A hybrid framework for enriching urban gazetteers by extracting geographic information from online housing listings

Mahdi Shakhesi, Ali Asghar Alesheikh*

Department of Geospatial Information System (GIS), Faculty of Geodesy and Geomatics Engineering, K. N. Toosi University of Technology, Tehran, Iran

* Corresponding Author: alesheikh@kntu.ac.ir

Abstract

Introduction: Place names, a common form of embedded geographic information in natural language texts, are used in various resources such as social media, news stories, historical archives, and property listings. The names are presented in different forms like business addresses, hashtags, or simple texts. Providing up-to-date data, carrying human experience and cognition, and containing types of geospatial information only available in textual resources make these resources precious for geospatial analyses. Therefore, mapping place names to their footprints is an essential task. One of the solutions for this task is using a digital gazetteer, a dictionary of place names. These precious resources enable Geographic Information Retrieval (GIR) systems to detect place names (geotagging) and convert the candidate ones to their geographic coordinates (geocoding). To fulfill ever-increasing geospatial demands, especially in GIR and LBSs, digital gazetteers should be enriched.

Materials and Methods: This paper presents a three-tier framework to extract urban geographic information from geotagged housing listings. The first tier is devoted to harvesting main street and neighborhood place names, which the authors usually write without any linguistic clue due to their well-knownness. Using a random forest model based on a set of spatial measures for each extracted n-gram from the textual content of real estate advertisements enables us to identify the main streets and neighborhoods. The first tier commences with the extraction of n-grams from the saved advertisements. After cleaning and standardizing the n-gram set, spatial clustering is applied, considering that each spatial n-gram can refer to multiple regions of the city. The defined spatial predictors are computed for each not-clustered n-gram or split n-gram from its generic cluster. Subsequently, a random forest model identifies the neighborhood and the main street n-grams. We developed a rule-based model to extract all urban place names in the second tier and a linguistic pattern-based model to extract spatial relationships in the third tier. This research focused on the Persian language and Tehran, Mashhad, Isfahan, and Shiraz metropolises from Iran as study regions.

Results and Discussion: The results are encouraging for the first tier, specifically achieving approximately 0.8 and 0.7, respectively, for recall and precision in predicting another metropolis's main streets and neighborhoods. However, differences in population levels and urban development patterns decreased the performance in identifying a neighborhood as a main street or vice versa. For the second tier, precision and recall are near 0.7. Although these results are notable compared to the performance of named entity recognition models in extracting urban place names which are often fine-grained, errors in this layer have led to reduced precision and recall in the third layer, spatial relation extraction.

Conclusion: Gazetteers are important geospatial resources in GIR tasks, especially in geoparsing. This paper presented a framework for extracting urban geographic information from online property listings. This geographic information includes the place names and the spatial relationships to enrich current gazetteers. Since main streets and neighborhoods as a part of place names are well-known, people mainly use them without any clue on property listing websites. Harvesting these place names can be done using a machine learning-based model. The next step is extracting all place names written in the property advertisement posts. To realize that, we developed a rule-based model to extract potential place names from the posts geographically located in the neighborhood/main street place name's convex-hull and remove the wrong identified cases. In the third step, we extracted spatial relationships between the place names extracted from each post text based on linguistic patterns. The framework has provided good results in harvesting main streets and neighborhoods and extracting place names. Extracting spatial relationships between the place names needs further work.

Keywords: gazetteer enrichment; spatial relationship extraction; geographic information retrieval; real estate advertisement; random forest

یک چارچوب ترکیبی برای غنی‌سازی جاینامه‌های شهری با استخراج اطلاعات جغرافیایی از آگهی‌های املاک

مهدی شاخصی، علی اصغر آل شیخ*

گروه سیستم‌های اطلاعات مکانی (*GIS*)، دانشکده مهندسی نقشه‌برداری، دانشگاه صنعتی خواجه نصیرالدین طوسی، تهران، ایران

*نویسنده عهده دار مکاتبات : alesheikh@kntu.ac.ir

چکیده

مقدمه: جاینام‌ها به عنوان یکی از رایج‌ترین اشکال اطلاعات جغرافیایی نهفته در متون زبان طبیعی، در منابع مختلفی همچون رسانه‌های اجتماعی، اخبار، آرشیوهای تاریخی و آگهی‌های املاک به‌کار می‌روند. این نام‌ها در قالب‌های گوناگونی نظیر نشانی کسب‌وکار، هشتگ، یا متن ساده ممکن است دیده شوند. به‌هنگام بودن داده‌ها، دربرگرفتن تجربه و شناخت انسانی، و دربرداشتن انواع خاصی از اطلاعات مکانی که صرفاً در منابع متنی موجود هستند، این منابع را برای تحلیل‌های جغرافیایی بسیار ارزشمند ساخته است. از این رو، نگاشت نام مکان‌ها به موقعیت جغرافیایی آن‌ها یک امر ضروری است. یکی از راه‌کارهای موجود، استفاده از جاینامه‌های رقمی است که در واقع فرهنگ لغتی از نام مکان‌ها هستند. این منابع ارزشمند به سامانه‌های بازیابی اطلاعات جغرافیایی (GIR)امکان شناسایی جاینام‌ها و تبدیل موارد شناسایی‌شده به مختصات جغرافیایی را می‌دهند. با توجه به کاربردهای روزافزون مکانی، به‌ویژه در GIR و خدمات مبتنی بر مکان(LBS) ، جاینامه‌های رقمی باید غنی‌سازی شوند.

مواد و روش‌ها: این مقاله چارچوبی سه‌لایه برای استخراج اطلاعات جغرافیایی شهری از آگهی‌های املاک که کدگذاری مکانی شده‌اند را ارائه می‌دهد. لایه نخست به استخراج نام‌های مکان مربوط به خیابان‌های اصلی و محله‌ها اختصاص دارد که به‌دلیل شناخته‌شده بودنشان، معمولاً بدون هیچ سرنخ زبانی توسط نویسندگان نوشته می‌شوند. با استفاده از یک مدل جنگل تصادفی مبتنی بر مجموعه‌ای از معیارهای مکانی برای هر ان‌گرم استخراج‌شده از محتوای متنی آگهی‌ها، می‌توان خیابان‌های اصلی و محله‌ها را شناسایی کرد. این لایه با استخراج ان‌گرم‌ها از آگهی‌های وبکاوی‌شده آغاز می‌شود. با توجه به اینکه هر ان‌گرم ممکن است به چند ناحیه از شهر اشاره کند، خوشه‌بندی مکانی پس از پاک‌سازی و استانداردسازی مجموعه ان‌گرم‌ها اعمال می‌گردد. معیارهای مکانی تعریف‌شده برای هر خوشه شناسایی‌شده ان‌گرم محاسبه می‌شوند. سپس یک مدل جنگل تصادفی برای شناسایی ان‌گرم‌های محله و خیابان اصلی به کار گرفته می‌شود. در لایه دوم، یک مدل مبتنی بر قواعد برای استخراج همه نام‌های مکان شهری توسعه یافته و در لایه سوم، یک مدل مبتنی بر الگوهای زبانی برای استخراج روابط مکانی طراحی شده است. این پژوهش بر زبان فارسی و کلان‌شهرهای تهران، مشهد، اصفهان و شیراز تمرکز دارد.

نتایج و بحث: نتایج برای لایه اول با دستیابی به حدود ۰/۸ و ۰/۷ به‌ترتیب برای بازیابی و دقت در پیش‌بینی خیابان‌های اصلی و محله‌ها در کلان‌شهری دیگر رضایت‌بخش است. با این حال، تفاوت در جمعیت و الگوهای توسعه شهری باعث شده است که به خاطر شناسایی خیابان‌های اصلی به عنوان محله و بالعکس، تعداد موارد شناسایی‌شده درست کاهش یابد. در شناسایی جاینام‌های شهری دقت و بازیابی نزدیک به ۰/۷ کسب شده است. هر چند این مقادیر در مقایسه با عملکرد مدل‌های شناسایی موجودیت‌های اسمی در

استخراج جاینام‌های شهری که اغلب ریزدانه هستند، قابل توجه است ولی شناسایی‌های اشتباه در این لایه موجب کاهش دقت و بازیابی در لایه سوم یعنی استخراج روابط مکانی شده است.

نتیجه‌گیری: این پژوهش چارچوبی برای استخراج اطلاعات جغرافیایی شهری از آگهی‌های املاک ارائه می‌کند. این اطلاعات شامل جاینام‌ها و روابط مکانی برای غنی‌سازی جاینامه‌های موجود است. از آنجا که خیابان‌های اصلی و محله‌ها بخشی از نام‌های مکان شناخته‌شده هستند، افراد عموماً آن‌ها را بدون هیچ سرنخی در وب‌سایت‌های آگهی ملکی استفاده می‌کنند. استخراج این نام‌ها را می‌توان با مدل یادگیری ماشینی انجام داد. گام بعدی، استخراج همه نام‌های مکان نوشته‌شده در متن آگهی‌هاست. برای تحقق این هدف، یک مدل مبتنی بر قواعد توسعه داده شده است تا جاینام‌های محتمل را از آگهی‌هایی که موقعیت جغرافیایی‌شان در محدوده پوش محدب نام خیابان اصلی یا محله قرار دارد، استخراج کرده و موارد نادرست را حذف کند. در گام سوم، روابط مکانی بین جاینام‌های شناسایی‌شده از متن هر آگهی با استفاده از الگوهای زبانی استخراج شدند. چارچوب عملکرد خوبی در استخراج خیابان‌های اصلی، محله‌ها، و نام‌های مکان نشان داده است، اما استخراج روابط مکانی نیاز به توسعه بیشتری دارد.

