561.
*Denotes statistical superiority according to the model confidence set.

*House purchase offers*

Next, we consider the results for the Los Angeles house purchase offers as shown in Table 6. Again, the gradient boosting model performs best among all predictive models. In addition, we find that the best performance is generally achieved by including contextualized BERT embeddings in the estimation. However, the results are less clear for the gradient boosting and random forest model, where, except for the RMSE of the gradient boosting and the MAE of the random forest with BERT, the error metrices for the LASSO-based dictionary by Nowak et al. (2019) are slightly lower. This finding differs from the analysis on the Berlin dataset, where the error metrics of gradient boosting were lowest when using BERT embeddings. One possible explanation could be that it is easier to extract meaningful keywords for descriptions with smaller vocabularies. The descriptions in the Los Angeles dataset exhibit approximately 20,000 fewer unique words than the descriptions in the Berlin dataset. Moreover, the approach by Nowak et al. (2019) was originally developed for English texts, while the rental apartment offers are written in German.

In general, it seems harder to predict the price per sqft of houses in Los Angeles than the monthly rent prices per sqm in Berlin. The MAPE on the Los Angeles dataset ranges from approximately 20 to 30 percent, while it only ranges from approximately 9 to 18 percent in the Berlin dataset. However, the Berlin dataset contains a larger number of features (13) than the Los Angeles dataset (9), which may explain the difference in MAPE scores. The relative improvement when including textual features of the property description for the best performing model is also lower in the Los Angeles dataset (5.66%) than in the Berlin dataset (17.09%). This seems unexpected given that the Los Angeles dataset has fewer features. At first, one could argue that the descriptions in the Los Angeles dataset are less informative than those of the Berlin dataset. In fact, the description texts in the Los Angeles data are approximately 42 words shorter than the German texts. Nevertheless, apart from the bag-of-words approach, the improvement effects over the models without any textual information are statistically significant with $p < 0.001$.

For all other predictive models and error metrics, the BERT model yields a greater error reduction than all other text representations. This finding is supported by the MCS results, where BERT is always selected within the set of superior models. However, considering the MAE and MAPE of gradient boosting, all error metrices of the random forest, and the MAPE of support vector regression, the hypothesis of equal predictive ability additionally cannot be rejected for the model including text features from the dictionary approach (Nowak et al., 2019), which suggests the joint superiority of these models.

### 4.2. Error reduction for different description lengths

We now analyze whether the error reductions of including the text representation depends on the length of the description. For instance, one could expect that short descriptions only offer little information,

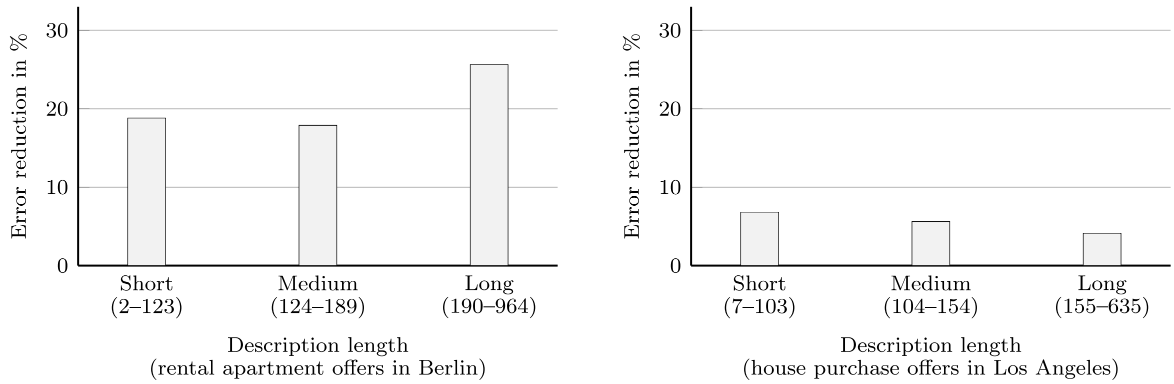



**Fig. 3.** MAE reduction in % when including textual descriptions into gradient boosting model using BERT embeddings for different lengths of property descriptions.

