# Using Text Mining to Analyze Real Estate Classifieds

Sherief Abdallah(✉) and Deena Abu Khashan

The British University in Dubai, Dubai, United Arab Emirates
shario@ieee.org

**Abstract.** There has been an explosion of websites that manage classifieds in general, and real estate listings in particular. Many brokers have adapted their operation to exploit the potential of the web. Despite the importance of the real estate classifieds, there has been little work in analyzing such data. In fact, we are not aware of any work that attempted to analyze the textual data in real estate classifieds using data mining techniques.

In this paper we propose a data mining process that exploits the textual data in real estate classifieds. We conduct the analysis on a large data set, which we gathered from three different property websites. We show that our process exploits the unstructured and ungrammatical textual features to significantly improve the prediction accuracy of a real estate unit price. We also illustrate how text mining combined with linear regression can be used to identify keywords that affect the price negatively or positively.

## 1 Introduction

There has been an explosion of websites that manage classifieds.[1] We focus in this paper on real estate classifieds, which describe a real estate unit that is available either for rent or sale. Although at some point it was feared that such a new trend in advertising properties may threaten the profitability of traditional brokerage companies, many brokers have adapted their operation to exploit what the web has to offer; and they integrated the web technology into their process [5]. Nowadays, many large brokerage companies developed their own websites to list their properties, in addition to listing their properties on 3rd party web portals. In Dubai, where the real estate market constitutes 12.5 % of the Gross Domestic Product (GDP),[2] several major website portals targeted real estate

[1] Criagslist (www.craigslist.com) is a clear example.

[2] "Dubai's GDP climbs 4.4 %", Khaleej Times, 12th June 2013.

A.E. Hassanien et al. (eds.), *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2016*, Advances in Intelligent Systems and Computing 533, DOI 10.1007/978-3-319-48308-5_19

classifieds.[3] Furthermore, major real estate brokering companies list classifieds on their own websites.[4]

Despite the importance of the real estate classifieds, there has been little work in analyzing such data. In particular, most of the previous work that applied data mining to real estate data focused on structured attributes such as number of bedrooms, area, location, etc. [7] (a broader view of related work is given in Sect. 5). However, for classifieds, the unstructured and ungrammatical[5] textual attributes (such as the classified title and description) are important components of a classified that should not be ignored.

In this work we apply text mining, along with regular data mining, to analyze real estate classifieds. Our aim is to answer two important research questions:

- Can the use of text mining improve the accuracy of predicting the price of a real estate classified?
- Can we identify which keywords affect the price of a real estate unit either positively or negatively?

Answering these questions will be very valuable for real estate agents, and even home owners who directly post classifieds. Predicting a more accurate price, for a real estate unit, that takes into account the textual unstructured data (not just the structured data) will prevent the stakeholder from overestimating or underestimating the price. Further more, by identifying the important keywords, the stakeholder can refine the unit description and title to better reflect the price being asked.

We conduct the analysis on a large data set (+50 K records) of real estate classifieds, which we gathered from three different websites that post real estate classifieds. We show that our proposed approach significantly improves the prediction of a property price. We also illustrate how text mining combined with a linear regression model can be used to identify keywords that affect the price negatively or positively.

The rest of the paper is organized as follows. The following section describes how the data was collected and prepared. We then explain our proposed data mining process for predicting the price of real estate units using text mining and linear regression model. This is followed by the evaluation and analysis of our proposed approach using the collected data. We then discuss the related work and conclude our paper.

## 2 Data Preparation

We extracted our data from three major websites that post on-line residential real estate classifieds in the city of Dubai, United Arab Emirates. The data

[3] Examples include Gulf News Ads (www.gnads4u.com), Dubizzle (www.dubbizle.com), Bayut (www.bayut.com), and Property Finder (www.propertyfinder.ae).

[4] Such as such as Better Homes (www.bhomes.com) and Hamptons (www.hamptons.ae).

[5] Ungrammatical means the text does not strictly conform to grammatical rules.