کلیدواژه‌ها: غنی‌سازی جاینامه، استخراج روابط مکانی، بازیابی اطلاعات مکانی، آگهی‌های املاک، مدل جنگل‌های تصادفی

1. Introduction

Place names, a common form of embedded geographic information in natural language texts, are used in various resources such as social media, news stories, historical archives, and property listings. The names are presented in different forms like business addresses, hashtags, or simple texts (Grace, 2021; Y. Hu, Mao, & McKenzie, 2019; Middleton & Krivcovs, 2016; Won, Murrieta-Flores, & Martins, 2018; W. Zhang & Gelernter, 2014). Providing up-to-date data, carrying human experience and cognition, and containing types of geospatial information only available in textual resources make these resources precious for geospatial analyses (Y. Hu & Adams, 2021). Therefore, mapping place names to their footprints is an essential task. One of the solutions for this task is using a digital gazetteer, a dictionary of place names. Digital gazetteers are the bridge between human-oriented place names and machine-oriented coordinates. These precious resources enable Geographic Information Retrieval (GIR) systems to detect place names (geotagging) and convert the candidate ones to their geographic coordinates (geocoding) (Goodchild & Hill, 2008; Hill, 2000). These two tasks together form the geoparsing stage in a GIR process (Gritta, Pilehvar, Limsopatham, & Collier, 2018; Karimzadeh, Pezanowski, MacEachren, & Wallgrün, 2019; Leppämäki, Toivonen, & Hiippala, 2024; Middleton, Kordopatis-Zilos, Papadopoulos, & Kompatsiaris, 2018). In addition to GIR-related applications like geo-enriched search engines (Acheson & Purves, 2021; Mai et al., 2020), geoparsing also is applied in extracting events (Dewandaru, Widyantoro, & Akbar, 2020), including hazard (Avvenuti, Cresci, Del Vigna, Fagni, & Tesconi, 2018; de Bruijn et al., 2019; Grace, 2021; Suwaileh, Elsayed, Imran, & Sajjad, 2022), disease (Gritta, 2019; Surano, Porfiri, & Rizzo, 2022), crime (Chopin & Caneppele, 2019; Haberman, Hatten, Carter, & Piza, 2021), and traffic (Idakwo, Adekanmbi, Soronnadi, & David, 2025; Mehta et al., 2023; Suat-Rojas, Gutierrez-Osorio, & Pedraza, 2022) events. Gazetteers can also be used in transforming geographic coordinates to an address or retrieving a list of nearby places. This task, which is known as reverse geocoding (McKenzie & Janowicz, 2015; Sabzali Yameqani & Alesheikh, 2025; Yameqani & Alesheikh, 2019), is applied in Location-Based Services (LBSs) (Deidda, Pala, & Vacca, 2013; Etemad et al., 2023), mapping and navigation tools (E. M. Delmelle, Marsh, Dony, & Delamater, 2021; Nallur, Elgammal, & Clarke, 2015), and discrete spatial analysis (D. Li, Cova, & Dennison, 2017; Qazi, Imran, & Ofli, 2020). To fulfill ever-increasing geospatial demands, especially in GIR and LBSs, digital gazetteers should be enriched.

Gazetteer enrichment strategies can be grouped into six categories: (1) place name harvesting and updating the gazetteer with new entries (Y. Hu et al., 2019); (2) developing a semantic structure for the gazetteer (Cardoso, Amanqui, Serique, dos Santos, & Moreira, 2016; Janowicz & Keßler, 2008; Machado, de Alencar, de Oliveira Campos, & Davis, 2011; Moura, Davis Jr, & Fonseca, 2017); (3) improving spatial dimension (Acheson, De Sabbata, & Purves, 2017; Adams, 2017; Jones, Purves, Clough, & Joho, 2008); (4) removing duplicates (Hastings & Hill, 2002; Martins, 2011); (5) making it temporal (Grossner, Janowicz, & Keßler, 2016; Manguinhas, Martins, Borbinha, & Siabato

Vaca, 2009), and (6) supporting multilingualism (Laurini, 2015). Our research study follows the two first strategies to enrich urban gazetteers.

For collecting place names, one approach is using the Volunteered Geographic Information (VGI) published on websites like OpenStreetMap and Wikimapia (De Longueville, Ostländer, & Keskitalo, 2010; Oliveira, Campelo, Baptista, & Bertolotto, 2016) or customized websites such as KiKojas[1] and YourPlace names[2] (Twaroch & Jones, 2010). Another approach is information exploitation from unstructured texts published on social media or the web (Gao, Li, Li, Janowicz, & Zhang, 2017; Lim, Nitta, Nakamura, & Babaguchi, 2019; Ying Zhang et al., 2019). The common aim of the two approaches is to enable digital gazetteers to play a desirable role in people-oriented applications such as disaster management (Cheng, Zhang, Su, Gao, & Shen, 2019; Y. Hu & Wang, 2020; Yang, Yu, Qin, Lu, & Yang, 2019; T. Zhang, Shen, Cheng, Su, & Zhang, 2021), tourism (Haris & Gan, 2017; Y. Hu et al., 2015), and decision-making processes (Huck, Whyatt, & Coulton, 2014; Keßler, Janowicz, & Bishr, 2009; Merschdorf & Blaschke, 2018). While the first approach has the edge in spatial accuracy, the latter has the advantage in thematic accuracy. Therefore, the two approaches complement each other (Bahrehdar, Adams, & Purves, 2020; X. Chen, Vo, Wang, & Wang, 2018; X. Hu et al., 2022). Regarding the second enrichment strategy, a gazetteer with a semantic structure has an ontology that makes it intractable with other geospatial resources (Liu et al., 2009). Moreover, it contains semantic relationships between place names to answer semantics queries (Keßler, Janowicz, et al., 2009). These semantic relationships can be divided into spatial (topological, metric, and directional) and aspatial (administrative hierarchy or town twinning).

Real estate advertising websites are informative for urban-level projects than other unstructured text resources like social media or travel blogs. In addition to property-related information, the contributors often refer to the locality, nearby facilities, and points of interest (POIs) visible from the property (Abbasi & Alesheikh, 2023; Blanchi, Fusco, Emsellem, & Cadorel, 2022; McKenzie, Liu, Hu, & Lee, 2018). Consequently, housing listings illuminate how citizens sense and experience the urban environment. Out of many urban entities, neighborhoods and main streets, due to their individual social, economic, and demographic characteristics, are essential for real estate market participants (E. C. Delmelle & Nilsson, 2021; Levy & Lee, 2011; Oto-Peralías, 2018; Samarin & Sharma, 2020; Talen & Jeong, 2019; Tewari & Beynon, 2018). Accordingly, property website users commonly mention them. However, the listing authors use these two urban place types without linguistic clues because of their well-knownness.

With more than one hundred million speakers and the official language of Iran, Persian is a branch of the Indo-Iranian group of Indo-European languages (Windfuhr, 2009). Although this language was the fifth most used on the Web by 2021 (W3Techs, 2021), a few works have been done in studying, creating tagged corpora (Shahshahani, Mohseni,

[1] https://www.kikojas.com

[2] https://www.yourplace names.com

Shakery, & Faili, 2018), and developing natural language processing tools (Asgari-Bidhendi, Janfada, Roshani Talab, & Minaei-Bidgoli, 2021; Poostchi, Borzeshi, Abdous, & Piccardi, 2016; Taher, Hoseini, & Shamsfard, 2020). A notable part of this issue refers to the properties of the language, such as free word order, having different forms of writing, and issues related to Arabic-like scripting (Ghayoomi & Momtazi, 2009; M. Shamsfard, 2011). Additionally, most researchers have only paid attention to formal Persian. Habib noted that recognizing nouns and pronouns, finding word boundaries, and Part Of Speech (POS) tagging and stemming are still unsolved (Habib, 2021). Although there is remarkable research on harvesting geospatial information from unstructured English texts (X. Hu et al., 2022; Y. Hu, 2018; Stock et al., 2022), there is very little research on Persian. In addition to the linguistic differences, different addressing and naming place mechanisms between Iran's study region and English-speaking countries make it necessary to review and develop the current methods appropriately.

To address the abovementioned issues, we propose a hybrid-driven framework to extract urban geospatial information from online real estate advertisement websites. Such information can be applied in gazetteer enrichment, mostly for regions with limited VGI. The contributions of this work are fourfold: (1) we extended a machine-learning method that was introduced by (McKenzie et al., 2018) to identify main street names in addition to neighborhood names from housing listings; (2) we developed a rule-based approach to extract urban place names from the listings; (3) we proposed a novel pattern-based method to infer spatial relationships between the identified place names; (4) our work focused on Persian texts despite the most of related studies which have addressed to English written texts.

2. Related Work

This section addresses related work in two subsections: Harvesting place names and extracting spatial relationships. Each subsection ends by highlighting the contribution of this work to the state of the art.

2.1. Harvesting Place Names

The resources utilized in the state-of-the-art studies can be divided into three groups: (1) social media platforms, (2) web search engines, and (3) real estate online listing websites.