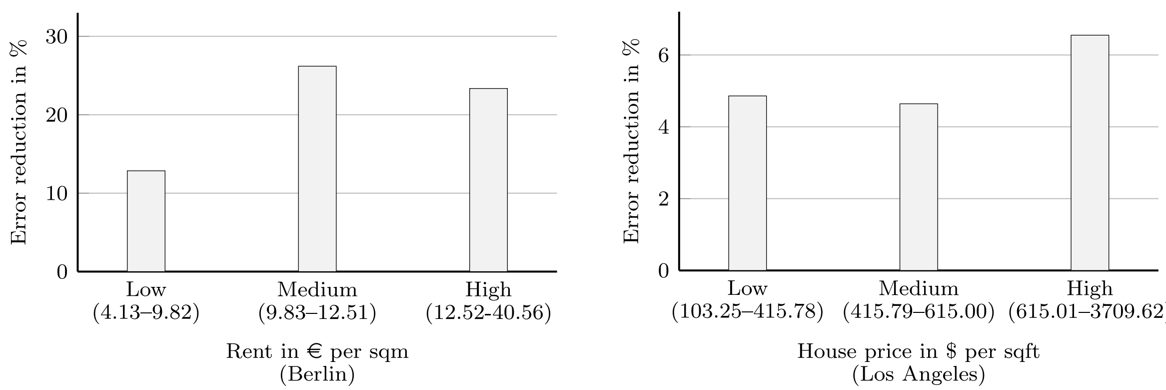



**Fig. 4.** MAE reduction in % when including textual descriptions into gradient boosting model using BERT embeddings for different price segments. The respective boundaries of the price segments are shown in parentheses.

while long descriptions contain a wide variety of additional information that is not reflected in the structured features. To test this idea, we split our datasets into short, medium, and long descriptions according to the respective terciles. We then focus on the gradient boosting model along with textual features from BERT as it performs best in our main analyses. Fig. 3 shows the relative improvement in percent for each subset. The corresponding length boundaries for the classes short, medium, and long are indicated below the bars. For the rental apartment offers, we observe a small U-shaped relation between the relative error reduction and the description length, so that the error reduction is slightly greater for short and long than for descriptions of medium length. The error reduction is greatest for long descriptions, with a relative MAE reduction of 25.63%. For the house prices, we observe smaller error reductions, which is consistent with our main analysis. The error reduction has its peak for short descriptions (6.82%). Altogether, we find few differences across the MAE reduction for short, medium, and long descriptions. This indicates that the benefit of including contextualized text representations in the prediction does not strongly depend on the description length.

### 4.3. Error reduction for different price segments

We also investigate whether the reduction in MAE depends on the price segment of the advertised real estate object. Therefore, we split our dataset according to the terciles of the rent per sqm and the price per sqft into low, medium, and high-priced real estate property offers. Again, we only consider the gradient boosting model as it showed the best overall performance.

Fig. 4 presents the relative reductions in MAE for low, medium, and high-priced real estate and the corresponding interval boundaries. For rental apartments in Berlin, we observe larger variation between the relative improvements across different price segments than for different lengths. The largest error reduction in predicting the price per sqm by including the real estate descriptions into a gradient boosting model is achieved on medium- (26.10%) and high-priced rent offers (23.35%), while the reduction is considerably lower when considering low-priced rental apartments. For house prices, the benefit of including the real description in the gradient boosting model is most pronounced for high-priced houses (6.55%). Like in our prior analyses, the overall benefit of including the textual description is smaller for house prices in Los Angeles than for rental apartments in Berlin.

### 4.4. Error reduction when omitting numerical features

We now perform additional analyses on both datasets to understand how the textual descriptions are linked to numerical features, i.e., to which degree the descriptions contain information that reflects other numerical attributes or additional useful information about the property that goes beyond the numerical attributes. Therefore, we omit each numerical feature individually and then assess the MAE of the gradient boosting model in two configurations, namely, (i) the model without the feature and without the description embedding against (ii) the model without the feature but with the description embedding. The different reductions between excluding both features and including the description while excluding the particular feature then provides additional insights about the benefit of textual descriptions given that data is missing.

The results are shown in Figs. 5 and 6 for the Berlin and Los Angeles dataset, respectively. Detailed prediction results can be found in Appendix E. We observe a greater benefit of including textual

