REAL-UP: Urban Perceptions From LBSNs Helping Moving Real-Estate Market to the Next Level

Frances A. Santos
University of Campinas
Campinas, São Paulo, Brazil
frances.santos@ic.unicamp.br

Thiago H. Silva
Univ. Tecnológica Federal do Paraná
Curitiba, Paraná, Brazil
thiagoh@utfpr.edu.br

Leandro A. Villas
University of Campinas
Campinas, São Paulo, Brazil
lvillas@unicamp.br

ABSTRACT

Finding the best short- or long-term accommodation is troublesome in unknown areas. Current tools provided by the real-estate market offer valuable information regarding the property, such as price, photos, and descriptions of the space; however, this market has little explored other relevant information regarding the surrounding area, such as what is nearby and users' subjective perception of the property's area. To address this gap, we propose REAL-UP, an interactive tool designed to enrich real-estate marketplaces. In addition to information commonly provided by such applications, e.g., rent price, REAL-UP also provides subjective neighborhood information based on Location-Based Social Networks (LBSNs) messages. This novel tool helps to represent complex users' subjective perceptions of urban areas, which could ease the process of finding the best accommodation.

CCS CONCEPTS

• **Information systems → Location based services**; *Sentiment analysis.*

KEYWORDS

Urban Perceptions, Generative AI, Sentiment Analysis, LBSN Data

ACM Reference Format:
Frances A. Santos, Thiago H. Silva, and Leandro A. Villas. 2024. REAL-UP: Urban Perceptions From LBSNs Helping Moving Real-Estate Market to the Next Level. In *Companion Proceedings of the ACM Web Conference 2024 (WWW '24 Companion), May 13–17, 2024, Singapore, Singapore.* ACM, New York, NY, USA, 4 pages.

1 INTRODUCTION

For several reasons, finding short/long-term accommodation is a common task in people's lives. Nevertheless, a typical problem people who move to another unknown area face is finding the best property. To deal with this, the real-estate marketplace has provided tools with detailed information regarding the property, such as price, parking spaces, description of rooms, photos, etc. However, this market has still little explored other relevant information regarding the property's surrounding area. Decisions to rent/buy could be influenced by further details regarding what amenities are nearby, such as schools and parks, and users' subjective perceptions regarding the area. While information regarding what is nearby is relatively easy to offer automatically at a large scale, users' urban perceptions are not.

Location-Based Social Networks (LBSNs), especially those focused on enabling users to review venues, such as Yelp, TripAdvisor, and Google Places, can be valuable in extracting urban perceptions [4, 5]. LBSNs are also attractive due to the significant volume of public data generated by their users in different areas around the globe [5]. Nevertheless, there are several challenges when using LBSNs for urban perception extraction. For instance, in LBSNs for reviewing venues, reviews are primarily for indoor venues – e.g., restaurants and hotels. Thus, opinions unlinked to a venue, i.e., perceptions regarding an area shared while transiting in the city, when they exist, tend to be mixed with topics regarding the venue, becoming more difficult for the proper perception extraction. Also, other LBSNs could enable geolocation of content about diverse topics, not only regarding urban areas, such as X – formerly Twitter. Thus, extracting and analyzing LBSN data demands sophisticated approaches to discovering interesting characteristics and dynamics of cities to better understand multiple urban phenomena and enable the development of smarter real-estate services for society.

To this end, in our previous work [4], we proposed a framework to enable the automatic extraction of urban outdoor perceptions from LBSNs, particularly messages containing natural language texts written in English. That work also validated the proposed method – please refer to the study for details. Taking advantage of urban perceptions extracted using this framework, we propose an interactive tool named Real-Estate Urban Perceptions (REAL-UP) in this work to enhance the real-estate marketplace by providing rich knowledge regarding urban areas in the form of interactive 2D maps. In addition to information commonly provided by such applications, such as rent price and property type, REAL-UP also provides subjective neighborhood information, more specifically, the emotion and sentiment perceived, and a short review generated by a Large Language Model (LLM) based on LBSN messages shared by users. By doing that, REAL-UP helps to describe complex subjective perceptions perceived by people in urban areas, which has the potential to help people in the decision-making process when selecting a place to rent/buy. To the best of our knowledge, REAL-UP is the first tool that shows the potential of considering users' urban perceptions to enrich real estate market tools.