In collecting place names from social media, some studies have focused on Twitter as a microblogging service (Gao, Janowicz, et al., 2017; Gao, Li, et al., 2017; Lim et al., 2019; Yang et al., 2019), while other studies have focused on picture-oriented social networks such as Panoramio (Popescu, Grefenstette, & Moë llic, 2008) and Flickr (Y. Hu et al., 2015; Keßler, Maué , Heuer, & Bartoschek, 2009; Rattenbury, Good, & Naaman, 2007). Like other microblogging social networks and news websites (Imani, Chandra, Ma, Khan, & Thuraisingham, 2017; Wang & Stewart, 2015), Twitter provides near real-time information suitable for extracting event-affected places (e.g., disaster management). On the other hand, picture-oriented social networks are ideal for harvesting touristic place names.

Utilizing web search engines is another way that attracted researchers' attention. Brindley et al. (2018) leveraged Bing API to extract neighborhood names from address texts (Brindley, Goulding, & Wilson, 2018). Their study focuses on the United Kingdom (UK). It uses the Ordnance Survey's Code-Point dataset to delineate neighborhoods interleaved in addresses between street and city names. Similarly, Zhang et al. (2019) proposed a framework for mining urban place names using the Google search engine from web pages containing addresses by focusing on American addressing styles. They used a machine learning-based classifier to filter out housing listings as noisy web pages according to their considered place types (Ying Zhang et al., 2019). Behind the scenes, what enables these methodologies is the standard addressing systems.

From the third group, McKenzie et al. (2018) presented a data-driven method to identify neighborhoods by defining a set of spatial statistics for n-grams extracted from the rental listing description and using random forest as the classifier model (McKenzie et al., 2018). Hu et al. (2019) suggested a framework to harvest place names using Named Entity Recognition (NER) tools and spatial clustering to remove wrong-labeled candidates (Y. Hu et al., 2019). These last two are similar to our work from data resource and methodology perspectives. However, there are differences in the language and the study regions between our work and those studies. While both are developed based on English-written housing listings of several cities in Canada and the U.S.A., we considered Persian and Iran's four most populated metropolises. These differences cover all language-related and urban development issues. In addition to neighborhoods, we have paid attention to gathering main streets. Concerning existing homonym sub-urban regions, we considered spatial clustering before identifying neighborhood and main street n-grams. Lastly, in addition to collecting urban place names from the listings, our proposed framework also extracts implicit spatial relationships between the place names.

2.2. Extracting spatial relationships

Spatial relationship extraction in the literature is studied for different applications. Generally, spatial relationships are defined between two things, whether non-geographic (Kordjamshidi, Van Otterlo, & Moens, 2011) or geographic entities. From the geographic point of view, some studies have used relative expressions to provide a better understanding of vague places (Yu Zhang, Wu, Wang, & Su, 2017) or to delineate the referred location with respect to the mentioned place name(s) (Cadorel, Blanchi, & Tettamanzi, 2021; Stock et al., 2022). Some studies modeled the meaning of spatial relational expressions using query tools (Derungs & Purves, 2016; Wallgrü n, Klippel, & Baldwin, 2014). Another group has focused on creating a place graph from place descriptions in which places would be nodes and the spatial relationships as edges (H. Chen, Vasardani, Winter, & Tomko, 2018; Kim, Vasardani, & Winter, 2015). Recent work by Zhang et al. (2023) has suggested a graph-based model for representing addresses that can be applied for address matching and toponym resolution (C. Zhang, He, Guo, & Ma, 2023).

Extracting spatial relationships from place descriptions is mainly realized by applying sentence-level natural language processing (NLP). In other words, the relationship between the locatum (also known as figure) and the relatum (also known as ground) is extracted according to their syntactic roles (J. Li, Liu, & Xiong, 2017; Stock & Yousaf, 2018). For example, in the sentence 'New York is the most populous city in the United States', the locatum is 'New York', which the description is written about it; the relatum is 'United States', a reference to 'New York'; and the spatial relationship is 'inside' concerning the preposition term 'in' before the relatum. On the other hand, the relationships can be extracted from descriptive addresses at the phrase level. Although applying a regular parsing method is more straightforward for standard addresses, descriptive addresses enable extracting more relationships, yet it needs further processes (Cruz, Vanneschi, Painho, & Rita, 2021; Javidaneh, Karimipour, & Alinaghi, 2020). While a notable number of studies have paid attention to inherent spatial relationships in addresses, most are limited to improving address-based geocoding systems (L. Li, Wang, He, & Zhang, 2018; Tian et al., 2016). Our work is based on descriptive addressing frequently used in online property listings. The framework combines place description and addressing aspects for gathering spatial relationships between the identified urban place names.

3. Materials and Methods

The methodology follows a structured three-tiered approach to processing geotagged housing listings, as illustrated in Figure 1. The duplicates and those outside the city's administrative boundaries are discarded. The first tier involves extracting neighborhood and main street names, often used without linguistic clues due to their commonness. Additionally, this tier enables us to distinguish between homonym place names in the second tier by assigning the listings to the most related neighborhood or main street place names by respecting the convex hulls. The first tier commences with the extraction of n-grams from the saved advertisements. After cleaning and standardizing the n-gram set, spatial clustering is applied, considering that each spatial n-gram can refer to multiple regions of the city. The defined spatial predictors are computed for each not-clustered n-gram or split n-gram from its generic cluster. Subsequently, a random forest model identifies the neighborhood and the main street n-grams. In the second tier, a rule-based approach extracts all remaining urban place names written in listings. Finally, in the third tier, spatial relationships between these place names are extracted based on defined linguistic patterns.

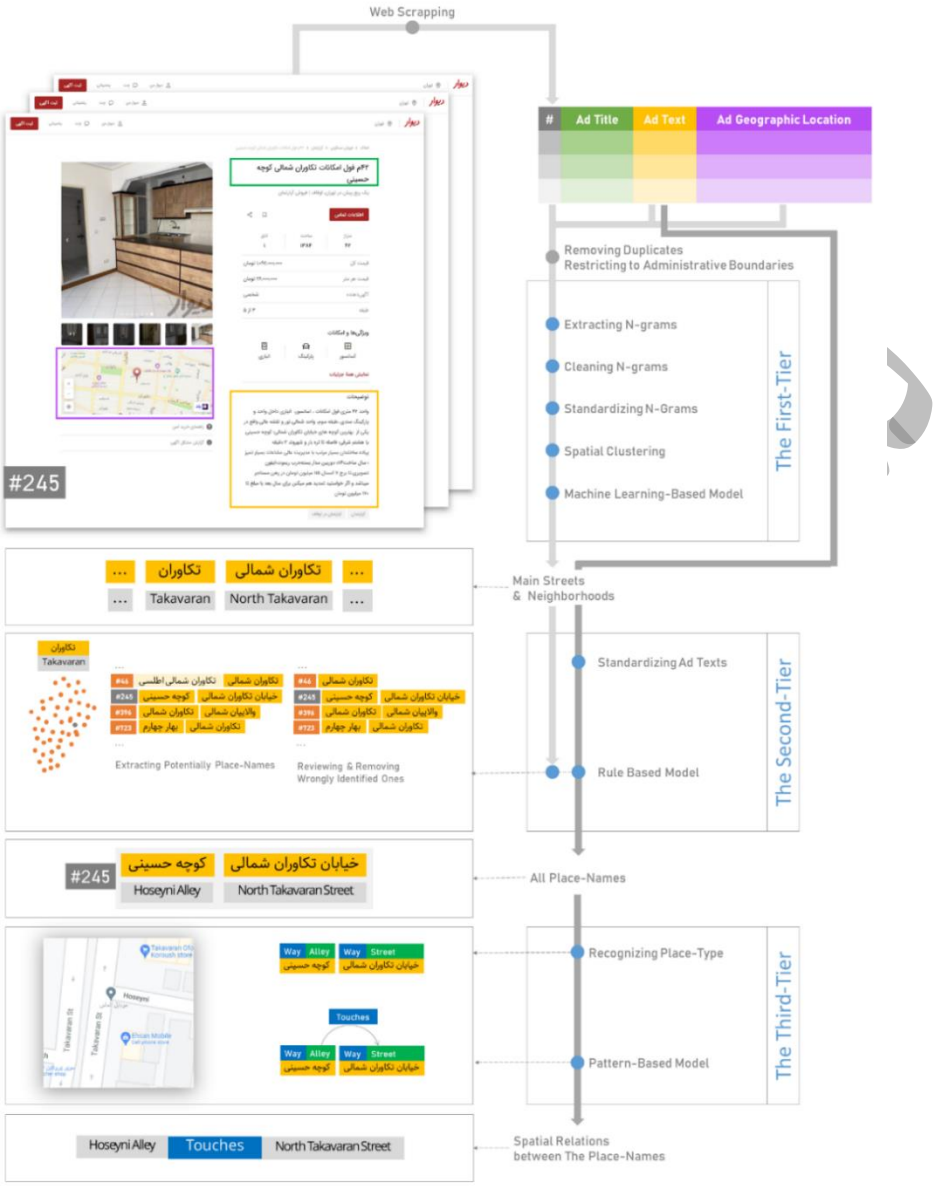

Figure 1 : The flowchart of the developed framework.

3.1. Study area and data used

We chose four Iranian metropolises, namely Tehran, Mashhad, Isfahan, and Shiraz, as our study areas. Tehran, the capital, is the most populous metropolis, with over 9 million people. Mashhad, with more than 3 million, Isfahan, with almost 3 million, and Shiraz, with nearly 2 million, are in the following ranks. We selected these cities due to their differences in area size, urban development patterns, and addressing structures. We chose Divar, a housing advertisement website in Iran, for its extensive geographical coverage and high rate of posts per day. We employed web

scraping for one month (January 2020) and retained only geotagged posts, including title, text, and geographical coordinates. We limited the collected posts to each metropolis's administrative boundaries and removed duplicate entries in textual and geographic contexts. Figure 2 depicts the geographic distributions of housing listings for each metropolis.

