A Comparative Study of Text Representations for French Real-Estate Classified Advertisements Information Extraction

Lucie Cadorel[1,2], Andrea G. B. Tettamanzi[1,*]

[1]Université Côte d'Azur, Inria, CNRS, I3S, Sophia Antipolis, France
[2]KCityLabs, Sophia Antipolis, France

Abstract

Text representations are widely used in NLP tasks such as text classification. Very powerful models have emerged and been trained on huge corpora for different languages. However, most of the pre-trained models are domain-agnostic and fail on domain-specific data. We perform a comparison of different text representations applied to French Real Estate classified advertisements through several text classification tasks to retrieve some key attributes of a property. Our results demonstrate the limitations of pre-trained models on domain-specific data and small corpora, but also the strength of text representation, in general, to capture underlying knowledge about language and stylistic specificities.

Keywords

Text Representations, Information Extraction, Real-Estate Market

1. Introduction

Real-estate classified advertisements provide great details and relevant information about a property that is valuable for the intelligence of the real-estate market. For example, price predictions are often based on attributes such as the type of property, the number of rooms or even the floor at which it is located. However, those key information are not always clearly specified in the ads and often differ from an ad to another and from an advertiser from another. Thus, the automatic extraction of key information from the text of real-estate classified advertisements is a challenging and promising task to help in some real-estate market applications.

Given the limited number of different values that economically relevant attributes of a real-estate property may take, this information extraction task can be viewed as a classification problem. For instance, if the number of rooms of a property is sought for, the problem can be stated as assigning the property described by a given advertisement to one of the classes labeled as $\{1, 2, 3, 4, 5+\}$, corresponding to single-room, two-room, three-room, four-room, and five-or-more-room dwellings. Nevertheless, texts of real-estate ads are often short with language and stylistic specificities and variabilities. Also, an additional challenge may be represented by the need to extract information from advertisements written in languages other than English, for which linguistic resources are thus harder to find or less well-developed.

Text representations models have emerged as very powerful approaches to learn useful features of a text and have been widely used for Machine Learning tasks such as classification. It is thus interesting to carry out a comparative study of the most prominent models found in the literature as they are applied to this specific text classification task, to understand their strengths and weaknesses.

Our main contributions may be summarized as follows:

- we apply different text representations to a classification task to retrieve key attributes of properties found in classified advertisements written in French, that we have collected and annotated manually;
- we propose a comparison of the strength and limitations of different text representations ;
- we analyse the vocabulary and register used by French real-estate agents to understand their impact in the classification.

The rest of the paper is organized as follows: Section 2 positions our contribution with respect to the literature; Section 3 provides a detailed description of the dataset and the method we propose to classify real-estate ads; Section 4 reports the results of the experiment and draws some conclusions.

2. Background and Related Work

From classical to the state-of-the-art methods, feature extraction models are the process of converting raw data

ENIGMA-23, September 03–04, 2023, Rhodes, Greece

[*] Corresponding author.

lucie.cadorel@inria.fr (L. Cadorel); andrea.tettamanzi@univ-cotedazur.fr (A. G. B. Tettamanzi)
https://www.i3s.unice.fr/~tettaman/ (A. G. B. Tettamanzi)
0000-0002-9397-5223 (L. Cadorel); 0000-0002-8877-4654 (A. G. B. Tettamanzi)

into numeric features.

One of the most common classical models is Bag-of-words (BoW), which represents a text by the occurrence of words. This model involves a vocabulary and a measure of the occurrence of words. The vocabulary captures all the words found in the corpus of texts and is fixed. Regarding the measure, it can be binary (presence or not of a word in the text) or weighted. For example, Term Frequency – Inverse Document Frequency (TF-IDF) is a weighted BoW that scores each word of a text by its frequency in the text (TF) and across the whole corpus (IDF). This measure penalizes very frequent words and highlights relevant one. The classical methods are easy to compute and customize for any language and text specificities. However, this approach suffers from a curse of dimensionality because of the size of the vocabulary and the sparsity. Also, it does not capture the position and meaning of a word in a text.