**Table 1.** The features of the data set.

| Name | Type | Description | Number of values |
|---|---|---|---|
| Type | Nominal | Type of classified (flat/villa, for sale/rent) | 4 |
| Beds | Integer | Number of bedrooms | 14 |
| Location | Nominal | Neighborhoods | 161 |
| Title | Text | Title of the classified | - |
| Description | Text | Description of the property | - |
| Price | Integer | The renting or selling price of the property | - |

was collected in the period from February 2011 to April 2011, using our own web crawler.[6] The collected data contained both apartments (flats) and villas (houses) that are offered for either sale or rent. A total of 66,388 records were extracted. Table 1 illustrates the extracted features.

The extracted data was then cleaned as follows. Records with unreasonably low price/rent were removed (less than AED 10,000).[7] Some unwanted texts were removed from the "Description" feature and replaced with white spaces, including HTML tags, email addresses, website addresses, and few irrelevant phrases related to contacting agents (please contact, for more information, for further information, for international call please dial).

In our analysis we partitioned the data in to 6 different subsets depending on the type: one data set for each of the four types (renting apartments, renting villas, selling apartments, and selling villas), one data set for all the ads for rent (apartments or villas), and one data set for all the ads for sale (apartments or villas).

## 3 Using Text Mining to Analyze Classifieds

Figure 1 illustrates the process we propose to analyze real estate classifieds data. The process was implemented using RapidMiner.[8] The following sub-sections describe the different components of our proposed approach.

### 3.1 Text Mining

The text mining stage converts the text attributes to numerical features. As we mentioned earlier, The purpose of applying text-mining is to discover the effect of the hidden information involved in "Title" and "Description" features that

[6] While more sophisticated crawlers do exist, [13], for our purposes we preferred a simple and efficient crawler.

[7] The rental price in Dubai is annual, and the currency is Arab Emirates Dirham (AED). US Dollar = 3.67 AED.

[8] RapidMiner is an open-source system for data mining that allows building a rich data-flow process of data-mining operators (http://rapid-i.com).

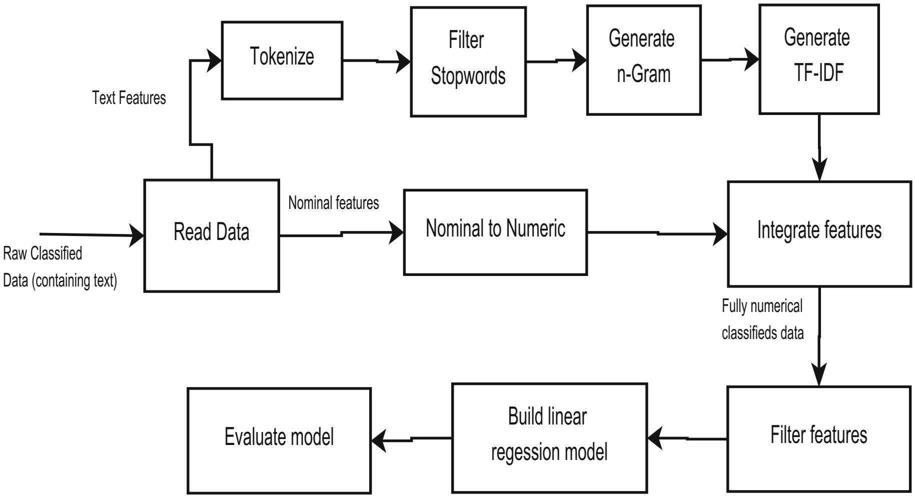

**Fig. 1.** The data mining process we used to analyze real estate classifieds.

might enhance the accuracy of predicting a property's price. First, the text is tokenized by splitting the text into sequence of tokens, words. The characters of every single word (token) is converted to lowercase. Then, stop-word tokens are removed using a predefined stop-word list. Also any word that is shorter than two characters is removed. This is followed by the generation of term n-Grams of tokens. A term n-Grams is a series of successive tokens of length n. The last step of the text mining process is generating the Term Frequency-Inverse Document Frequency (TF-IDF) for each token. TF-IDF counts how many times a particular token (n-Gram) appears in text, which is then inversely weighted by how common the token is across different classifieds. The output of the text-mining model is thousands of new numerical features. Each feature corresponds to a token, where the feature value is the TF-IDF of the token.