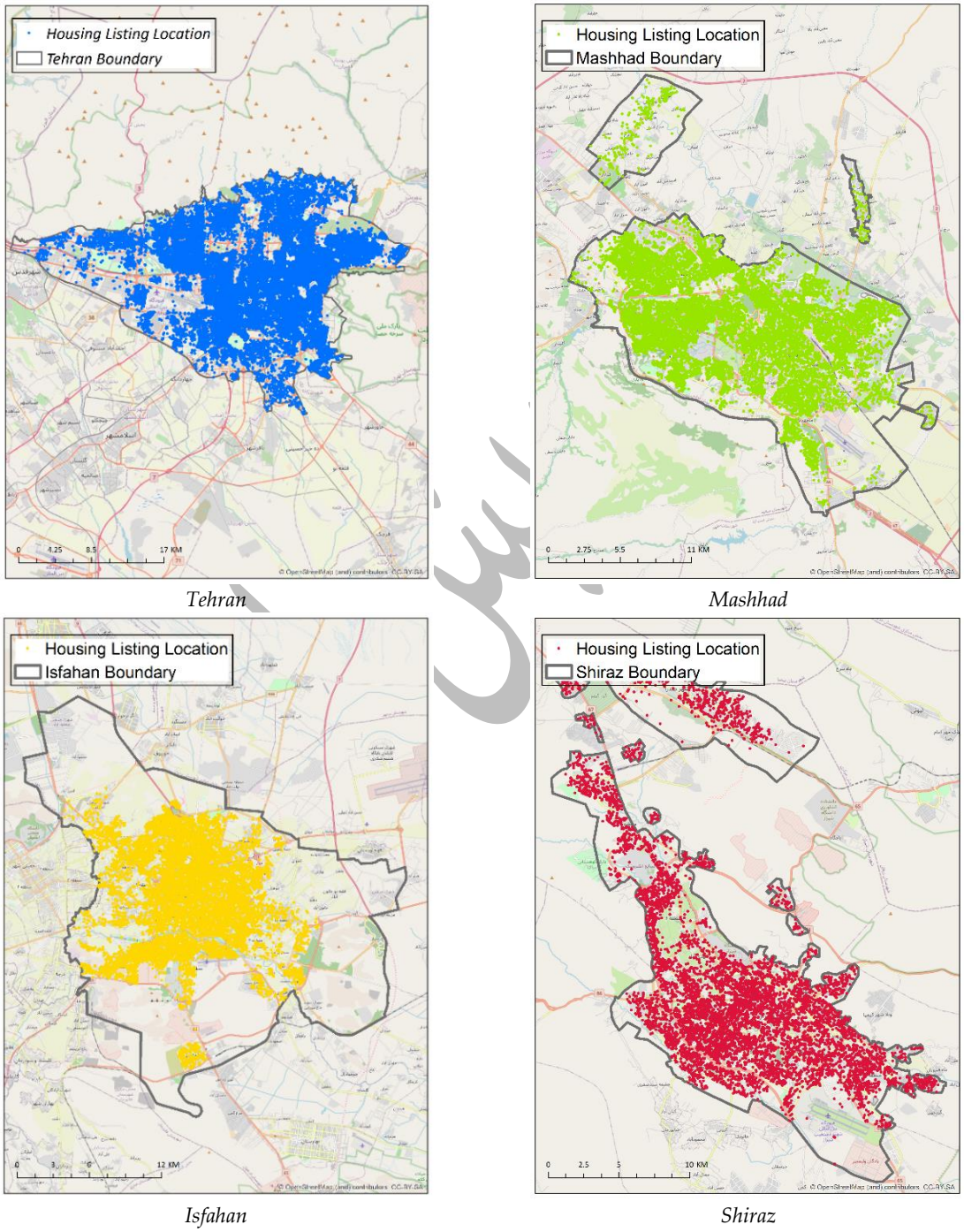

Figure 2: The geographic distributions of housing listings in the four metropolises

3.2. First tier

The first tier of the process employs a machine learning technique to harvest neighborhoods and main streets, which are often written without linguistic clues and may be overlooked by named entity recognition (NER) tools.

3.2.1. N-gram extraction

Before extracting n-grams, a normalization process was applied. This process involved removing non-textual characters, converting one or two-digit numbers to their alphabetical style, omitting three or more digit numbers, converting characters to their Persian equivalent, and changing abbreviated forms to their complete forms. We extracted n-grams (with lengths ranging from 1 to 5) from each advertisement's textual content, including the title and the description text. Figure 3 shows the extraction of n-grams for a text written in English. For this step and the other steps that involve NLP tasks, we utilized Hazm[3], a Python package developed by Roshan, and the dependent package, NLTK[4].

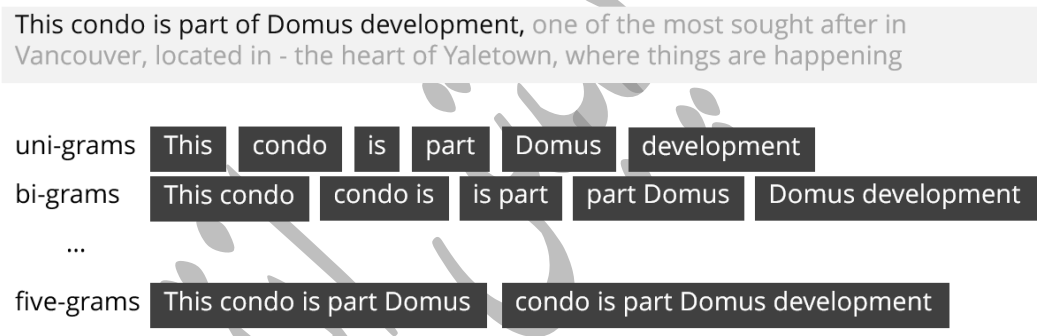

Figure 3: A sample housing listing from Craigslist (a housing advertisement website) and its extracted n-grams

3.2.2. Cleaning and standardizing

We performed statistical and non-statistical cleaning of the n-gram datasets. In statistical cleaning, we removed n-grams with frequency values (i.e., the number of geotagged listings containing the n-gram) higher than three deviations above the mean, which included the most frequently used words, such as grammatical particles in listings. Based on the histogram, we chose 30 as the threshold to retain as many spatial n-grams as possible and enhance the models' robustness. In non-statistical cleaning, we removed n-grams with one or two characters, those containing only numbers, and common real estate listing keywords. In standardizing, we attempted to convert phrases to their commonly used counterparts.

[3] https://www.roshan-ai.ir/hazm/

[4] https://www.nltk.org/

3.2.3. Spatial clustering

Considering that two different places in a city can have the same name, the spatial clustering of the point set of each n-gram is necessary. For instance, several places (either neighborhoods or main streets) located in Tehran share the name 'گلستان' (Golestān) (Figure 4). As a uni-gram, this name has a point set with a distribution like other non-spatial n-grams. Spatial clustering (Aldstadt, 2010; Madhulatha, 2012) allows such n-grams to save their included clusters. After spatial clustering using DBSCAN (Schubert, Sander, Ester, Kriegel, & Xu, 2017), we considered clusters with a size greater than 30 as a new n-gram record and removed others. Additionally, due to clustering, outlier points from the point set of each n-gram were deleted. To this end, we created an instance of DBSCAN object defined in the scikit-learn Python package[5] with Euclidean method and Ball tree algorithm to compute distances and find nearest neighbors.

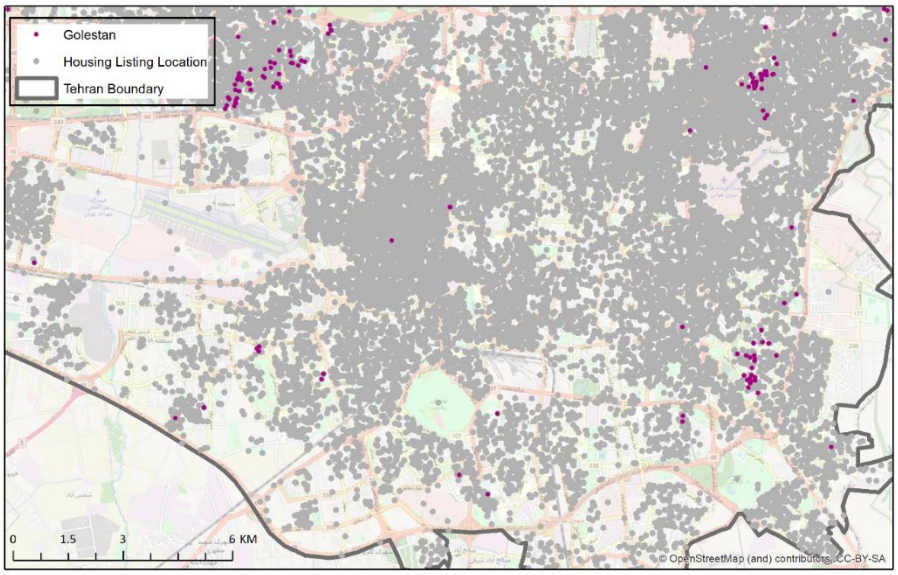

Figure 4: Golestan (گلستان), one of the n-grams which include two or more spatial clusters and several outliers

3.2.4. Labeling

During this step, n-grams were manually checked based on their linguistic aspects. We identified the place type for each n-gram based on the spatial distribution of the points on the OpenStreetMap base map. We labeled the n-gram in three classes: neighborhood, main street, and non-spatial (including non-spatial n-grams such as 'this condo' and any spatial n-grams other than the two first classes like intersections and landmarks). We encountered two challenges in labeling n-grams: (1) distinguishing between a neighborhood and a main street and (2) multipart place names.