Those limitations led to a feature learning technique in which words are mapped to a vector of N dimensions, with N smaller than the size of the vocabulary. This approach is called Word Embedding. The most well-known model is Word2Vec, developed by Mikolov et al. [1] and based on neural networks. Two architectures have been released: continuous bag-of-words (CBOW) and Skip-gram. Both methods train a word against its neighboring words in the input corpus. The main difference is CBOW uses neighbors to predict the target word, while Skip-gram uses the target to predict its neighbors.

A variant of Word2Vec is Doc2vec [2], which creates a numeric representation for the whole document. As Word2Vec, two architectures have been built. The first one is based on CBOW with an extra input which is the ID of the document. This model is called Distributed Memory version of Paragraph Vector (PV-DM). The other one is inspired by Skip-gram and is called Distributed Bag of Words version of Paragraph Vector (PV-DBOW).

Finally, there exist other models similar to Word2Vec, such as Global Vectors (GloVe) [3] or FastText [4]. Glove focuses on the global context instead of local one and uses a word-word co-occurrence matrix computed from the entire corpus. FastText is based on the CBOW architecture but using n-grams as input instead of full words. N-grams help to prevent the Out-of-Vocabulary problem that suffer Word2Vec, Glove or classical representations. However, this method requires a huge storage memory.

All of the feature learning techniques presented above have limitations such as the need for a huge corpus to train and failure to capture contextual information.

Context-based models try to tackle those limitations. Contextual embeddings help to distinguish a same word with a different meaning according to the semantics and grammar. The first contextual embeddings models are mainly based on the (Bi-)LSTM architecture. For example, Embeddings from Language Models (ELMo) [5] is a two-layer bidirectional language model using Bi-LSTM. This means that the left and right contexts are taken into account in the predictions. Also, it uses a character-level representation of words. Nevertheless, those kinds of modes do not improve performance significantly and are computationally very expensive.

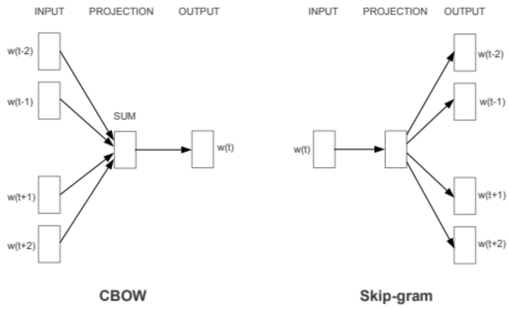


Figure 1: CBOW and Skip-gram architectures [1]

The last state-of-the-art context-based model uses a Transformer architecture [6]. It has been proven that Transformers are faster and more efficient than (Bi)-LSTM (ELMo) or CNN (Word2Vec). For example, Bidirectional Encoder Representations from Transformers (BERT) [7] and its variants are one of those models. BERT uses parallel attention layers instead of sequential recurrent neural networks as ELMo does. Also, it is trained on a huge corpus with two specific tasks: masked language model (MLM) and next sentence prediction (NSP). For MLM, some tokens are masked and the model has to predict them in a sentence. The other task (NSP) is to try to predict, between two sentences, which one follows the other one. Models like BERT reach high performance compared to the other embedding models. However, some limitations have arisen: the need for a huge training set and a pre-trained model on general domain limits their application to specific domains or tasks. Thus, some specific domain models have been trained such as BERTTweet for Twitter or SciBERT for the biological domain.

The advantages and limitations of the different existing word embedding models for low-quality data are discussed in a recent survey of word representation models [8].

All the methods described above have been widely used and compared for different tasks and types of text. However, a few of them are focused on small data and French documents, as in our work.