### 3.2 Feature Selection and Preprocessing

In the second phase, the features are filtered to reduce their numbers. Here we use the correlation between each feature and the target attribute (the property price) as the feature weight. Only the features with minimum weight threshold are then selected (we have experimented with different threshold values as we show later). Also to facilitate linear regression, nominal features (such as the "Type" and "Location") are converted to binary features, where each feature corresponds to a value of the original nominal attribute. For example, "Dubai Marina" is a possible value of the location attribute. After pre-processing we have a binary feature "loc_dubai_marina", which equals 1 if and only if the location attribute equals "Dubai Marina".

### 3.3 Linear Regression

Finally the linear regression model learning algorithm is applied to the preprocessed dataset. The linear Regression (LR) model assumes the price of a property is a linear combination of the property features.[9] The learning algorithm finds the best weight for each feature based on the dataset. The weight intuitively reflects the feature's effect on the price. For example, when a feature has a positive weight value, it means that the feature works toward increasing the price. Similarly, having a feature with negative weight has the effect of decreasing the price. We show in the following section how this intuition helps in identifying important keywords.

## 4 Analysis

After building the LR model and using it for prediction, the evaluation metrics are calculated. Table 2 displays the Root Mean Squared Error (RMSE), and the Correlation Coefficient (CC) for the different datasets without using the textual features (i.e. relying primarily on structured features).

**Table 2.** Performance measurements of linear regression using only structured features.

| Dataset | RMSE | CC |
|---|---:|---|
| Type 0 | 110,263.85 | 0.479 |
| Type 1 | 204,650.55 | 0.543 |
| Type 2 | 8,676,287.33 | 0.296 |
| Type 3 | 4,756,663.73 | 0.802 |
| Renting | 144,483.34 | 0.586 |
| Selling | 7,738,338.21 | 0.501 |

As shown in Table 2, the RMSE for Type0 and Type1 are much lower than the RMSE for Type2 and Type3. This is expected because Type0 and Type1 subsets represent renting price of real estate property and Type 2 and Type3 subsets represent selling price. On the other hand, Type2 dataset has the greatest RMSE and the lowest linear correlation between regular features and the price feature. Also, although the dataset of Type3 has the highest correlation, its RMSE is relatively high. Having high correlation means that there is a strong linear relationship between the predicted price and the other regular (structured) features so that when values of the regular features increase (decrease) the price value increases (decreases) as well.

Looking at the LR model for each data subset and analyzing it, we found (not surprisingly) that the features related to locations have the biggest effect on

[9] A common assumption in automatic real estate valuation.

**Table 3.** Results of linear regression experiments with and without text mining

| Type0 Renting apartments dataset | | | |
|---|---|---|---|
| | LR | Text Mining + LR | |
| | | Uni-gram | Bi-gram |
| Selected weight | - | *w* >= 0.2 | *w* >= 0.3 |
| No. of features | 122 | 321 | 343 |
| RMSE | 110,263.85 | 82,044.01 | 61,715.12 |
| CC | 0.479 | 0.757 | 0.871 |
| Type1 Renting villas dataset | | | |
| | LR | Text Mining + LR | |
| | | Uni-gram | Bi-gram |
| Selected weight | - | *w* >= 0.7 | *w* >= 0.16 |
| No. of features | 92 | 558 | 548 |
| RMSE | 204,650.55 | 99,021.76 | 89,710.59 |
| CC | 0.543 | 0.914 | 0.93 |
| Type2 Selling apartments dataset | | | |
| | LR | Text Mining + LR | |
| | | Uni-gram | Bi-gram |
| Selected weight | - | *w* >= 0.06 | *w* >= 0.25 |
| No. of features | 87 | 503 | 407 |
| RMSE | 8,676,287.33 | 5,490,825.94 | 3,853,490.15 |
| CC | 0.296 | 0.797 | 0.906 |
| Type3 Selling villas dataset | | | |
| | LR | Text Mining + LR | |
| | | Uni-gram | Bi-gram |
| Selected weight | - | *w* >= 0.15 | *w* >= 0.2 |
| No. of features | 76 | 331 | 508 |
| RMSE | 4,756,663.73 | 3,778,524.95 | 3,750,734.50 |
| CC | 0.802 | 0.88 | 0.882 |
| Renting apartments and villas dataset | | | |
| | LR | Text Mining + LR | |
| | | Uni-gram | Bi-gram |
| Selected weight | - | >= 0.2 | >= 0.2 |
| No. of features | 143 | 237 | 361 |
| RMSE | 144,483.34 | 105,307.09 | 96,334.73 |
| CC | 0.586 | 0.807 | 0.841 |
| Selling apartments and villas dataset | | | |
| | LR | Text Mining + LR | |
| | | Uni-gram | Bi-gram |
| Selected weight | - | >= 0.1 | >= 0.1 |
| No. of features | 102 | 239 | 283 |
| RMSE | 7,738,338.21 | 5,788,073.98 | 4,440,108.38 |
| CC | 0.501 | 0.762 | 0.868 |