[5] https://scikit-learn.org/

A significant portion of n-grams that are main streets can also be referred to as neighborhoods. Some are initially neighborhoods, but the name later becomes assigned to the main street that crosses through the neighborhood. Conversely, the neighborhood formed around a main street takes the name used for the street. We labeled most of them as street (commonly with a linear distribution and high density in the ridge) unless those had consistent distributions like ordinary neighborhood n-grams.

A multipart place name generates a full n-gram and several partial n-grams. Sometimes, these partial n-grams have the same point set as the full n-gram. The most common case involves neighborhoods that end with the suffix 'اباد' (-ābād). Partial n-grams such as 'سعادت' (Sa'ādat) and 'مهر' (Mehr) have independent meanings. However, for the full n-grams 'سعادت اباد' (Sa'ādat ābād) and 'مهراباد' (Mehr ābād), they are not used alone. Such partial n-grams can be easily filtered out, but in some cases, a few partial n-grams can be more frequent than the full n-gram. That often occurs for places named after famous personalities that the vernacular people prefer to use the simplest form of their names. For example, the place name 'حکیم شفایی' (Hakim Shafāei[6]) also can be used as 'شفایی' (Shafāei) by people. Depending on the case, such partial n-grams were either kept or removed.

3.2.5. Spatial predictors

The spatial predictors used in the study include 11 measures for spatial dispersion, 12 measures for spatial homogeneity, and three other measures, namely convex-hull area, convex-hull density, and the width-to-length ratio of the rotated minimum bounding rectangle. In the following, we explain more about these predictors.

From the spatial dispersion group, compared to (McKenzie et al., 2018), we already used standard distance, average pairwise distance, average distance between points, and their first nearest neighbors (NN1). The average 5th and 10th nearest neighbor distances were used as the substitutes for NN2 and NN3, respecting the low difference between them and NN1 in our data. We also defined the peak-to-average ratio, which represents the ratio of the maximum NN1 to mean NN1. Like (McKenzie et al., 2018), we used Ripley's L function to determine the spatial homogeneity of each n-gram's point set. The function's values were calculated for ten 100-m distances starting from 50-m and ending at 950-m distance. Additionally, the kurtosis and the skewness for each Ripley's function were computed as the measures representing the overall status of the function. Besides the convex hull measures, we also paid attention to the rotated minimum bounding rectangle. We calculated the ratio of the width to the length, hypothesize that the main

[6] A great poet and physician of Safavid era

street n-grams have higher values for that measure compared to other n-grams. The complete list of the spatial predictors is presented in Table A.1, Appendix A. We employed SciPy[7], Shapely[8], and Astropy[9] to compute the predictors.

3.2.6. Modeling

The study utilized the random forest, an ensemble learning classification method for modeling, based on the findings of (McKenzie et al., 2018). Random forest is a supervised machine learning method. It can solve classification and regression problems by generating a plurality of decision trees (Biau & Scornet, 2016; Breiman, 2001). Due to Its simplicity, few parameters to tune, high efficiency, and accuracy, this method is popular in various tasks, including geospatial issues (Ayala-Izurieta et al., 2017; Chang et al., 2020). The data suffer from imbalancement, as shown in Table 1. To some extent, training the model with imbalanced data can be resolved by oversampling the positive classes. Oversampling was performed for each modeling process on train data to equalize the participation of neighborhood, main street, and non-spatial records. The scikit-learn python package was used to handle data imbalancement and generate random forest models.

Table 1 : Number of housing listings (After preliminary data cleaning); Number of n-grams after extracting, cleaning, standardizing, spatial clustering, and labeling (removing partial n-grams); and the number of labeled n-grams in each class.

#	Tehran	Mashhad	Isfahan	Shiraz
Raw Listings	35451	15894	12590	10474
Extracting	1820492	926717	748812	620888
Cleaning	1273	603	542	443
Standardizing	1258	596	534	442
Clustering	654	290	294	213
Labeling	373	169	164	125
Main Street	35	25	33	37
Neighborhood	155	54	55	46
Non-Spatial	183	90	76	42

Figure 5 illustrates the feature importance results obtained using permutation-based feature importance with confidence intervals. The evaluation was conducted over 50 iterations, and the reported values represent the average across

[7] https://scipy.org/

[8] https://github.com/shapely/shapely

[9] https://www.astropy.org/

five-fold cross-validation. While the plots suggest that certain features may introduce noise and could be candidates for removal to improve model performance in identifying neighborhoods and major streets within the same city, further analysis revealed that excluding these features leads to a noticeable drop in performance when the model is tested on data from other cities. This issue is particularly evident for features that exhibit interdependence due to their inherent characteristics.

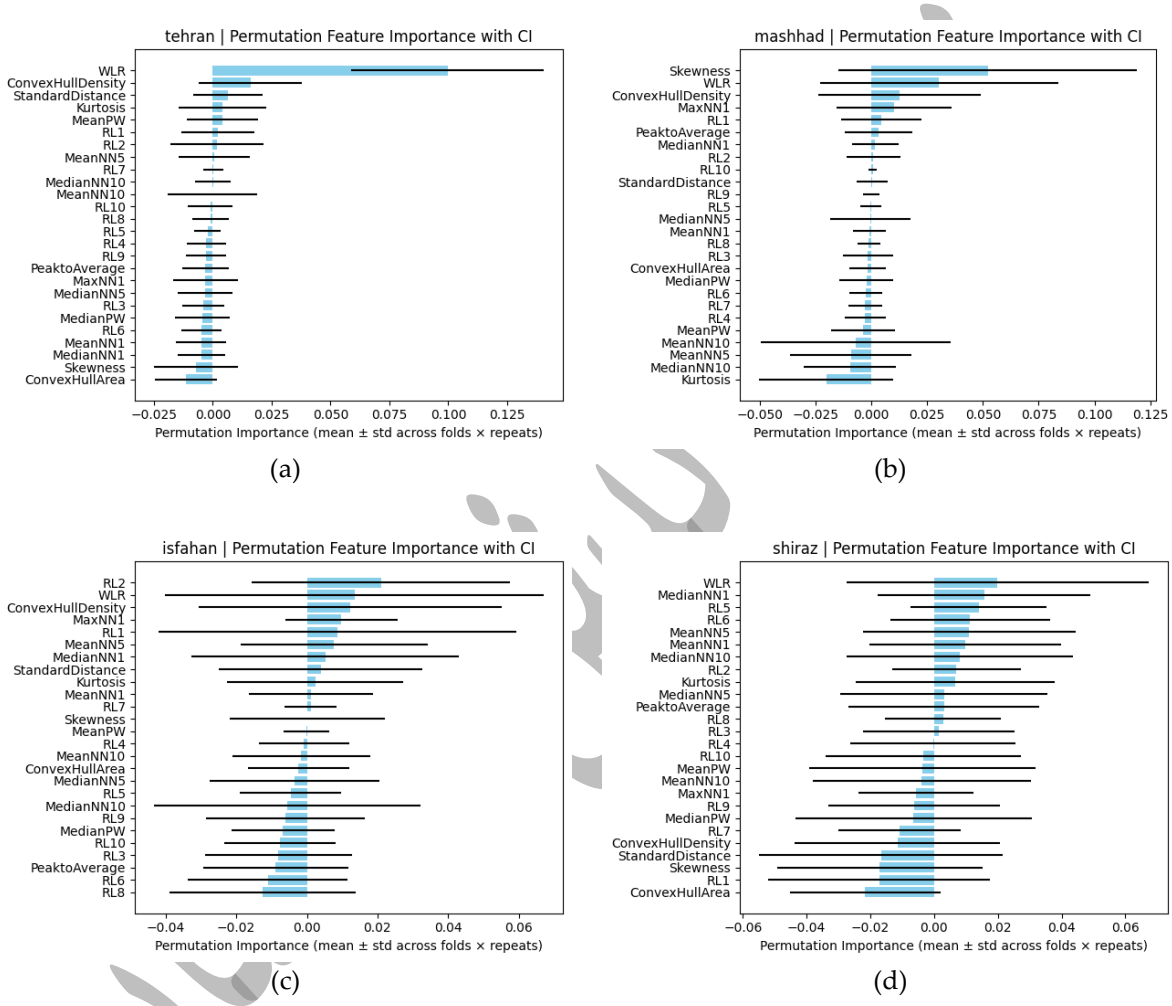

Figure 5: Feature importance charts of the model of all four metropolises: (a) Tehran, (b) Mashhad, (c) Isfahan, and (d) Shiraz.

3.3. Second tier

The second tier's main objective is to extract all urban place names from the advertisements. We developed a rule-based model to accomplish this goal for two reasons. Firstly, the Persian NER tools that are currently available do not effectively identify place names in informal texts like housing advertisements. Secondly, creating a suitable corpus

to generate a machine-learning model is time-consuming. Although some parts of the developed model are similar to geoparsing tools, it also includes other parts to enhance precision.