Dynomant et al. [9] compare mainly Word2Vec (Skip-gram, CBOW), FastText and Glove on a specific dataset (health-related documents) written in French. They also

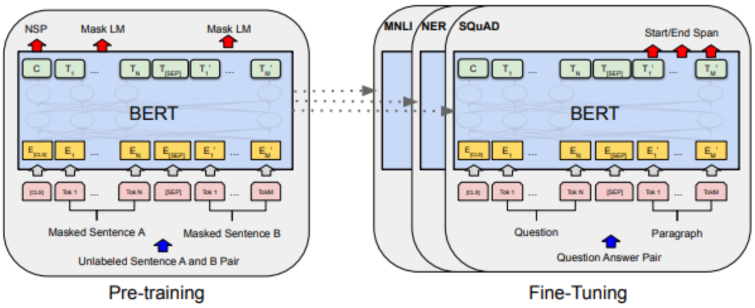


Figure 2: BERT architecture [7]

have to come to grips with the issue of the quality of language. According to their evaluation, which is more qualitative (similarity, word clustering, etc.) than quantitative, they find Word2Vec with a Skip-gram architecture to be the most promising model.

Another work [10], comparing Word2Vec to latent semantic analysis (LSA), shows that LSA gives better results for small training corpora, while Word2Vec works best with medium-sized training corpora. This study too carries out a qualitative, rather than quantitative, analysis.

A recent survey of text classification algorithms [11] illustrates essentially the same methodology we followed, consisting of pre-processing, feature extraction, and classification, by covering a variety of models of text representations (BoW, TF-IDF, Word2Vec, BERT, etc.) and classification (Naive Bayes, Decision Trees, Random Forests, Gradient Boosting, Logistic Regression, and Deep Learning). It compares the advantages and limitations of each model. Also, the authors mention several works in different domains, including health, business, and social, and their applications.

Finally, Gupta and Waseem [12] try different word embedding models to detect hate speech from Twitter and use Logistic Regression to classify tweets. The authors find Word2Vec to perform better than Glove, and FastText to be the worst model in that setting. They also provide compelling evidence that, while domain-specific models outperform domain-agnostic ones, their combination yields the best results. In addition, they find Doc2Vec to perform poorly and Tweet2Vec (specifically designed for Twitter micro-blogs) to perform rather well on imbalanced data, but less so on balanced data.

3. Method

In this section, we will first describe the dataset, then the preprocessing and feature extraction applied on the advertisments and, finally, the different classification tasks that we have used for the comparison.

3.1. Dataset

We gathered 5,440 real-estate classified advertisements of residential and commercial properties, luxury homes and garages/parkings, all located on the French Riviera, from various French online advertisers. The ads are written in French and composed of a title, description, pictures, and some metadata about the property (e.g type of property, number of rooms, price, etc.). Metadata include the most important and relevant information to summarize a property. However, metadata differ from an ad to another and from an advertiser to another. On the other hand, all ads contain a textual description which is a great source of information to infer missing metadata.

In our study, we focused on extracting the type of property, the number of rooms, and the apartment's floor. As we have this information in metadata for a sufficient number of ads, it was relatively easy to label the texts and to apply classification for each type of information. Also, we created artificial classes to normalize our targets. For example, for the type of property, we did not distinguish a luxury villa from a small house : both have been classified as 'House'. Regarding the number of rooms, it goes from '1' to '5 and more' rooms which is a popular discretization in the Real Estate market. Finally, the floor of the property is divided in 5 classes : from 'Ground floor' to 'third floor' and then, 'High floor' and 'Last Floor'. The last two classes might be confusing as 'Last Floor' is also a 'High Floor' but the information 'Last Floor' could be important in price predictions.

Tables 1, 2, and 3 show the class distribution for each target. The type of property is complete (0 missing data) but not really balanced. It is very easy to get the type of property in the metadata since it is essential to sell the product. However, the number of rooms and the floor of a property are not complete. This is the reason why it

Table 1
Type of property

Label	Distribution
Apartment	2402
House	1679
Commercial property	988
Garage/Parking	349
Missing data	0

Table 2
Number of rooms

Label	Distribution
1	188
2	694
3	1104
4	690
5+	1342
Missing data	63

Table 3
Floor of the property

Label	Distribution
Ground floor	311
1	165
2	131
3	96
High floor	97
Last floor	41
Missing data	1560

might be interesting to predict this kind of information from the text. Also, we only predict the number of rooms for apartments and houses, and the floor of the property only for apartments.