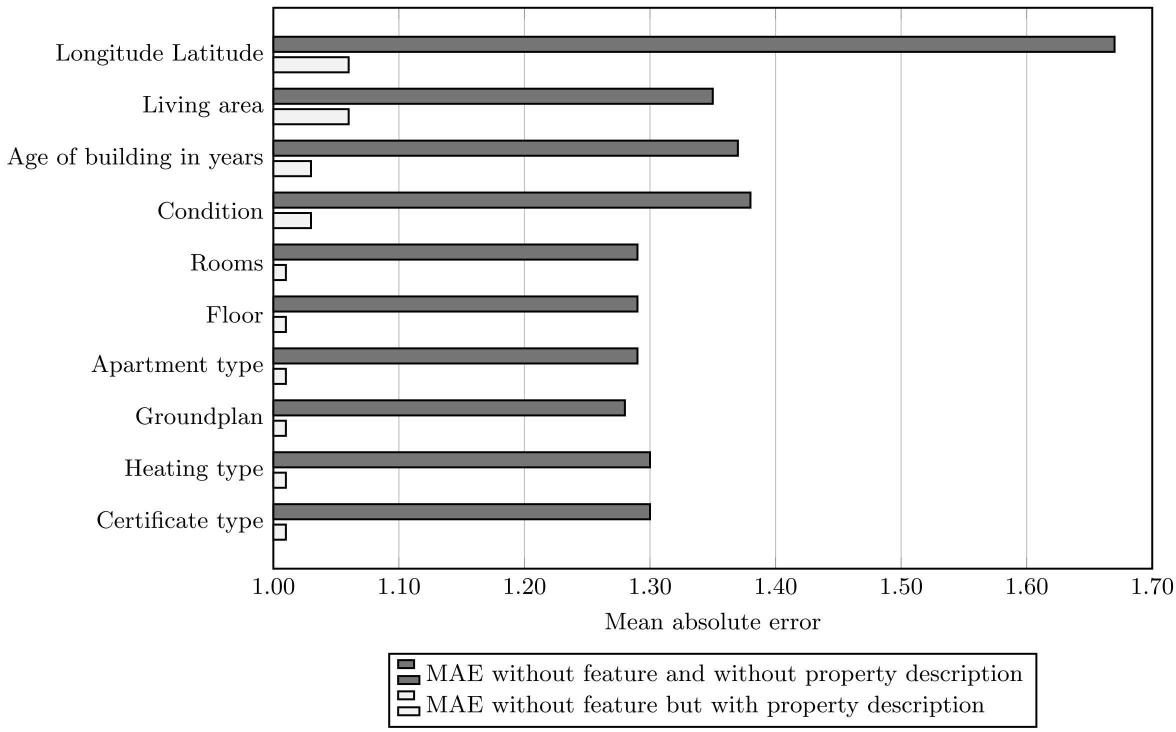



**Fig. 5.** Predictive error (MAE) when omitting different variables (Berlin dataset).

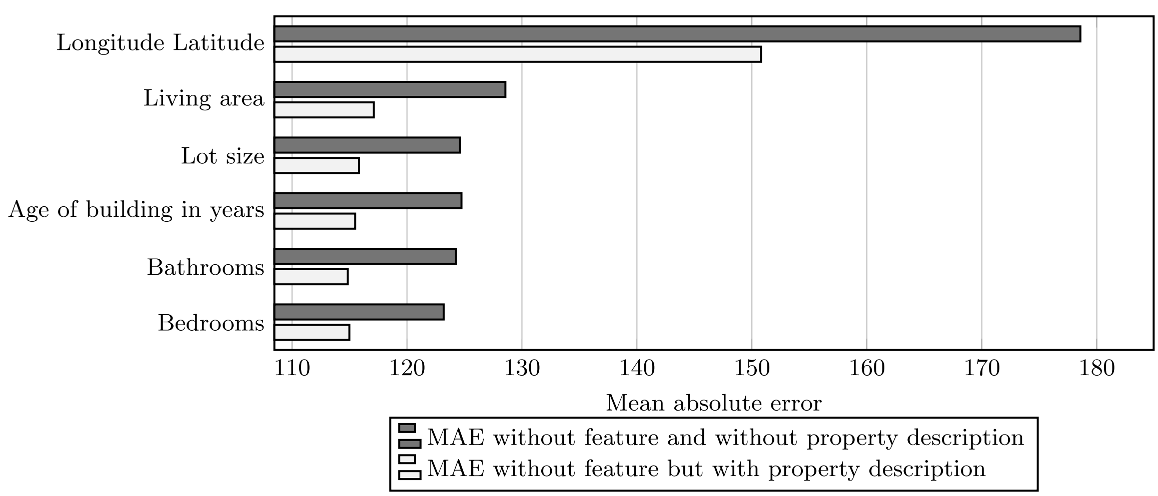



**Fig. 6.** Predictive error (MAE) when omitting different variables (Los Angeles dataset).

descriptions to missing attributes in the Berlin dataset than in the Los Anageles dataset. For instance, the MAE equals 1.67 in the Berlin dataset when the location-based features (Longitude and Latitude) *and* the description are excluded. (Recall that the target variable of the Berlin dataset is the monthly rent per square meter in euros). However, when the description embedding is added while the location-based features remain excluded, the MAE is reduced to 1.06, which indicates a relative reduction of 36.53%. This is much more than the relative error reduction in the LA dataset when the textual description is added while location-based features are missing (15.56%). In particular, both relative reductions are greater than the relative reductions in MAE for the best performing model when all features are present (17.09% in Berlin dataset, 5.66% in Los Angeles dataset).