In order to develop the rule-based model, we utilized 4000 records from the Isfahan dataset and 100 records from the other metropolises' datasets due to enhance the generalizability of the model. From the new dataset, we kept the records that referred to places within the metropolis and located within the union of the convex hulls of the place names identified in the first tier. Additionally, we removed the records containing multiple points of the metropolis. The developed rule-based model was tested separately on 4000 records from Tehran, Mashhad, and Shiraz datasets.

3.3.1. Rule-based model

The rule-based model developed for extracting place names from housing advertisement texts includes two steps: (1) Identifying potential place names from each text and (2) Reviewing identified place names and deleting wrongly identified ones. To simplify, we will refer to the main street/neighborhood place name as SN and the housing advertisement records as AD. Each SN collected in the first tier takes the set of ADs geographically located in the convex hull of the SN. After editing, tokenizing, and POS tagging, potential place names are extracted from each AD text. Once all the ADs assigned to the SN are extracted, the model removes the wrongly identified place names based on other rules (respecting the frequency and length of the extracted ones).

3.3.2. Clue phrases and stop phrases

In addition to place suffixes, we used three groups of clue phrases: (1) phrases in which place names come after, (2) place type phrases, and (3) honorifics. These are the starters of the token series when extracting potential place names from AD, giving them a chance to be identified as a place name. The expansion continues until a stop phrase appears in the token series. We categorized these stop phrases into several groups, such as the following: (1) building elements, facilities, and home furniture; (2) phrases used for expressing the property's geometry and other attributes; (3) transaction-related phrases; and (4) generic stop words.

3.3.3. Advertisement text editing

It is necessary to edit and standardize text to extract the maximum number of place names, respecting the nature of texts. We accomplished this phase in four ways: (1) changing Arabic and English characters to their related Persian characters; (2) changing abbreviated forms to their full forms (e.g., 'خ' to 'خیابان' similar to changing 'St' to 'Street'); (3) resolving issues related to the space character (the improper space between a word and adherent phrases); and (4) correcting badly typed words (e.g., 'شهزک' to 'شهرک' analogous to correcting 'Tpwn' to 'Town').

3.3.4. Place name identification (Rules)

If the n-gram (generated based on the current token) begins with a clue word/phrase, ends with a place name suffix, or contains a formerly identified NS place name, the defined conditions will be checked to decide whether to expand the n-gram or to stop the expansion of the n-gram. Otherwise, the next token will be examined. The unaffected n-grams will be added to the AD's temporary potential place names list. After surveying all tokens, collected entries are standardized.

After surveying all ADs assigned to the NS, each AD's extracted potential place names are reviewed. Suppose a place name (A) would be a part of another place name (B). In that case, different actions can be taken depending on which of the following conditions is true: (1) If the rest of B is either a number, a place type phrase, or a directional name, then A should be removed; (2) If B is as frequent as A, then A as a partial place name should be removed (e.g., 'خیابان نیرو' and 'خیابان نیرو هوایی', Khiyābān-e Niroo Havāyi); and (3) If A is more frequent than B (e.g., 'خواجه ربیع' and 'خواجه ربیع بهمن', Khāje Rabi Bahman), then B should be removed. This final step is analogous to the named entity boundary detection (Ahmadi & Moradi, 2015; Bokaei, Nouri, & Sepahvand, 2021; M. M. Shamsfard, Puneh-Sadat, 2009). However, in our case, we first extract potential location entities (place names) and then determine which one has the correct boundaries.

3.4. Third tier

In the third tier, we aimed to extract spatial relationships between extracted place names from each AD based on a set of linguistic patterns. Descriptive addressing enables us to extract these relationships by respecting the place names' order and the specific words that link them together. We also considered place types in the defined patterns. The process was applied to the place names without any interruptions between them, regardless of some specific phrases. The extractable spatial relationships from the defined patterns include the following: (1) topological relationships (including inside, touches, intersects), (2) directional relations (in front of), and (3) distance relationships (nearby). Each pattern includes 2 or 3 place name participants. Depending on the number of participants and the pattern itself, a minimum of one spatial relationship and a maximum of four spatial relationships can be extracted.

3.4.1. Place type recognition

We applied a rule-based linguistic approach to categorize the place names into four classes of urban place types. These classes were defined based on (Lynch, 1960) and include: (1) Ways (defined as paths in (Lynch, 1960)) including alleys, streets, boulevards, and highways; (2) Intersections (analogous to nodes) such as squares and junctions; (3) Residential areas (referred as districts) like neighborhoods and residential complexes; and (4) Landmarks such as marketplaces and hospitals.

3.4.2. Spatial relationships

Error! Reference source not found. presents the spatial relationships that can be extracted from housing listings. To simplify, we used W, R, I, and L for Ways, Residential areas, Intersections, and Landmarks. The three relationships 'inside', 'touch', and 'cross' are the only extractable ones from topological relationships. From directional relationships, since relative directional relationships are mainly used in the listings, we only considered this group (only including *front of* concerning our data) and excluded absolute ones (such as 'north of' and 'east of'). We considered two ways to extract a 'nearby' relationship. The first is when there are no logical topological or directional relationships between the two associated place names. For example, assume that one of the place names is a landmark and the other is an intersection. Mentioning these together means that the two places are close to each other but, topologically, are disjoint from each other, and directionally, there is no clue to extract a relationship. The latter is based on the existence of a clue word like 'near to' in the text.

Table 2: The spatial relationships and extracted statements for each relationship depending on the place type of participants.

Spatial Relationship Group	Spatial Relation	Extracted Statement
Topological	Inside	W/R/I/L is inside R
		I is inside W
		L is inside L
	Touch	W touches W
		R/L touches W
		L touches L
	Cross	W crosses W
		W crosses W in I
Directional	In front of	W/L is in front of L
		W/L is in front of W
Distance	Nearby	W/I/L is nearby I
		W/L is nearby L

3.4.3. Patterns

To represent the patterns formally, we used ISO-Space (Pustejovsky, 2017). In general, the common patterns in extracting spatial relationships are as follows:

1. [PLACE 1] [PLACE 2]
2. [PLACE 1] [SPATIAL_SIGNAL 1] [PLACE 2]
3. [PLACE 1] [SPATIAL_SIGNAL 1] [PLACE 2] [PLACE 3]
4. [PLACE 1] [SPATIAL_SIGNAL 1] [PLACE 2] [SPATIAL_SIGNAL 2] [PLACE 3]

ISO-Space defined PATH tag as 'to capture locations where the focus is on the potential for traversal or functions as a boundary' whether for nominals such as '[railroad] between Boston and New York' or proper names like 'Pacific Coast Highway'. However, since the path references in our case are only urban ways (proper names), we annotated them with a PLACE tag to simplify.

Spatial signals might (1) refer to a part of the next place element, for instance, 'the beginning of', (2) implicitly/explicitly mean in the direction of the next place element like 'a few steps to', (3) be used to refer a region between the place elements (in company with another spatial signal), or (4) refer to a place related to the next place element such as 'nearby'.

Based on the PLACE and SPATIAL_SIGNAL entities, the extracted spatial relationships would be annotated as follows:

1. QSLINK (id, relType = IN, figure,ground)
2. OLINK (id, relType = FRONT, figure,ground)
3. MLINK (id, relType = Distance, figure,ground, val = NEAR)

That QSLINK, OLINK, and MLINK are used for qualitative spatial links (topological), orientation links (directional), and measure links (including distance, width, height, and length). To shorten it, we only mentioned 'IN' from topological relationships. Regarding the limitations of the Region Connection Calculus (RCC) as the base of ISO-Space for topological relationships, we followed the calculus-based method proposed by Clementini et al. to define topological relationships in a formal way (Clementini, Di Felice, & Van Oosterom, 1993) and used it to specify the relType.

3.5. Evaluation

We used recall, precision, and f-score measures to evaluate all three tiers. Recall expresses the participation of the positive records, which the model predicts, among other positive records. Conversely, precision indicates from the records predicted as positive how many of them are positive in reality. F-score, the harmonic mean of those two measures, also was used for evaluation (Equations 1, 2, and 3).

$$\text{Precision} = \frac{T\_P}{(T\_P + F\_P)} \tag{1}$$

$$\text{Recall} = \frac{T\_P}{(T\_P + F\_N)} \tag{2}$$

$$\text{F}_{\text{score}} = \frac{2 \cdot (Precision \cdot Recall)}{(Precision + Recall)} \tag{3}$$

That *T_P*, *F_P*, and *F_N* represent the number of true positives, false positives, and false negatives, respectively.

For the first tier, we used a multi-classification manner that needed an averaging task to get the measures' final values. While macro average calculates the measure by summing the measure's value for each class without considering their proportion, weighted average gives the final value by considering their proportion. According to the imbalancement of the data, we used the macro average. Thus, first recall and precision measures were computed for each class, followed by the f-score. For each metropolis, the model was evaluated using 5-fold cross-validation performed exclusively on the data from that metropolis. The final performance metric was calculated as the average across the five folds. To assess the model's generalizability to other cities, the model trained using 70% of the data from one metropolis (following a 70-30 split) was applied to datasets from other metropolises.

For each AD entry in the subdatasets of the three metropolises, we labeled the place names by marking their start and end character indices. In addition to our rule-based model, we employed two ParsBERT-based NER models (Farahani, Gharachorloo, Farahani, & Manthouri, 2021) developed by Hooshvare Research Lab[10]—each fine-tuned separately on the PEYMA (Shahshahani et al., 2018) and ARMAN corpora (Poostchi, Borzeshi, & Piccardi, 2018)—to extract urban place names from the advertisement texts. This allowed for a more robust evaluation of the rule-based model's performance by comparing it against transformer-based baselines trained on standard Persian NER datasets.