Although we plan on making our dataset publicly available, for the time being we could not, for legal reasons, as it is expressly forbidden by the real-estate agencies that own those advertisements to republish them in any form, as long as the relevant properties are still on the market.

3.2. Text Preprocessing

Preprocessing and cleaning texts are a crucial step since French real-estate ads are full of noisy, repetitive words and abbreviations.
First, we removed noise such as elongated punctuations (“......”, “!!!!”, etc.) and URLs. We replaced symbols such as “€” by “euros”, “m2”, “M2”, and “M^2” by “m^2”. Also, some abbreviations are of common use in the real-estate market, e.g., ‘apt.’ stands for apartment, ‘balc.’ for balcony. We tried to remove proper nouns found in phrases such as “*contactez Paul Martin*” (*contact Paul Martin*), which refer to the advertiser, thanks to a regex.
Then, we lemmatized texts with a French lemmatizer from **spaCy**. This lemmatizer has been trained on the French Sequoia corpus[1] and WikiNER. The lemmatizer performs pretty well for French, as we can see in Figure 3. Nevertheless, the syntax of real-estate ads is slightly different from the general French syntax and the lemmatizer fails to assign the correct lemma. For example, the noun “*nuit*” (*night*) in the phrase “*coin nuit*” (*sleeping area*) is erroneously lemmatized in the verb “*nuire*” (*nuit* is also the 3rd person sg. form of “*nuire*”, *to harm*).
The final step was to remove punctuation, numbers and French stopwords. We tuned the stopword list with very frequent words in the real estate vocabulary, e.g., “*honoraires*” (*fees*), “*agent immobilier*” or “*agent commercial*” (*real-estate agent*).

3.3. Feature Extraction

After cleaning the texts, we applied different text representations methods presented in Section 2.

Two classical Bag-of-Words methods have been tested : TF and TF-IDF. We set maximum features to 5,000 and we chose to take a range of n-grams from 1 to 3, that is to say from 1 to 3 tokens.

Then, we tried non-contextual embedding with Word2Vec (Skip-Gram), Doc2Vec (PV-DM), and FastText. We trained those 3 models on our corpus of ads. We chose a smaller number of features for FastText because of its need of memory storage. We wanted to compare our own trained non-contextual embeddings with a pre-trained model, but only few models have been trained on a French corpora. Thus we only found a pre-trained Word2Vec model [13].

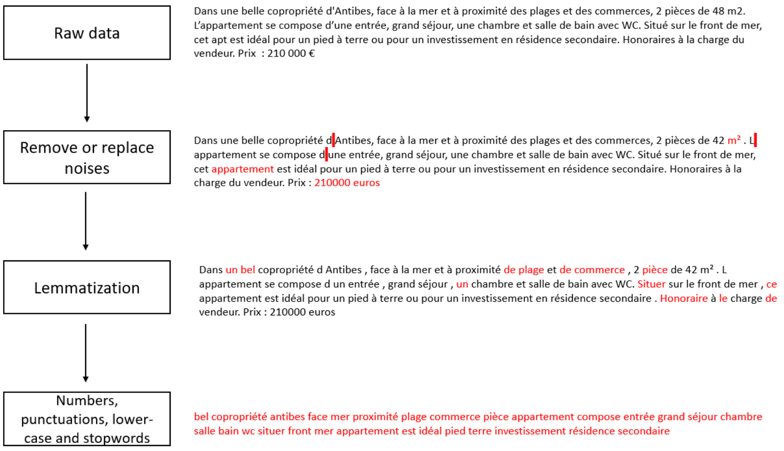
Finally, we wanted to apply contextual embedding. However, although pre-trained (Bi-)LSTM models such as ELMo have recently become available for French,[2] we were not aware of them at the time we planned our experiments. Now, it is very long to train a model and a huge corpus is needed, which we lack. Therefore, we decided to try pre-trained Transformer models for French, like CamemBERT [14] and FlauBERT [15]. These French models chiefly differ for their training data. CamemBERT was trained on OSCAR [16], whose size is 138 GB after cleaning, while FlauBERT used 24 corpora from different sources (Wikipedia, books, Common Crawl, etc.), whose overall size is only half as the CamemBERT training data (71 GB). They also differ for their tokenizer: FlauBERT uses a basic Byte Pair Encoding [17], whereas CamemBERT prefers its extension, called SentencePiece [18]. Finally, the models use different strategies for the masking task: FlauBERT masks sub-word, whereas CamemBERT masks the whole word.