Overall, this analysis shows that the importance of textual descriptions increases when numerical information is not available. Moreover, the results confirm our previous findings that the real estate descriptions in the Berlin dataset contain more useful information than those in the Los Angeles dataset.

### 4.5. Visualization of description embeddings

Next, we analyze whether the embeddings of the property descriptions can be clustered according to their price segment. For this purpose, we focus on the gradient boosting model with textual features from BERT. We reduce the dimension from 768 (as generated by the BERT model) to 2 dimensions using "Uniform Manifold Approximation and Projection" (UMAP) (McInnes et al., 2018). Fig. 7 shows the results in different colors according to the price segment. Low-priced properties are plotted in purple, medium-priced properties are plotted in blue, and high-priced properties are plotted in yellow. Evidently, Fig. 7 shows that description embeddings from the same price segment

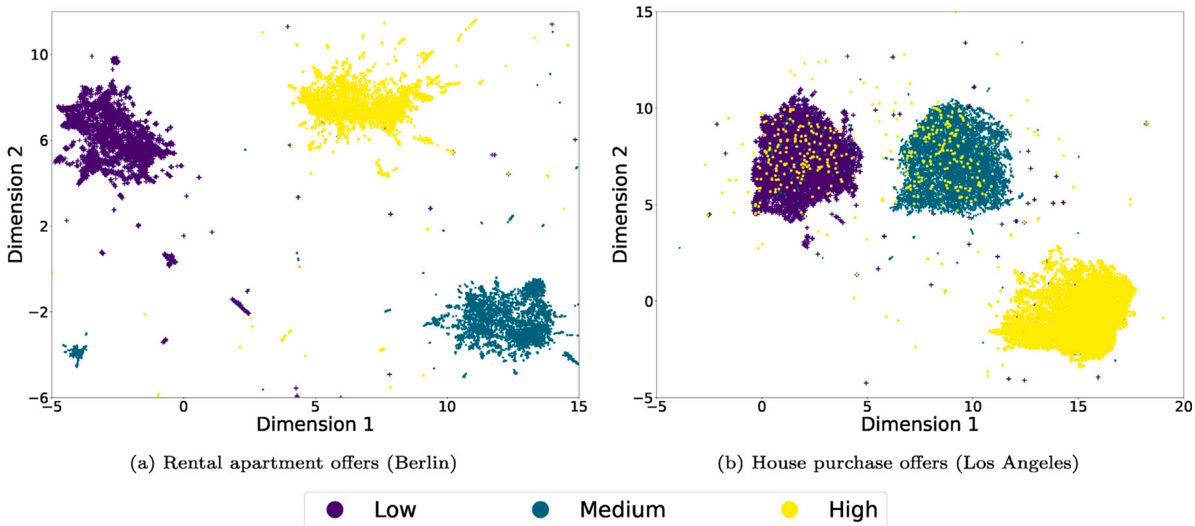



**Fig. 7.** Description embeddings after dimension reduction with UMAP for real estate properties of low, medium, and high price segments.

form individual clusters. Moreover, the results suggest that embeddings are more dissimilar in regard to the difference in prices as the clusters of low- and high-priced properties are more distant to each other than to the description embeddings from the medium price segment. In addition, we observe that the differences between the embedding clusters are greater for German rental apartment offers from Berlin as they exhibit greater distances than the clusters of the English house purchase offers from Los Angeles. Overall, this analysis shows that differences in real estate prices are also reflected in their descriptions.

### *4.6. Feature importance*

We now calculate Shapley additive explanations (SHAP values, Lundberg & Lee, 2017) to analyze the specific influence of real descriptions on the model's predictions. Specifically, the SHAP value of a feature is the average change in the prediction that the model makes when the feature is added as an input (Kraus et al., 2020). The scores are measured in the scale of the target variable. We present the SHAP only for the gradient boosting model as it performs best. Since the property descriptions are reflected in 768 individual embedding dimensions, we aggregate the SHAP values of all embedding dimensions to obtain a more meaningful measure of the importance of the real estate descriptions.