2 METHODOLOGY

2.1 Data

Using the X API, we collected tweets shared by X users in Chicago and New York City (NYC), United States, and London, United Kingdom, from January 9th to August 13th, 2018. Besides the texts, X

WWW '24 Companion, May 13–17, 2024, Singapore, Singapore

data contains useful metadata: timestamp, geolocation, and a unique identifier. We collected 9,210,925 tweets for Chicago, 17,646,407 for NYC, and 12,043,181 for London. From the total, we only consider the geotagged data (i.e., those with geolocation information) that were predicted as "Urban Perception" by the framework proposed in [4].

Besides that, to obtain accurate information regarding properties, we collected data from the initiative Inside Airbnb[1], which provides data and advocacy about Airbnb's impact on residential communities in several cities around the world, including those evaluated in this work: Chicago, IL; New York City, NY; and, Greater London, London. Among the data available on this website, we used the file called "listings.csv," which contains summary information (quarterly data for the last 12 months) and metrics for properties of each city, such as latitude, longitude, price, and room type (entire home/apt, private room, and shared room). We collected this file for three cities evaluated on January 25th, 2022.

Ethical considerations. This study does not involve experiments with human subjects. All sensitive information in our datasets (i.e., account name and account ID) has been previously anonymized to ensure users' privacy. Besides, the data explored is publicly available.

2.2 Sentiment and Emotion Analysis

We explore a Google Research dataset called GoEmotions to enable machines to understand human emotions [1]. GoEmotions is a human-annotated dataset for fine-grained emotion classification with 27 emotion categories [1]. Different from previous emotion datasets, which are closed-domain (i.e., focus on a specific domain, such as news, headlines, movie subtitles, and fairy tales), the GoEmotions dataset has a broader scope, once it contains 58K Reddit comments, all in English language, about diverse contexts, which were extracted from popular subreddits [1]. Besides fine-grained emotion categories, GoEmotions also groups its emotions in terms of their sentiment (negative, positive, and ambiguous) and the six basic emotions (anger, surprise, disgust, joy, fear, and sadness) [1]. There is also an additional category called "neutral" for sentiment and basic emotions, which indicates the absence of sentiment or emotion in the text analyzed [1].

Thus, we leverage labeled data from the GoEmotions dataset to train a supervised text classification model to predict emotions and sentiments. But, before that, we made two minor changes: (*i*) GoEmotions was built for the multi-label classification task, where one or more labels can be attributed to a single expression. So, we removed all occurrences with more than one label from the train and test datasets. This operation reduced the training dataset from 43,410 to 36,308 (around 16%), and the test dataset from 5,427 to 4,590 (around 15%); and, (*ii*) GoEmotions dataset uses a special tag "[NAME]" to mask the name of authors of Reddit comments. We replaced it with the tag "USERNAME," used in this work.

As the performance reported in [1] using the 27 emotion categories is significantly lower than the Elkman group model (i.e., six basic emotions), F1-Score= 0.46 against F1-Score= 0.64, respectively, we decided to use the six basic emotions instead of 27 emotions to training our model, as described below.

Table 1: Our model performance of predicting emotions in the GoEmotions dataset.

Ekman Emotion	Precision	Recall	F1	Support
anger	0.45	0.32	0.37	520
fear	0.39	0.44	0.41	76
neutral	0.44	0.57	0.50	77
joy	0.76	0.69	0.72	1,603
surprise	0.54	0.73	0.62	1,606
sadness	0.49	0.42	0.45	259
disgust	0.45	0.22	0.30	449
macro-average	**0.50**	**0.48**	**0.48**	**4590**
weighted-average	**0.59**	**0.59**	**0.58**	**4590**

Initially, we use the *SentenceBERT* (SBERT) [3] model to generate the sentence embeddings for every sentence available in the GoEmotions dataset. SBERT allows the generation of embeddings with the same generalization capacity, but with computational complexity orders of magnitude lower than its predecessor models, such as InferSent and USE [3]. SBERT is available through an open-source Python library called Sentence Transformes[2], which provides more than 100 pre-trained models for different NLP tasks. Among those models, we selected the model called "all-MiniLM-L6-v2," where the term "all" indicates that the model has been trained with all available data (over 1 billion training pairs) and is designed as a general-purpose model. Also, the term "Mini" refers to the small version of the base model ("all-mpnet-base-v2"), reducing the model size from 420MB to 120MB, becoming 5x faster without degrading its performance significantly[3].