3.4. Classification Task

The final step was to classify our real-estate ads according to three labels (Type of property, Number of rooms

[1] https://github.com/UniversalDependencies/UD_French-Sequoia

[2] Cf., for example, https://github.com/HIT-SCIR/ELMoForManyLangs.



Dans une belle copropriété d'Antibes, face à la mer et à proximité des plages et des commerces, 2 pièces de 48 m2. L'appartement se compose d'une entrée, grand séjour, une chambre et salle de bain avec WC. Situé sur le front de mer, cet apt est idéal pour un pied à terre ou pour un investissement en résidence secondaire. Honoraires à la charge du vendeur. Prix : 210 000 €

Dans une belle copropriété d Antibes, face à la mer et à proximité des plages et des commerces, 2 pièces de 42 m^2 . L appartement se compose d une entrée, grand séjour, une chambre et salle de bain avec WC. Situé sur le front de mer, cet appartement est idéal pour un pied à terre ou pour un investissement en résidence secondaire. Honoraires à la charge du vendeur. Prix : 210000 euros

Dans un bel copropriété d Antibes , face à la mer et à proximité de plage et de commerce , 2 pièce de 42 m^2 . L appartement se compose d un entrée , grand séjour , un chambre et salle de bain avec WC. Situer sur le front de mer , ce appartement est idéal pour un pied à terre ou pour un investissement en résidence secondaire . Honoraire à le charge de vendeur. Prix : 210000 euros

bel copropriété antibes face mer proximité plage commerce pièce appartement compose entrée grand séjour chambre salle bain wc situer front mer appartement est idéal pied terre investissement résidence secondaire

Figure 3: Example of text pre-processing

and Floor of the property) in order to retrieve missing metadata. We used our different feature extractions as input and we applied a classifier. Different classifiers have been tried (e.g. Naive Bayes, Logistic Regression, or Random Forest) for every text representations, but in the following, we will only present results with Logistic Regression as it gave higher score. CamemBERT and FlauBERT have been already trained for classification, so we used their classifier based on neural networks.

We also compared our models to a simple Regex, which is our baseline. For instance, to find the type of property, we crafted rules that classify a property as an apartment if the words "apartment" or "studio" are found; else if "house" or "villa" are found, it is classified as house, and so on. For the number of rooms, we searched for a number (in words or digits) before the word "pièce" (room) or "chambre" (bedroom); if we find a number before "pièce" then we take it as the label for classification; else, if we find a number before "chambre", we take that number plus one. The same idea is followed for the floor of a property, but the word "étage" (floor) is used instead.

We used the F1-Score as measure of accuracy to handle imbalanced classes.

4. Experiments and Results

In this section, we will present the results of our experiments and the comparison of the different text representations. Afterward, we will discuss the possible explanations of the results and infer knowledge for the real-estate market.

4.1. Classification results and Comparison

The results of the classification show that most of the text representation methods combined with a classifier outperform the Regex baseline. It shows that Real Estate ads are too noisy to retrieve information easily. Nevertheless, we noticed that the predictions of number of rooms are slightly better with a regex for 2, 3 or 4 rooms. The advertisers seem to write more often the number of rooms for those classes than for '1' or '5 and more' rooms in their description.