The SHAP values of the top-10 most important features for the rental apartment offers in Berlin are shown in Fig. 8. Evidently, location-based features, living area, and first-time occupancy have the strongest influence on the predicted rents from the gradient boosting model. The description has greater influence on the predicted rent per square meter than the floor and age of the apartment. Furthermore, the plot indicates that adding BERT embeddings of the property description to the gradient boosting model changes the model's predictions on average by 45.21 Euro cents for the monthly rent per square meter (mean rent 11.56 €).

The SHAP values for house purchase offers from Los Angeles dataset are presented in Fig. 9. The plot shows the importance of all nine features in the dataset. Latitude, living area, and age of the building have the strongest influence on the predicted price of the property. The description ranks fourth place. The plot indicates that adding BERT embeddings of the description changes the predictions of the gradient boosting model on average by 46.7 $ per square foot (mean price 583.04 $).

### *4.7. Similarities between property descriptions*

One noticeable difference between the prediction results of our two datasets is the smaller performance improvement in the Los Angeles data when including the property descriptions into the price predictions. Therefore, we conduct a further analysis to clarify the lower error improvement for the Los Angeles descriptions. We assume that the lower usefulness of the Los Angeles descriptions may stem from a smaller variation within the description texts. Property descriptions are generally a good way to provide information about special characteristics of the real estate objective that cannot be reflected by the standard numerical features. Having a higher similarity between the descriptions of the real estate listings may not capture different characteristics as precisely as in the Berlin descriptions. To investigate the variation of embeddings we compute the cosine similarity between the embeddings between all descriptions in the datasets. Given two description embeddings $v_1$ and $v_2$, we calculate the cosine similarity via

$$sim(v_1, v_2) = \frac{\langle v_1, v_2 \rangle}{\|v_1\| \, \|v_2\|}, \quad (5)$$

where $\langle v_1, v_2 \rangle$ denotes the dot product and $\|v\|$ the Euclidean norm.

Fig. 10 graphically shows the distribution of the similarities for the Los Angeles and the Berlin dataset as a boxplot. We find that the interquartile range is much larger among the similarities of the Berlin descriptions, which stems from a substantially smaller first quartile in the Berlin dataset (0.28) than in the Los Angeles dataset (0.75). Moreover, the median (0.76) and mean (0.66) of the similarities among Berlin descriptions are smaller than median (0.79) and mean (0.77) of the similarities among Los Angeles descriptions. The difference of the means is also statistically significant with $p < 0.001$ according to a two-sided $t$-test. Taken together, this suggests that the Los Angeles descriptions are more similar to each other than the Berlin descriptions. This finding is also supported in regard to the different vocabulary sizes of both datasets. While the Los Angeles corpus consists of 26,671 unique words, the Berlin corpus with 45,620 unique words is almost twice as large.

## 5. Discussion

### *5.1. Implications for research*

Our study contributes to research on automated real estate valuation by showing how predictive models can benefit from including numerical text representations of real estate descriptions. Our evaluation