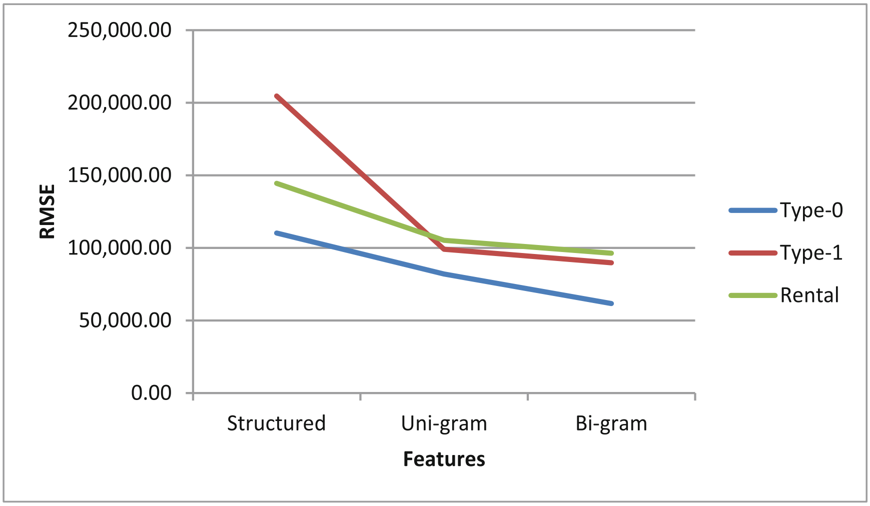

**Fig. 2.** The RMSE, for the three datasets related to renting property, declines as more sophisticated textual features are added.

increasing or decreasing the price. In the second part of our experiments we apply the linear regression model after the text mining process. The experiments were repeated twice for each dataset. First experiment considered uni-grams tokens and the second experiment considered bi-grams tokens. Table 3 shows the results.

For better illustration, Figs. 2 and 3 visualize the decline in RMSE for the different datasets. As more textual features are incorporated, from only structured features to uni-grams and finally to bi-grams, the RMSE is consistently declining across the different datasets. This clearly confirms that exploiting textual features in real estate classifieds can greatly improve the accuracy of price prediction.

Few points are worth noting. The RMSE (even after the decline) is relatively high. This is due to the wide variety of offered units. For example, the price of a 2-bedroom apartment for sale in Dubai Marina (a neighborhood in Dubai) ranges from just above 1 million, to over 25 million AED (one of the 25+million unit is a luxurious furnished penthouse). Given the variety, an RMSE of less than 5 million AED (as shown in Fig. 3 with bi-gram textual features) is actually very good.

To understand why text mining reduced the RMSE, we investigated the linear regression model to identify which words affected the price positively or negatively. Some of the words that affected the price are related to the location. While the location attribute in the original dataset[10] did specify the location of the unit in a structured manner, the words that were discovered through text-mining were at a finer grain. For example, there is a location code reserved for

[10] Recall that the location attribute was converted to binary attributes corresponding to each location value, as we explained in Sect. 3.2.