Comparing text representations, classical methods (mainly TF-IDF) and CamemBERT achieved the best F1-Score. Non-contextual methods such as Word2Vec, Doc2Vec or FastText lagged behind. FlauBERT fared even worse. FastText and FlauBERT are more based on a character or sub-word level. However, the vocabulary of French real-estate ads is pretty poor and might not be suitable for this level. On the contrary, Doc2Vec embeds the whole paragraph and is less precise. Most paragraphs contain the same kind of vocabulary and syntax, so Doc2Vec fails to discriminate them. In general, classical methods gave better results, except for CamemBERT, which has similar results. Nevertheless, the training dataset was really small and the performance of more complex methods might improve with a larger corpus.

4.2. Discussion

The comparison of the text representations for the classifications tasks highlights that classical representations such as TD-IDF have better results than non-contextual word-embbedings (Word2vec, FastText, etc.) or quite

Table 4
Classification Type of product

Model	F1-Score (Total)	F1-Score (Apartment)	F1-Score (House)	F1-Score (Commercial)	F1-Score (Parking)
Regex	71%	74.2%	52.8%	81%	53.1%
BoW	97.8%	97.8%	98%	98%	**97.1%**
TF-IDF	**98.3%**	**98.5%**	98.5%	**98.2%**	**97.1%**
Word2Vec (frWac2Vec)	96.4%	97.7%	96.4%	95.5%	93%
Word2Vec (corpus)	96.5%	97.5%	96%	95.8%	94.3%
Doc2Vec	89.8%	93.4%	86.4%	90%	74.3%
FastText	94.4%	95.3%	94.3%	94%	91.2%
CamemBERT	98%	98.1%	**99%**	97.6%	95.8%
FlauBERT	44.2%	61.2%	0%	0%	0%

Table 5
Classification Number of Rooms

Model	F1-Score (Total)	F1-Score (1 room)	F1-Score (2 rooms)	F1-Score (3 rooms)	F1-Score (4 rooms)	F1-Score (5+ rooms)
Regex	59%	55.6%	**76%**	**75%**	**61.7%**	51%
BoW	66%	**90%**	67.1%	62%	45%	77.1%
TF-IDF	69.3%	85%	69.7%	62.6%	54%	80%
Word2Vec (frWac2Vec)	65.3%	71.4%	67.1%	57.3%	48.4%	80.6%
Word2Vec (corpus)	66.8%	77%	69.7%	61%	47.8%	80%
Doc2Vec	59.1%	69.8%	57.3%	54.3%	35%	76.7%
FastText	61.6%	66.7%	58%	52.8%	40.2%	81.5%
CamemBERT	**70.8%**	86.5%	70%	67.2%	43.4%	**85.2%**
FlauBERT	32%	0%	0%	0%	0%	48.4%

similar with state-of-the-art models (CamemBERT). This result might be explained by the lack of data and a very narrow vocabulary found in the ads.

Also, it could be surprising that text representations combined with classification algorithms outperform simple regular expression. Thus, we analysed the odds ratio of the Logistic Regression used for the classification combined with TF-IDF representations since it gave good results. Odds ratio helps to know if a variable (e.g., a word) is increasing or decreasing the probability of the text belonging to one class or another. In Tables 7, 8 and 9, we can see the words that increase or decrease the most the probability for each class. For example, the probability is obviously increased by the word that describes the class, such as "apartment", "house", for the type of product, or "two rooms", "three rooms" for the

Table 6
Classification Floor

Model	F1-Score (Total)	F1-Score (ground floor)	F1-Score (1^{st} floor)	F1-Score (2^{nd} floor)	F1-Score (3^{rd} floor)	F1-Score (high floor)	F1-Score (last floor)
Regex	40.7%	49%	67%	63.2%	18.2%	47%	**60%**
BoW	61.8%	75%	64%	**63.6%**	25%	75%	35.3%
TF-IDF	68.4%	84.8%	**76.2%**	57.1%	**26.7%**	**83.3%**	35.3%
Word2Vec (frWac2Vec)	56.6%	82%	46.7%	41.7%	14.3%	75%	26.7%
Word2Vec (corpus)	57.8%	79.4%	38.5%	41.7%	40%	75%	37.5%
Doc2Vec	50%	68.8%	48%	46.1%	0%	22%	37.5%
FastText	43.4%	74.6%	36.4%	25%	9.5%	0%	30%
CamemBERT	**72.5%**	**91.4%**	44.4%	0%	0%	0%	0%
FlauBERT	69.7%	82.1%	0%	0%	0%	0%	0%