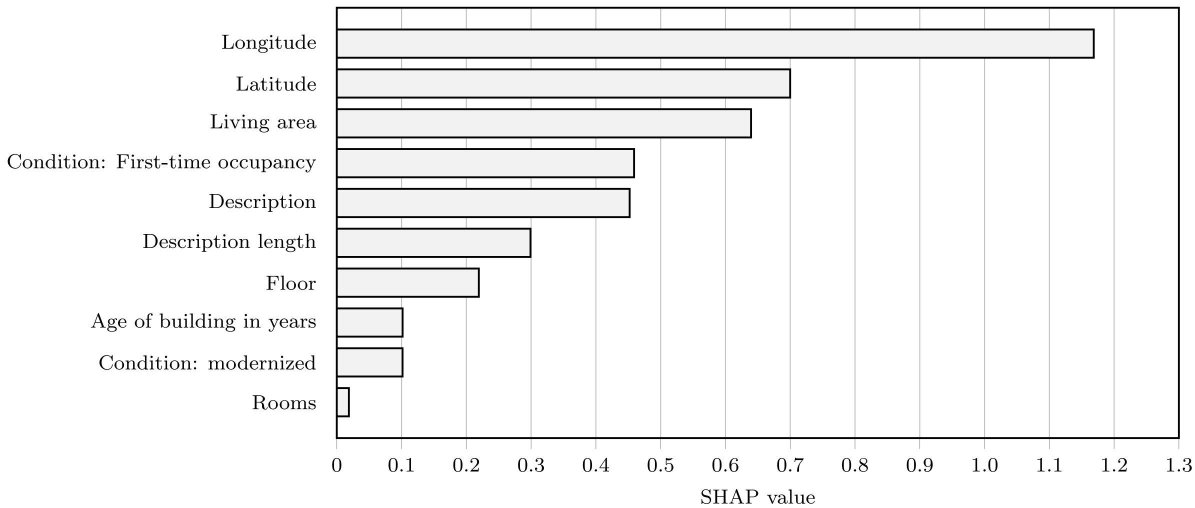



**Fig. 8.** SHAP values for rental apartment offers (Berlin).

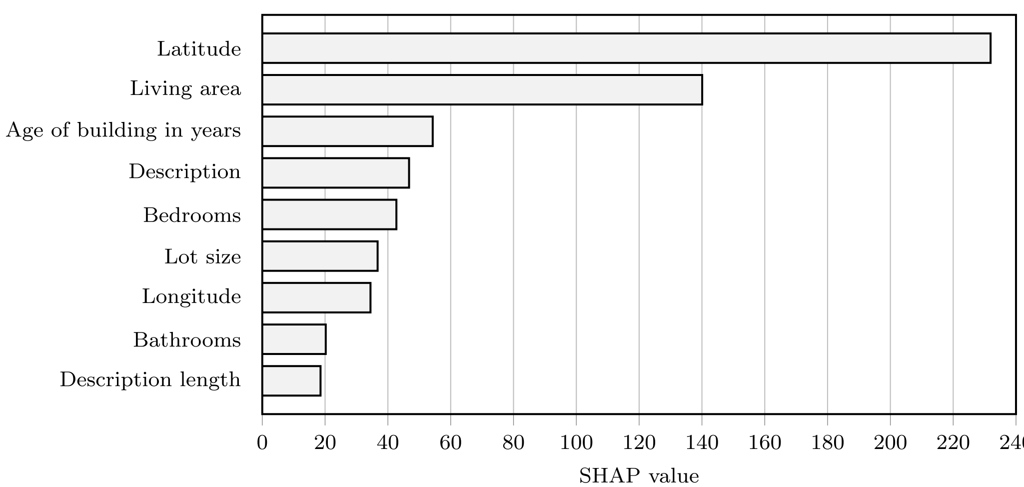



**Fig. 9.** SHAP values of house purchase offers (Los Angeles).

showed error reductions of up to 17.09% for rental apartment offers from Berlin and 5.62% for English house purchase offers from Los Angeles. While prior studies (Goodwin et al., 2014; Liu et al., 2020; Nowak et al., 2019; Nowak & Smith, 2017) have primarily focused on establishing causal relations between real estate descriptions and the corresponding prices, we specifically analyze how including numerical representations of the full textual description benefits non-linear predictive models.

Furthermore, we made an interesting observation regarding the LASSO approach by Nowak et al. (2019). For the Los Angeles dataset, the performance of the LASSO model is similar to the performance of BERT. However, for the German rental apartments offers in Berlin, the features extracted by the LASSO model are strongly outperformed by the BERT representations. This appears at first unexpected given that the Berlin dataset contains more structured features. As a consequence, one could have expected that the textual descriptions of the Los Angeles house purchase dataset contain more additional useful information that is not reflected in the structured features. We find two major explanations for this difference in performance. First, the English descriptions of Los Angeles property are written with a much smaller vocabulary with only 26,671 words compared to the vocabulary of the German descriptions with 45,620 words. Second, our analysis of description similarities suggests that descriptions of houses in Los Angeles are written in a significantly more generic way than the descriptions of rental apartments in Berlin. Taken together, the descriptions of Berlin rental apartments exhibit a larger vocabulary and greater variability in regard to their similarities.

### 5.2. Implications for practice

Real estate market participants can greatly benefit from automated and accurate tools for real estate appraisal (Kok et al., 2017). We have proposed a novel approach that uses state-of-the-art methods from NLP to estimate more accurate real estate valuations than established methods. The proposed model could be integrated into a software application, which allows landlords to estimate the monthly rent they can achieve for their apartments. In addition, buyers and sellers of houses could use this application to quickly obtain an initial estimate of a property's price. Furthermore, landlords and sellers can use the proposed approach to analyze which textual description is likely to achieve a higher rent or price.

### 5.3. Limitations and future research

The main goal of our study is to improve real estate valuation approaches in predicting rent and purchase prices by incorporating textual real estate descriptions. For this purpose, contextualized embedding representations from BERT provide a state-of-the-art method with high predictive performance. The best results were achieved by a gradient boosting model in combination with BERT representations.