Then, we use the Lazy Predict library[4] to assess which is the best supervised classifier with the GoEmotions dataset. The classifiers Logistic Regression, Linear Discriminant Analysis (LDA), and SVC obtained the best performance in terms of F1-Score: 0.59, 0.59, and 0.58, respectively. However, LDA takes less time than others: 3.37 seconds against 8.01 seconds of Logistic Regression and 1325.02 seconds of SVC. For this reason, we choose the LDA classifier for emotion classification.

Table 1 shows the performance our model achieves. Note that our performance is a bit worse than the performance in [1] (although not far off); however, our model demands less computational resources and should be much faster because it uses a BERT-base model, favoring scalability and (near) real-time tasks.

Based on the emotion predicted, we also define the polarity of sentiment for every given sentence according to the taxonomy proposed by the GoEmotions dataset [1], where: *positive* sentiment is associated with the emotion *joy*; *negative* is associated with emotions *anger, disgust, fear, and sadness*; *ambiguous* is associated with the emotion *surprise*; and, when the emotion is *neutral*, the sentiment also is defined as *neutral*.

2.3 Zero-Shot Prompting

In addition to sentiment and emotion analysis, we also explore LLMs to elaborate meaningful reviews about urban areas based on crowd

[1] http://insideairbnb.com/get-the-data.html.

[2] https://www.sbert.net.

[3] SBERT's pre-trained models: https://www.sbert.net/docs/pretrained_models.html.

[4] https://lazypredict.readthedocs.io/en/latest/.

knowledge, helping to better understand cities' neighborhoods, and enabling a smarter real-estate service. In this sense, first, we elaborate the following prompt:

What perceptions does {{URBAN AREA}} trigger in citizens and visitors?

, where {{URBAN AREA}} is a tag that is replaced by some real urban area.

For instance, using OpenAI ChatGPT (version gpt-3.5-turbo-0613 and $temperature = 0.2$) and replacing the {{URBAN AREA}} by "Lincoln Park Zoo", "Central Park", or "St. James's Park", we obtained good answers, including the main activities people execute in such urban areas, which might be very useful to know when you are planning to visit these places. Nevertheless, the results are often biased (e.g., only informing the positive perceptions), which might be a consequence of bias in the training data, where the model tries to please those who are interacting with it to receive more rewards, due to the learning process using Reinforcement Learning with Human Feedback (RLHF), as discussed in [2]. Also, most of the answers we got are too generic, with slight changes for distinct urban areas, which does not help us understand the singularities of every place.

Aiming to mitigate such problems, we included samples of urban perceptions in the prompt, considering sentences related to urban perceptions associated with some neighborhoods and shared by LBSN users (as described in Section 2.1). Moreover, we rewrite the prompt to better instruct the LLM to generate more specific and trustworthy answers. Thus, the new prompt template is:

Elaborate a detailed review of experiences and activities that can happen in {{neighborhood}}, {{city}}, based on comments below giving some examples to fundament it. Don't be repetitive.

− − −

Comments:

{{comment 0}}, ..., {{comment N}}

, where {{neighborhood}}, {{city}} define the desired place, and {{comment 0}}, ..., {{comment N}}, is the list of sentences containing the urban perceptions. To avoid extrapolating the context window size of LLM (i.e., the maximum number of input tokens – 4,096 tokens for gpt-3.5-turbo-0613) and for cost savings, we limited to 100 samples of sentences (i.e., $N < 100$). When we had more than 100 sentences available for the {{neighborhood}}, {{city}}, we selected a random sample of sentences after removing duplicated ones.