Table 7
Word importance for the type of property

Label	Odds ratio > 1	Odds ratio < 1
Apartment	apartment, cellar, apartment complex, room, living-room	land, house, floor, single-storey, location
House	house, villa, bedroom, swimming-pool, property	apartment complex, parking, building, located, lot
Commercial property	business, hotel, wall, commercial space	apartment complex, living room, garage, quiet, view
Garage/Parking	garage, box, apartment complex, parking, basement	terrace, room, bedroom, sea

Table 8
Number of rooms

Label	Odds ratio > 1	Odds ratio < 1
1	studio apartment, investment, lot, living-room, rental	bedroom, room, house
2	two rooms, bedroom, living-room, kitchen, investment	two bedrooms, parental bedroom, villa, house, studio apartment
3	three rooms, two bedrooms, children's bedroom, crossing apartment, corner apartment	studio apartment, living-room, large room, rare, awesome
4	three bedrooms, villa, house, terrace, bourgeois building	rental, cellar, furnished, to renovate, garage
5+	villa, swimming-pool, land, house, property	apartment complex, shared property, closet, parking, small

number of rooms. However, it is interesting to see that unexpected words turn out to have an important impact. Indeed, we can see that real-estate agents target different people according to the number of rooms. They often use words about investment for one- or two-room dwellings, while they target parents with children for three-room properties. Furthermore, the vocabulary used to classify the floor of the property is also different. An agent will describe more the view for a high floor than for the ground floor. For instance, the expression "panoramic view" is associated to "Last Floor".

In a nutshell, this study points out the specific and stylistic vocabulary used by French real-estate agents. As we have shown, basic information is not always clearly and explicitly written in the ads. However, other words can help to infer such key information for the real-estate market.

5. Conclusion and Further Work

In this paper, we carried out a comparison of various text representations models with respect to their application to the classification of real-estate classified advertisements in French, having the ultimate goal of extracting key information about the properties they advertise.

In summary, we found out that, among the models we tested, the classic representation TF-IDF and the most state-of-the-art model CamemBERT are the ones that stand out for the classification task.

Another interesting finding is that pre-trained models, despite having been trained on a much larger corpus and with impressive computational resources, turn out not to be much useful, due to their being domain-agnostic, whereas models trained on our small domain-specific corpus clearly outperform them, thus confirming similar findings in other domains [12].

Finally, we provided knowledge about the vocabulary and the type of register use by the real-estate agents according to the property they advertise.

We believe that our results will provide useful guidance to anybody willing to engage in real-estate classified ads classification, in general, and information extraction in particular.

A promising research direction which is, in our opinion, worth investigating is to combine different text representations models, each capable of capturing different details, in order to obtain a higher overall accuracy. Another quite obvious extension of our investigation would be to gather a bigger corpus of advertisements and study how using it to train the model increases the classification accuracy. Large language models could also be considered, as a possible end-to-end solution to this information extraction task.

As future work, it could be interesting to capture the syntax of real-estate ads in order to understand even better their language. Also, property pictures, which of-

Table 9
Floor of the property

Label	Odds ratio > 1	Odds ratio < 1
Ground floor	ground floor, quiet, living conditions	view, entry, balcony, last floor, shared property
1	first floor, storage area, living-room, standing, zone	last floor, view, sea, second floor
2	second floor, owner, shared property, lot, storage	
3	third floor, cellar, crossing apartment, beautiful view, orientation	space, bathroom, secure
High floor	fourth floor, high, hills, small sea view, last floor	shared property, gate, children, visit
Last floor	last floor, large living-room, residential area, sea, panoramic view	shared property, lot

ten come together with the advertisements, are a major source of information that can contribute to the intelligence of the real-estate market.