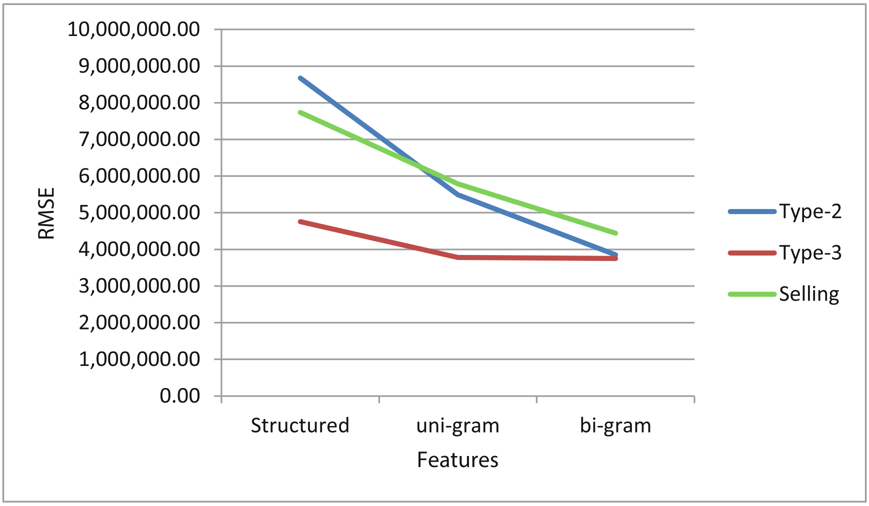

**Fig. 3.** The RMSE, for the three datasets related to buying property, declines as more sophisticated textual features are added.

**Table 4.** Sample of words (uni-gram and bi-gram) that affect the price of a real estate classified (either positively or negatively)

| Dataset | Sample of words affecting positively | Sample of words affecting negatively |
|---|---|---|
| Type-0 | investment, balcony | bus, unfurnished |
| Type-1 | wardrobes, driving_mall | fronds_palm, truly_luxurious |
| Renting | shores_palm, signature_villas | european_miditranian,building |
| Type-2 | hotel_barsha, penthouse_car | road_plot, ground_floor |
| Type-3 | views_beachfront, billiard | palm_offer,school_springs |
| Buying | full_building, mixed_jumeirah | apartment |

Palm Jumeirah. Through text mining and linear regression, the textual feature “fronds_palm” (which refers to a neighborhood within Palm Jumeirah) was discovered to affect the price negatively. Table 4 lists a sample of the discovered important words that affect the price of a real estate unit either positively or negatively.

## 5 Related Work

Due to the importance of valuating a real estate property, there has been extensive research on automatic valuation [2,3,9,11,14,15]. Most of that work used hedonic models, which assumed the price can be predicted from a combination (usually linear) of the property (structured) features. Traditional hedonic models are based on human expertise, where the model parameters are usually hand-coded by experts, unlike our proposed method here, which uses data mining. There has been growing literature on the use of data mining techniques to analyze real estate data [4,6–8,10], however, most of the previous work focused only on structured features and ignored textual features. We review sample of these works in the remainder of this section.

One of the early works [10] used decision tree and neural network techniques to predict the sale price of a house. The analysis used data with 15 numerical features that represent the houses' characteristics plus a categorical feature that corresponds to the address. The dataset consisted of 1000 records that were collected from the houses' sales transactions in Miami, US. Unlike our work, the analysis focused only on properties for sale (did not include rentals), used only structured features (no text mining) and relied on a much smaller dataset (compared to our +50,000 records). A broader analysis was conducted in [17], covering 295,787 transactions from four cities in the US. Again, only numerical features were used (although more extensive features were used, almost 200) and no textual features were used (also despite attempting to predict the price, no performance criterion was reported). A more recent work [7] proposed Adaptive Neuro Fuzzy Inference System (ANFIS) and tested the system over 360 records of past sales properties in Midwest, US. The dataset had 14 numerical features and again no textual feature was used.

Another research paper [6] focused on studying the prediction of prices of apartments in city in Macedonia. Among the three data mining techniques that were applied on a dataset of 1200 sales transactions, the logistic regression (very similar to linear regression) was found to be the superior in prediction accuracy over decision tree and neural network techniques. Like the other earlier mentioned papers, there was no use of textual data. Some attempted to add structure to unstructured and ungrammatical data. However, this requires domain knowledge to build a reference structure (model) which can be used to extract the corresponding features [12]. Our proposed approach does not require deep domain knowledge (aside from simple data cleansing, the whole process is mostly automated).