In the third tier, we identified the spatial relationships that occurred in the real world to evaluate the framework's performance in spatial relationship extraction. Table 3 depicts the participation ratio of each pair place type in the labeled spatial relationships for Tehran, Mashhad, and Shiraz, respectively. For instance, 'touches' for 'LW' represents the landmark place name (L) touches the way place name (W).

Table 3: Labeled spatial relationships for each pair place type

		WW	WL	WR	WI	LW	LL	LR	LI	RW	RR	IW	IL	IR	II
Tehran	Inside	0	0	110	0	0	3	13	0	0	18	29	0	23	0
	Touches	364	0	0	0	48	0	0	0	11	0	0	0	0	0
	Crosses	10	0	0	0	0	0	0	0	0	0	0	0	0	0
	Nearby	0	6	0	40	0	3	0	8	0	0	1	1	0	2
	In front of	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Mashhad	Inside	0	0	77	0	0	1	7	0	0	4	17	0	8	0
	Touches	335	0	0	0	57	2	0	0	15	0	0	0	0	0
	Crosses	11	0	0	0	0	0	0	0	0	0	0	0	0	0
	Nearby	0	2	0	27	0	1	0	11	0	0	0	0	0	0

[10] https://huggingface.co/HooshvareLab/models

	In front of	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Shiraz	Inside	0	0	175	0	0	0	27	0	0	42	23	0	2	0
	Touches	354	0	0	0	64	0	0	0	26	0	0	0	0	0
	Crosses	3	0	0	0	0	0	0	0	0	0	0	0	0	0
	Nearby	0	3	0	24	0	0	0	18	0	0	0	0	0	0
	In front of	0	5	0	0	0	3	0	0	0	0	0	0	0	0

4. Results and Discussion

This section presents the result of implementing the framework in the three tasks.

4.1. Identifying main street and neighborhood place names

The results of the first tier in identifying main street and neighborhood n-grams are shown in Figure 6, Tables 4, 5, and 6. Figure 6 displays the trained model of each metropolis how identified the main streets, neighborhoods, and non-spatial n-grams of other metropolises. The rows represent the metropolises considered for training, and the columns represent those considered for testing. The prediction performance evaluated by recall, precision, and f-score are presented in Tables 4, 5, and 6.

Table 4: Recall values for the four random forest models trained on each metropolis's dataset and tested on the same dataset (the diagonal cells) and tested on the other datasets (the off-diagonal cells). Each value is the macro average of the recall values computed for all three classes.

		Testing			
		Tehran	**Mashhad**	**Isfahan**	**Shiraz**
Training	**Tehran**	0.80	0.70	0.74	0.72
	Mashhad	0.72	0.77	0.73	0.69
	Isfahan	0.81	0.67	0.71	0.75
	Shiraz	0.81	0.64	0.76	0.73

Table 5: Precision values for the four random forest models trained on each metropolis's dataset and tested on the same dataset (the diagonal cells) and tested on the other datasets (the off-diagonal cells). Each value is the macro average of the precision values computed for all three classes.

		Testing			
		Tehran	**Mashhad**	**Isfahan**	**Shiraz**
Training	**Tehran**	0.79	0.73	0.78	0.73
	Mashhad	0.70	0.86	0.73	0.68
	Isfahan	0.76	0.67	0.74	0.75
	Shiraz	0.76	0.64	0.76	0.70

Table 6: F-score values for the four random forest models trained on each metropolis's dataset and tested on the same dataset (the diagonal cells) and tested on the other datasets (the off-diagonal cells). Each value is the macro average of the f-score values computed for all three classes

		Testing			
		Tehran	Mashhad	Isfahan	Shiraz
Training	Tehran	0.79	0.70	0.75	0.69
	Mashhad	0.70	0.80	0.73	0.68
	Isfahan	0.77	0.67	0.71	0.75
	Shiraz	0.77	0.64	0.76	0.71

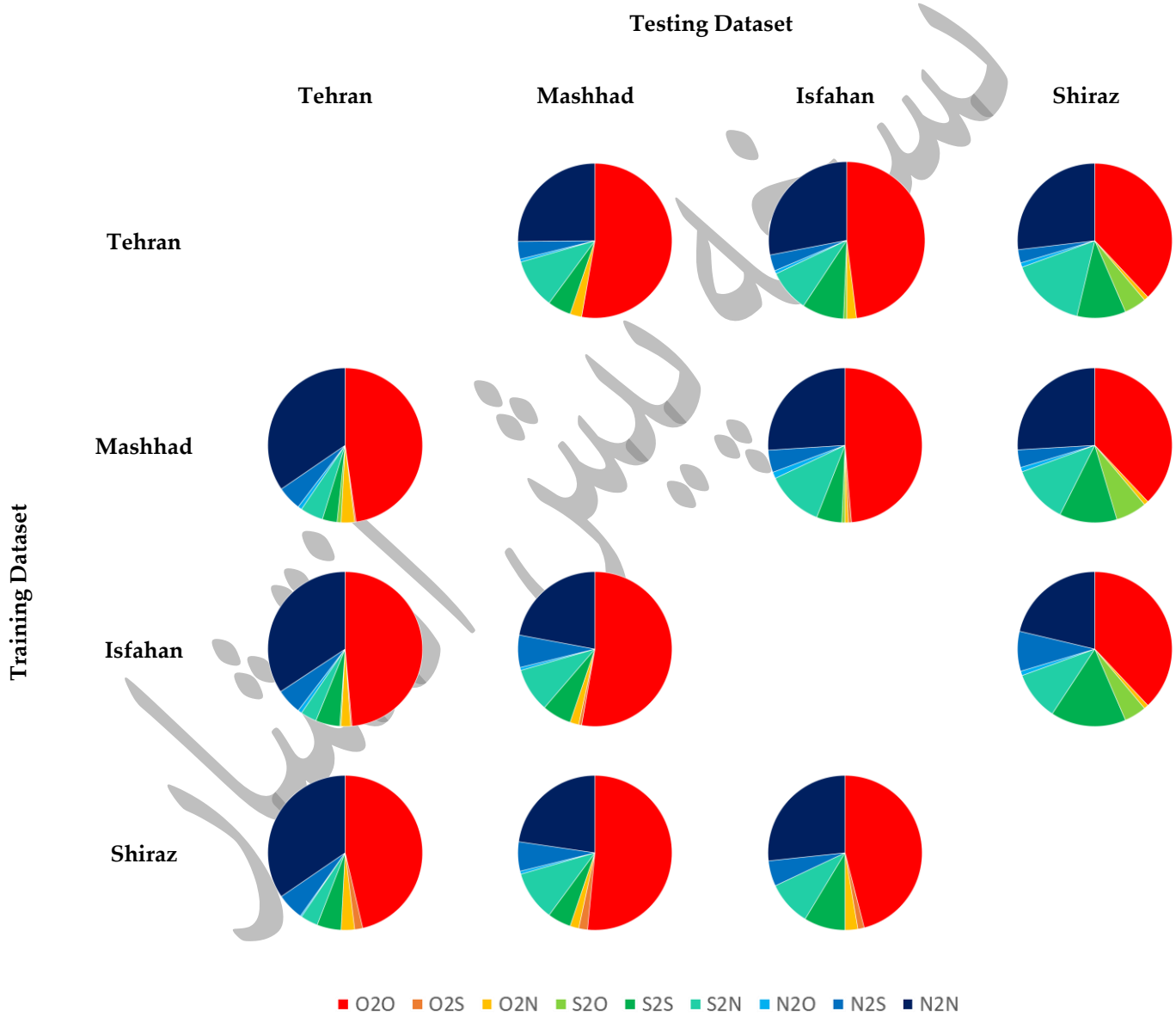

Figure 6: Pie charts of labels and predictions of each metropolis' dataset considered for testing. O represents n-grams, neither are neighborhoods nor main streets; N represents neighborhood n-grams, and S represents main street n-grams. For example, O2S indicates n-grams labeled as non-spatial but predicted as main street.

4.1.1. Changes and improvements

As a general occurrence, consideration of the same metropolis for training and testing results in the highest number of true predictions compared to when the train and test datasets are selected from different metropolises. However, regardless of the issues expressed in Section 4.1.1, the model trained on a metropolis's dataset and tested on others provides a good prediction. The study provides some new findings compared to (McKenzie et al., 2018). Firstly, based on a set of spatial measures and conducting a machine learning-based method, both main street and neighborhood n-grams can be identified. Secondly, the results show that such a data-driven method is successful for Persian-speaking cities. Thirdly, while the base study chose nearly similar cities except for Montreal for its bilinguality, four metropolises at different population levels and various urban development patterns were considered in this study. As the results indicate, such differences decreased the performance in identifying a neighborhood as a main street or vice versa. We addressed this issue in the following subsection. Lastly, prediction measures show improvement in identifying neighborhoods of other cities.

4.1.2. Sources of misclassifications

An important issue is that the difference between main street and neighborhood n-grams from their spatial perspective is sometimes unclear. Most neighborhoods are structured along main streets, whether they are homonyms or not, and a few numbers of them have a structure like a town. According to Figure 6, most false predictions are labeled as a main street but predicted as a neighborhood and vice versa. Assume one dataset includes various neighborhood n-grams, which notably overlap main street n-grams. When testing on other datasets, the model trained on that dataset predicts the main streets as a neighborhood. As shown in Figure 6, this has happened to the Isfahan dataset considered for testing Tehran's random forest model.