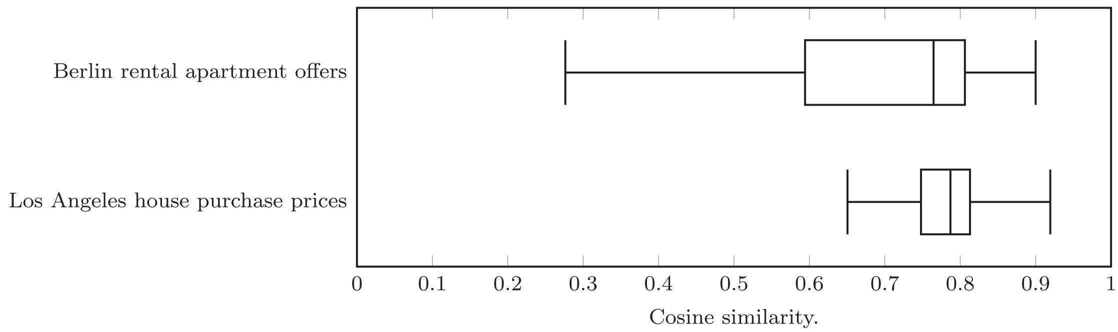



**Fig. 10.** Boxplots of cosine similarities between embeddings of real estate descriptions within both datasets.

However, embedding representations and decision tree ensembles are considered as black-box models, which may limit their use in practice. As a remedy, we visualized the description embedding representations using dimensionality reduction techniques to give the reader an intuition of how these representations capture semantic relations between different texts. Furthermore, we calculated SHAP values (Lundberg & Lee, 2017), which provide at least post-hoc explanations of how the model arrives at decisions.

Moreover, our results are limited by the fact that real estate listings only provide offers for rent or purchase, but they do not necessarily reflect the actual price that was finally achieved. However, it is common practice in the related literature to use asking prices of rental offers as they often present the best available data (Antipov & Pokryshevskaya, 2012; Solovev & Pröllochs, 2021). A study from real estate pricing research (Beracha & Seiler, 2014) found that listing and actual house prices in a large US dataset differ, on average, only by 1.19%.

Another limitation is that our approach requires description texts in digital form, which may not be available for all existing real estate offers. Thus, our proposed approach is mainly applicable for real estate offers from online platforms.

Our findings also provide several avenues for future research. First, it seems intriguing to validate the influence of different descriptions in a controlled laboratory experiment. Given two descriptions of the same real property and the predicted prices from our predictive model, one could analyze whether participants are truly willing to pay the higher price. Second, as we could not fine-tune BERT to the real estate domain due to a small dataset, we use the pretrained version of BERT for our models. With no fine-tuning we already show a clear accuracy benefit of using contextualized embeddings compared to more traditional NLP approaches. Research, therefore, could try to obtain a much larger dataset and show that the difference between non-contextualized methods and contextualized embeddings increases even more through fine-tuning a pre-trained BERT model. Third, future research could look into real estate descriptions of other languages than English and German to see where including textual representations is beneficial. One problem for non-English real estate listings is that BERT models are mostly trained based on English texts, while there are fewer pre-trained models for non-English languages. Fourth, it would be interesting to extend our approach to the textual descriptions of short-term rent offers like hotel rooms or AirBnB offers (Zhang et al., 2020).

## CRediT authorship contribution statement

**Katharina Baur:** Methodology, Software, Formal analysis, Writing – original draft. **Markus Rosenfelder:** Conceptualization, Methodology, Software, Formal analysis, Writing – original draft. **Bernhard Lutz:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

The authors do not have permission to share data.

## Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.eswa.2022.119147.

## References

Antipov, E. A., & Pokryshevskaya, E. B. (2012). Mass appraisal of residential apartments: An application of random forest for valuation and a CART-based approach for model diagnostics. *Expert Systems with Applications*, *39*(2), 1772–1778.

Baldauf, M., Garlappi, L., & Yannelis, C. (2020). Does climate change affect real estate prices? Only if you believe in it. *The Review of Financial Studies*, *33*(3), 1256–1295.

Beracha, E., & Seiler, M. J. (2014). The effect of listing price strategy on transaction selling prices. *The Journal of Real Estate Finance and Economics*, *49*(2), 237–255.

Bourassa, S. C., Hoesli, M., & Oikarinen, E. (2019). Measuring house price bubbles. *Real Estate Economics*, *47*(2), 534–563.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 conference on empirical methods in natural language processing* (pp. 1724–1734). Doha, Qatar: Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Retrieved from https://arxiv.org/abs/1810.04805.

Diebold, F. X., & Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *20*(1), 134–144.