By using OpenAI ChatGPT again with the same conditions (i.e., version gpt-3.5-turbo-0613 and $temperature = 0.2$), we could observe a significant improvement in the answers obtained, showing ChatGPT was able to incorporate knowledge from comments inserted in the prompt to generate responses more specific, including possible issues that people might face in such areas, and reducing possible hallucinations, since it pays more attention to the content inputted instead of creatively generate the response without any prior context.

For instance, Figure 1(e) shows an example of ChatGPT's response considering the neighborhood Loop, Chicago, IL. As we can see, the response contains a reference to religious events and the architecture of the chapel, which was motivated by the comment *"sky chapel travel chicago skychapel methodist religion church URL,"* where a person shared that he/she traveled to Chicago and visited the Methodist Church Sky Chapel. Similarly, we can see references to some popular places at Loop for the diverse experiences (culture, art, nightlife, breathtaking views, etc.) they can offer to their visitors, which are commonly shared by them on LBSNs, as shown in some comments present in the prompt: *"Chicago on a perfect evening with fun peeps at smss cindy s rooftop URL," "just chilling on the willis tower en skydeck chicago URL",* and *"mural on back of chicago cultural centrer chicago cultural center URL."* Finally, the response also contains relevant traffic information, warning people to plan and check for traffic updates before visiting the region. Such information is very often on comments shared by people, for example, *"closed due to a bridge lift in lakeshoredrive on lk shore dr nb," "between wacker dr and grand ave traffic chicago URL,"* and *"closed due to accident in loop on e wacker dr both eb wb between la salle st and state st traffic chicago URL."*

Besides OpenAI ChatGPT, we also conducted experiments using some open-source LLMs, such as Mistral 7B and Falcon 7B; however, the responses generated by them were not satisfactory, since they generated unintelligible texts and repeated some comments present in the prompt, showing they failed to understand and execute the task properly, by using zero-shot prompting.

3 SYSTEM OVERVIEW

REAL-UP[5] is a RESTful API developed in Python 3.7.6 using a modern and fast web framework called FastAPI[6]. Then, we use a Docker[7] container running locally, where the hardware used to execute it was a processor I7-7600U with 2 cores and 4 threads, 16 gigabytes of RAM with 2400 MT/s, to interact with REAL-UP via web browsers.

As we can see in Figure 1(a), REAL-UP's homepage displays a short description and requires users to input two information: select one of the cities available, by default Chicago is selected; the period granularity, where "Last month" (default value) consider data from July 2018 to August 2018, "Last 3 months" consider data from May 2018 to August 2018, and "Whole period" consider data from January 2018 to August 2018.

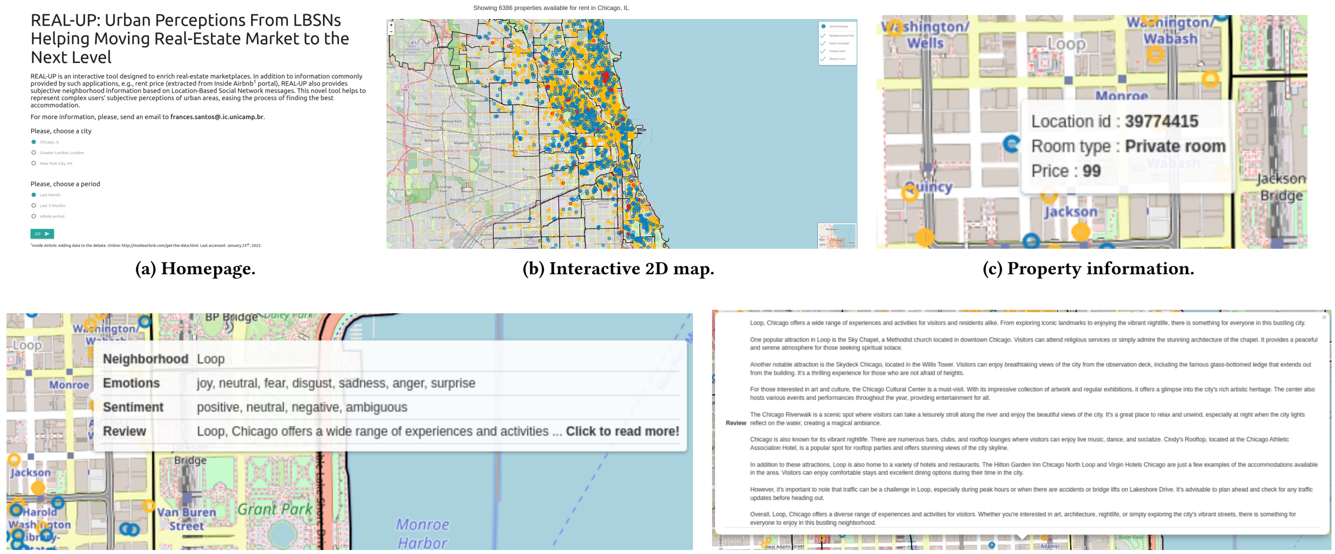
After providing this information and clicking on the button "Go," users are redirected to a new web page, where an interactive 2D map, built using the Folium library[8], is displayed. Also, a message informing the number of properties available for rent is displayed on top of the page, as shown in Figure 1(b) for Chicago, IL. As we can see, the neighborhoods' boundaries are highlighted with strong black lines. Small colorful circles represent the properties, where each color refers to the room type: red for shared rooms, yellow for the entire home/apartment, and blue for private rooms. Moreover, we can use the small map in the right-bottom corner to facilitate our localization in the whole city, which can be helpful when we increase the zoom. Also, we can apply filters to remove or add information on the map at any time by disabling/enabling them using the top-right menu.