The most related work to ours was concurrently and independently developed for analyzing real estate classifieds in the United States [16].[11]

## 6 Conclusion and Future Work

We propose in this paper a data mining process that uses text mining along with linear regression to improve the prediction of the price of real estate classifieds. We show that using text mining significantly reduces the RMSE of prediction. We also show how our proposed approach can identify keywords that affect the price positively or negatively.

[11] Our work was actually published as an MSc dissertation a couple of months earlier than Dick Stevens' dissertation. However, we can not provide further evidence due to the double-blind review process.

One of the direction we want to pursue is extending our analysis to the Arabic language (which is commonly used in our region). We are also considering the integration of our system with a named-entity recognition component [1], particularly for identify locations in ungrammatical text, to improve accuracy.

## References

1. Abdallah, S., Shaalan, K.F., Shoaib, M.: Integrating rule-based system with classification for arabic named entity recognition. In: Gelbukh, A. (ed.) Computational Linguistics and Intelligent Text Processing. LNCS, vol. 7181, pp. 311–322. Springer, Heidelberg (2012)
2. Bourassa, S.C., Cantoni, E., Hoesli, M.: Predicting house prices with spatial dependence: a comparison of alternative methods. J. Real Estate Res. **32**(2), 139–159 (2010)
3. Case, B., Clapp, J., Dubin, R., Rodriguez, M.: Modeling spatial and temporal house price patterns: a comparison of four models. J. Real Estate Finan. Econ. **29**(2), 167–191 (2004)
4. Chen, T.H., Chen, C.W.: Application of data mining to the spatial heterogeneity of foreclosed mortgages. Expert Syst. Appl. **37**(2), 993–997 (2010)
5. Crowston, K., Wigand, R.T.: Real estate war in cyberspace: an emerging electronic market? Int. J. Electron. Markets **9**(1–2), 1–8 (1999)
6. Gacovski, Z., Kolic, J., Dukova, R., Markovski, M.: Data mining application for real estate valuation in the city of skopje. In: ICT Innovations 2012, Web Proceedings, pp. 537–538 (2012). ISSN 1857-7288
7. Guan, J., Zurada, J., Levitan, A.S.: An adaptive neuro-fuzzy inference system based approach to real estate property assessment. J. Real Estate Res. **30**(4), 395–422 (2008)
8. Helbich, M., Brunauer, W., Hagenauer, J., Leitner, M.: Data-driven regionalization of housing markets. Ann. Assoc. Am. Geog. **103**(4), 871–889 (2013)
9. Helbich, M., Jochem, A., Mcke, W., Hfle, B.: Boosting the predictive accuracy of urban hedonic house price models through airborne laser scanning. Comput. Environ. Urban Syst. **39**, 81–92 (2013)
10. Jaen, R.D.: Data mining: an empirical application in real estate valuation. In: Haller, S.M., Simmons, G. (eds.) FLAIRS Conference, pp. 314–317. AAAI Press (2002)
11. McGreal, S., de La Paz, P.T.: An analysis of factors influencing accuracy in the valuation of residential properties in spain. J. Property Res. **29**(1), 1–24 (2012)
12. Michelson, M., Knoblock, C.A.: Creating relational data from unstructured and ungrammatical data sources. J. Artif. Intell. Res. (JAIR) **31**, 543–590 (2008)
13. Pera, M.S., Qumsiyeh, R., Ng, Y.K.: Web-based closed-domain data extraction on online advertisements. Inf. Syst. **38**(2), 183–197 (2013)
14. Rossini, P.: Accuracy issues for automated and artificial intelligent residential valuation systems. In: International Real Estate Society Conference (1999)
15. Rossini, P., et al.: Using expert systems and artificial intelligence for real estate forecasting. In: Sixth Annual Pacific-Rim Real Estate Society Conference, Sydney, Australia, pp. 24–27. Citeseer (2000)
16. Stevens, D.: Predicting real estate price using text mining. Master's thesis, Tilburg University School of Humanities, The Netherlands (2014)
17. Wedyawati, W., Lu, M.: Mining real estate listings using ORACLE data warehousing and predictive regression. In: Proceedings of the 2004 IEEE International Conference on Information Reuse and Integration, IRI, pp. 296–301 (2004)