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature*, *57*(3), 535–574.

Goodwin, K., Waller, B., & Weeks, H. S. (2014). The impact of broker vernacular in residential real estate. *Journal of Housing Research*, *23*(2), 143–161.

Hansen, P. R., Lunde, A., & Nason, J. M. (2011). The model confidence set. *Econometrica*, *79*(2), 453–497.

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2–3), 146–162.

Ho, H.-P., Chang, C.-T., & Ku, C.-Y. (2015). House selection via the internet by considering homebuyers' risk attitudes with S-shaped utility functions. *European Journal of Operational Research*, *241*(1), 188–201.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems: vol. 30*, (pp. 3146–3154).

Kofner, S. (2014). The german housing system: Fundamentally resilient? *Journal of Housing and the Built Environment*, *29*(2), 255–275.

Kok, N., Koponen, E.-L., & Martínez-Barbosa, C. A. (2017). Big data in real estate? From manual appraisal to automated valuation. *The Journal of Portfolio Management*, *43*(6), 202–211.

Kraus, M., Feuerriegel, S., & Oztekin, A. (2020). Deep learning in business analytics and operations research: Models, applications and managerial implications. *European Journal of Operational Research*, *281*(3), 628–641.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the 31st international conference on machine learning* (pp. 1188–1196).

Liu, C. H., Nowak, A. D., & Smith, P. S. (2020). Asymmetric or incomplete information about asset values? *The Review of Financial Studies*, *33*(7), 2898–2936.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems* (pp. 4768–4777). Red Hook, NY: Curran Associates Inc..

McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, *3*(29), 861.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119). La Jolla, CA: Neural Information Processing Systems Foundation.

Nowak, A. D., Price, B. S., & Smith, P. S. (2019). Real estate dictionaries across space and time. *The Journal of Real Estate Finance and Economics*, 1–25.

Nowak, A. D., & Smith, P. S. (2017). Textual analysis in real estate. *Journal of Applied Econometrics*, *32*(4), 896–918.

Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Systems with Applications*, *42*(6), 2928–2934.

Pavlov, A. D. (2000). Space-varying regression coefficients: A semi-parametric approach applied to real estate markets. *Real Estate Economics*, *28*(2), 249–283.

Pérez-Rave, J. I., Correa-Morales, J. C., & González-Echavarría, F. (2019). A machine learning approach to big data regression analysis of real estate prices for inferential and predictive purposes. *Journal of Property Research*, *36*(1), 59–96.

Rico-Juan, J. R., & Taltavull de La Paz, P. (2021). Machine learning with explainability or spatial hedonics tools? An analysis of the asking prices in the housing market in Alicante, Spain. *Expert Systems with Applications*, *171*, Article 114590.

Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy*, *82*(1), 34–55.

Scholz, M., Franz, M., & Hinz, O. (2017). Effects of decision space information on MAUT-based systems that support purchase decision processes. *Decision Support Systems*, *97*, 43–57.

Sirmans, S., Macpherson, D., & Zietz, E. (2005). The composition of hedonic pricing models. *Journal of Real Estate Literature*, *13*(1), 1–44.

Solovev, K., & Pröllochs, N. (2021). Integrating floor plans into hedonic models for rent price appraisal. In *Proceedings of the web conference 2021* (pp. 2838–2847).

Trinh, T., Dai, A., Luong, T., & Le, Q. (2018). Learning longer-term dependencies in RNNs with auxiliary losses. In J. Dy, & A. Krause (Eds.), *Proceedings of the 35th international conference on machine learning, vol. 80* (pp. 4965–4974). PMLR.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 6000–6010). Red Hook, NY, USA: Curran Associates Inc..

Vicente, A., & Falkum, I. L. (2017). Polysemy. In A. Vicente, & I. L. Falkum (Eds.), *Oxford research encyclopedia of linguistics*. Oxford University Press.

Yao, Z., Sun, Y., Ding, W., Rao, N., & Xiong, H. (2018). Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the eleventh ACM international conference on web search and data mining* (pp. 673–681). New York, USA: Association for Computing Machinery.

Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing [review article]. *IEEE Computational Intelligence Magazine*, *13*(3), 55–75.

Zhang, L., Yan, Q., & Zhang, L. (2020). A text analytics framework for understanding the relationships among host self-description, trust perception and purchase behavior on airbnb. *Decision Support Systems*, *133*, Article 113288.

Zhao, H., Sinha, A. P., & Bansal, G. (2011). An extended tuning method for cost-sensitive regression and forecasting. *Decision Support Systems*, *51*(3), 372–383.