[5] REAL-UP's repository: https://doi.org/10.5281/zenodo.10802458.

[6] https://fastapi.tiangolo.com/.

[7] https://www.docker.com/.

[8] https://python-visualization.github.io/folium/latest/.



REAL-UP: Urban Perceptions From LBSNs Helping Moving Real-Estate Market to the Next Level

REAL-UP is an interactive tool designed to enrich real-estate marketplaces. In addition to information commonly provided by such applications, e.g., rent price (extracted from Inside Airbnb[1] portal), REAL-UP also provides subjective neighborhood information based on Location-Based Social Network messages. This novel tool helps to represent complex users' subjective perceptions of urban areas, easing the process of finding the best accommodation.

For more information, please, send an email to **frances.santos@.ic.unicamp.br**.

Please, choose a city

● Chicago, IL
○ Greater London, London
○ New York City, NY

Please, choose a period

● Last Month
○ Last 3 Months
○ Whole period

GO ➤

[1] Inside Airbnb: Adding data to the debate. Online: http://insideairbnb.com/get-the-data.html. Last accessed: January 25th, 2022.

(a) Homepage.

Showing 6386 properties available for rent in Chicago, IL

(b) Interactive 2D map.

Location id : **39774415**
Room type : **Private room**
Price : **99**

(c) Property information.

Neighborhood	Loop
Emotions	joy, neutral, fear, disgust, sadness, anger, surprise
Sentiment	positive, neutral, negative, ambiguous
Review	Loop, Chicago offers a wide range of experiences and activities ... **Click to read more!**

(d) Neighborhood information.

Review

Loop, Chicago offers a wide range of experiences and activities for visitors and residents alike. From exploring iconic landmarks to enjoying the vibrant nightlife, there is something for everyone in this bustling city.

One popular attraction in Loop is the Sky Chapel, a Methodist church located in downtown Chicago. Visitors can attend religious services or simply admire the stunning architecture of the chapel. It provides a peaceful and serene atmosphere for those seeking spiritual solace.

Another notable attraction is the Skydeck Chicago, located in the Willis Tower. Visitors can enjoy breathtaking views of the city from the observation deck, including the famous glass-bottomed ledge that extends out from the building. It's a thrilling experience for those who are not afraid of heights.

For those interested in art and culture, the Chicago Cultural Center is a must-visit. With its impressive collection of artwork and regular exhibitions, it offers a glimpse into the city's rich artistic heritage. The center also hosts various events and performances throughout the year, providing entertainment for all.

The Chicago Riverwalk is a scenic spot where visitors can take a leisurely stroll along the river and enjoy the beautiful views of the city. It's a great place to relax and unwind, especially at night when the city lights reflect on the water, creating a magical ambiance.