Acknowledgments

This research was carried out in the Wimmics team, which is a joint research team of Université Côte d'Azur, Inria, and I3S. Our research motto: AI in bridging social semantics and formal semantics on the Web.

This work has been partially supported by the French government, through the 3IA Côte d'Azur "Investments in the Future" project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002.

References

[1] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K. Q. Weinberger (Eds.), Advances in Neural Information Processing Systems, volume 26, Curran Associates, Inc., 2013. URL: https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.

[2] Q. V. Le, T. Mikolov, Distributed representations of sentences and documents, in: Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014, volume 32 of *JMLR Workshop and Conference Proceedings*, JMLR.org, 2014, pp. 1188–1196. URL: http://proceedings.mlr.press/v32/le14.html.

[3] J. Pennington, R. Socher, C. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1532–1543. URL: https://www.aclweb.org/anthology/D14-1162. doi:10.3115/v1/D14-1162.

[4] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, arXiv preprint arXiv:1607.04606 (2016).

[5] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: Proc. of NAACL, 2018.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[7] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://www.aclweb.org/anthology/N19-1423. doi:10.18653/v1/N19-1423.

[8] N. Usman, R. Imran, K. S. Khalid, P. Mukesh, A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models, 2020. URL: http://arxiv.org/abs/2010.15036.

[9] E. Dynomant, R. Lelong, B. Dahamna, C. Masson-

naud, G. Kerdelhué, J. Grosjean, S. Canu, S. J. Darmoni, Word Embedding for the French Natural Language in Health Care: Comparative Study, JMIR Med Inform 7 (2019). URL: http://www.ncbi.nlm.nih.gov/pubmed/31359873. doi:10.2196/12310.

[10] E. Altszyler, M. Sigman, S. Ribeiro, D. F. Slezak, Comparative study of LSA vs Word2vec embeddings in small corpora: a case study in dreams database, Consciousness and Cognition (2017) 178–187. URL: http://arxiv.org/abs/1610.01520. doi:10.1016/j.concog.2017.09.004.

[11] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. E. Barnes, D. E. Brown, Text Classification Algorithms: A Survey, arXiv:1904.08067 [cs, stat] (2020). URL: http://arxiv.org/abs/1904.08067. doi:10.3390/info10040150, arXiv: 1904.08067.

[12] S. Gupta, Z. Waseem, A Comparative Study of Embeddings Methods for Hate Speech Detection from Tweets (2018).

[13] J.-P. Fauconnier, French word embeddings, 2015. URL: http://fauconnier.github.io.

[14] L. Martin, B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, É. de la Clergerie, D. Seddah, B. Sagot, CamemBERT: a tasty French language model, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7203–7219. URL: https://www.aclweb.org/anthology/2020.acl-main.645.

[15] H. Le, L. Vial, J. Frej, V. Segonne, M. Coavoux, B. Lecouteux, A. Allauzen, B. Crabbé, L. Besacier, D. Schwab, FlauBERT: Unsupervised language model pre-training for French, in: Proceedings of the 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 2479–2490. URL: https://www.aclweb.org/anthology/2020.lrec-1.302.

[16] P. J. Ortiz Suárez, B. Sagot, L. Romary, Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures, Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, Leibniz-Institut für Deutsche Sprache, Mannheim, 2019, pp. 9 – 16. URL: http://nbn-resolving.de/urn:nbn:de:bsz:mh39-90215. doi:10.14618/ids-pub-9021.

[17] R. Sennrich, B. Haddow, A. Birch, Neural machine translation of rare words with subword units, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 1715–1725. URL: https://www.aclweb.org/anthology/P16-1162. doi:10.18653/v1/P16-1162.

[18] T. Kudo, J. Richardson, SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 66–71. URL: https://www.aclweb.org/anthology/D18-2012. doi:10.18653/v1/D18-2012.