Chicago is also known for its vibrant nightlife. There are numerous bars, clubs, and rooftop lounges where visitors can enjoy live music, dance, and socialize. Cindy's Rooftop, located at the Chicago Athletic Association Hotel, is a popular spot for rooftop parties and offers stunning views of the city skyline.

In addition to these attractions, Loop is also home to a variety of hotels and restaurants. The Hilton Garden Inn Chicago North Loop and Virgin Hotels Chicago are just a few examples of the accommodations available in the area. Visitors can enjoy comfortable stays and excellent dining options during their time in the city.

However, it's important to note that traffic can be a challenge in Loop, especially during peak hours or when there are accidents or bridge lifts on Lakeshore Drive. It's advisable to plan ahead and check for any traffic updates before heading out.

Overall, Loop, Chicago offers a diverse range of experiences and activities for visitors. Whether you're interested in art, architecture, nightlife, or simply exploring the city's vibrant streets, there is something for everyone to enjoy in this bustling neighborhood.

(e) Full review.

Figure 1: System interfaces. [Best in color]

To know more information about any property, we just need to point the mouse at the circle; then a pop-up window will open with the following information: *Location id*, a unique identifier of property in Inside Airbnb data; *Room type*; and, *Price*, as shown in Figures 1(c). Similarly, as Figures 1(d) shows, we can point the mouse at any neighborhood to obtain the following information regarding it: *Neighborhood*; *Emotions*; *Sentiment*; and, *Review*. As we can see, all sentiments and emotions occur in Loop, Chicago, which are displayed in an orderly manner, from the most relevant to the less relevant one. Then, as indicated in Figure 1(d), we can click on the left mouse button to access the full review of the neighborhood, as shown in Figure 1(e).

In this way, our real-estate demo can provide a better understanding of the urban areas of surrounding properties, helping people choose the best places to stay by providing richer information than information commonly provided by real-estate applications.

4 CONCLUSION

We propose REAL-UP, the first interactive tool to enhance the real-estate marketplace with subjective urban perceptions. REAL-UP combines the emotion and sentiment users perceive, besides a short review generated by OpenAI ChatGPT based on LBSN messages for the city’s neighborhood to provide rich knowledge regarding urban areas through interactive 2D maps. Thus, users can better understand the urban areas of surrounding properties while choosing a place to stay. Nevertheless, we can also mention some challenges we faced in building this tool. First, defining a good way to present all relevant information about properties and urban areas without affecting the user experience was challenging. Moreover, LLMs can often write plausible sounding but incorrect or nonsensical answers (phenom known as hallucinations), making it difficult for people with no prior knowledge about the urban area to distinguish whether or not the information is accurate. In this sense, it would be advisable for applications leveraging our tool to address these points to guarantee great user use experiences and avoid misinformation.

ACKNOWLEDGMENTS

FAS and LAV were financed by the Ministry of Science, Technology, and Innovation of Brazil, with resources granted by the Federal Law 8.248 of October 23, 1991, under the PPI-Softex. The project was coordinated by Softex and published as Intelligent agents for mobile platforms based on Cognitive Architecture technology [01245.013778/2020-21]. THS was supported by the project SocialNet (grant 2023/00148-0 from Sao Paulo Research Foundation - FAPESP) and CNPq (process 314603/2023-9 and 441444/2023-7).

REFERENCES

[1] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547* (2020).

[2] Emilio Ferrara. 2023. Should chatgpt be biased? challenges and risks of bias in large language models. *arXiv preprint arXiv:2304.03738* (2023).

[3] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

[4] Frances A Santos, Thiago H Silva, Antonio AF Loureiro, and Leandro A Villas. 2020. Automatic extraction of urban outdoor perception from geolocated free texts. *Social Network Analysis and Mining* 10 (2020), 1–23.

[5] Thiago H. Silva, Aline Carneiro Viana, Fabrício Benevenuto, Leandro Villas, Juliana Salles, Antonio Loureiro, and Daniele Quercia. 2019. Urban Computing Leveraging Location-Based Social Network Data: A Survey. *ACM Comput. Surv.* 52, 1, Article 17 (Feb. 2019), 39 pages. https://doi.org/10.1145/3301284