Some reasons can be expressed about the records labeled as non-spatial but predicted as a neighborhood or main street. One reason is the side effect of spatial clustering. We used spatial clustering to remove outlier points from each n-gram's point set and consider valid clusters as the new n-gram point sets. However, some non-spatial n-grams have been clustered due to spatial clustering and took a spatial distribution similar to neighborhood or main street n-grams. Besides, real estate agencies usually use constant phrases in their housing listings, whether about the property or their company. Although most common real estate phrases are removed in data cleaning, some phrases remain that the n-grams extracted from them are associated with main street or neighborhood names and partly take their point patterns.

4.2. Extracting all place names

As presented in Table 7, the rule-based model has successfully extracted place names. In addition to streets and residential areas frequently used in housing listings, many other place types can be extracted from these texts, such as markets, schools, hospitals, and recreation centers.

Table 7: Recall, Precision, and F-score for models of each metropolis in extracting all place names

Metropolis	Model	Recall	Precision	F-Score
Tehran	ParsBERT (PEYMA)	0.28	0.38	0.32
	ParsBERT (ARMAN)	0.38	0.53	0.44
	Our Rule-Based Model	0.73	0.75	0.74
Mashhad	ParsBERT (PEYMA)	0.20	0.32	0.25
	ParsBERT (ARMAN)	0.19	0.38	0.25
	Our Rule-Based Model	0.68	0.67	0.68
Shiraz	ParsBERT (PEYMA)	0.29	0.37	0.32
	ParsBERT (ARMAN)	0.42	0.54	0.47
	Our Rule-Based Model	0.76	0.78	0.77

4.2.1. Linguistic ambiguities

Housing listings, like other texts, suffer from ambiguity. One typical example is the different usage of the conjunction 'and' ('و', va in Persian). Authors usually write the common place type to shorten a list of place names; then, the place names would be written. For example, one might write 'مترو صادقیه و طرشت' (Metro-e Sādeghiye va Tarasht). In this case, two place names are subways. However, the second-place name's type sometimes differs from the first one (Tarasht may refer to a neighborhood). In another usage, 'and' only connects two parts of a place name, for instance, 'خیابان برنامه و بودجه' (Khiyābān-e Barnāme va Boodje, Barnameh va Boudjeh Street) is only one place name. The ambiguity is also valid for two nouns with no stop word between them, and the first one starts with a clue word. The second noun can refer to the property owner, the real estate agency, another place, or an alternative name for the first one. Such ambiguities cause partially correct place names to be extracted, which increases false positives and false negatives. Sometimes, the place type is more important than the place name itself for the people. Allegorically, it is enough for the author to only write 'near to metro station' in the description. However, that also happens to some place names unique to each city (or a big part of a city). For example, rather than 'ترمینال کاوه' (Terminā̄l-e Kāve, Kaveh (bus) terminal), one might use 'ترمینال' (the terminal) in the listing. This way of referring places is usually used for real estate or booking websites. However, it can be considered a cul-de-sac for extracting urban place names.

4.2.2. Strengths of the rule-based model

The results for the second tier are promising. However, there is still room for improvement in inferring maximum possible spatial relationships. In the following, we explain more about the advantages of using a rule-based model for the second tier.

The second tier of the framework requires a model that can identify various types of urban place names and accurately detect entity boundaries. Naming places after people or organizations makes this task even more challenging, especially with increasing spatial scale. Most existing NER tools are trained with texts chiefly at the super-urban level (especially for Persian) and recognize place names by 'Location', 'GPE' (geopolitical entity), or 'FAC' classes (facility) (Abdollah Pour & Momtazi, 2022; Farahani et al., 2021; Mohseni & Tebbifakhr, 2019; Taher et al., 2020). Moreover, the entity type 'Location' does not necessarily refer to places (Berragan, Singleton, Calafiore, & Morley, 2023; Stock et al., 2022).

These limitations are further evidenced in the evaluation of ParsBERT (ARMAN) and ParsBERT (PEYMA) models, as presented in Table 7, where both models underperform compared to our rule-based approach. Although these models are based on the BERT architecture and fine-tuned on Persian NER corpora, the datasets they rely on—namely ARMAN and PEYMA—primarily include coarse-grained location entities such as country and city names, and rarely cover fine-grained urban place names. As a result, their performance in identifying detailed urban entities mentioned in housing listings is notably weaker.

Another challenge observed in these transformer-based models is their tendency to merge multiple adjacent urban place names into a single entity. This behavior stems from the general annotation schemes used in their training corpora. For instance, a street name followed by an alley name is often recognized as one unified entity. Such entity grouping inevitably leads to reduced precision and recall, especially when accurate boundary detection is critical for downstream spatial analysis.

4.3. Extracting spatial relationships between the place names

The result of applying the pattern-based model in extracting spatial relationships is provided in Table 8. The model achieved recall values ranging from 0.42 to 0.52 and precision values ranging from 0.50 to 0.60.

Table 8: Recall, Precision, and F-score for models of each metropolis in extracting spatial relationships between the place names

Metropolis	Recall	Precision	F-Score
Tehran	0.45	0.59	0.51
Mashhad	0.42	0.50	0.45
Shiraz	0.52	0.60	0.55

Even though the preliminary step of the place type recognition was almost successful, some exceptions in using linguistic clues have caused false predictions in extraction. For example, we considered 'located in' as a clue for residual areas. However, it is also used for way place names in some cases. Consequently, the extracted spatial relationship would be 'LinsideR' instead of 'LtouchesW'. To extract correct spatial relationships, we defined the process of identifying the most important place name based on the place type and the ordering of the participants. Thus, misordering leads to a false negative and a false positive prediction. For instance, assume a residential area place name (R) is written after a way place name (W). If the R is considered the more important one, the extracted statement would be 'W is inside R'; otherwise, 'R touches W'. That is also valid for those that are common in place type. Remembering that we incorporated the 'flows from' with the 'touches' relationship, the vast number of false predictions resulting from this issue is way place names had 'touches' relationships but in opposite forms.

4.4. Limitations and future work

The study focused on providing a framework to gather geographic information that could be useful for enriching current gazetteers. However, the study has some limitations. Firstly, it did not address issues such as developing a geographic knowledge base (GKB) and comparison to the current gazetteers and GKBs, which may be addressed in future work. Secondly, data unavailability at both urban and suburban levels limits the consideration of more cities as study regions. It causes some parts of a considered city to be ignored due to the low participation rate in publishing geotagged online listings. Lastly, existing NER tools developed for Persian do not provide sufficient performance in detecting location entities (urban place names), which led to the development of a rule-based model. Although the ad-hoc model gives good results, applying fine-grained NER tools can significantly empower the framework's performance (Ringland et al., 2019; Schiersch et al., 2020; Tao et al., 2022).

We focused on extracting geographic information from each housing listing for the two last tiers. However, organizing information at a metropolis level requires further steps, such as place name resolution. Organizing extracted spatial relationships in a semantic structure could also lead to extracting secondary relationships like 'located in the same street' and 'located in the same neighborhood', besides new ordinary relations. In addition to harvesting place names, there is potential for delineating their location and boundary from housing listings. However, inattention to surrounding words (e.g. around and towards) leads to a crude boundary for residential areas. Additionally, the geographic location tagged for the property could be a few hundred meters away from the written place name in the description. Future work should respect these geographic footprint issues.

Although housing listing websites are good resources for harvesting main streets and neighborhoods, the frequency and spatial distribution of housing listings can be affected by various factors such as the residents, the age of buildings, the included POIs, and the population of a region of a city. For instance, assume a neighborhood in the downtown of

a city. However, due to its commercial importance, the number of published housing listings is deficient in filtering out during data cleaning. Additionally, geotagged listings are a small proportion of all publishing housing listings. In some regions, people often prefer to skip the geographic location field (by selecting a point on the base map) and represent it in other related parts like the urban district or only mention it in the description field. Hence, some regions' neighborhood and main street n-grams would have a handful of points. In some cases, the point set of a main street can be split into two populated parts by natural or artificial features, making it difficult to identify it as a usual main street. The point pattern of a main street n-gram can also be affected by the nearby places homonym to the main street name, especially a square connected to it.

5. Conclusion

Gazetteers are important geospatial resources in GIR tasks, especially in geoparsing. This paper presented a framework for extracting urban geographic information from online property listings. This geographic information includes the place names and the spatial relationships to enrich current gazetteers. Since main streets and neighborhoods as a part of place names are well-known, people mainly use them without any clue on property listing websites. Harvesting these place names can be done using a machine learning-based model. The main challenge is that residents can use the main street place name as a neighborhood. The next step is extracting all place names written in the property advertisement posts. To realize that, we developed a rule-based model to extract potential place names from the posts geographically located in the neighborhood/main street place name's convex-hull and remove the wrong identified cases. In the third step, we extracted spatial relationships between the place names extracted from each post text based on linguistic patterns. The framework has provided good results in harvesting main streets and neighborhoods and extracting place names. Extracting spatial relationships between the place names needs further work. However, the first results are still hopeful by ignoring related issues.

Appendix A

Table A1: Spatial Predictors

	Spatial Predictor	**Short Form**
Spatial Dispersion	Standard Distance	StandardDistance
	The average distance from the 1st nearest neighbor by mean	MeanNN1
	The average distance from the 5th nearest neighbor by mean	MeanNN5
	The average distance from the 10th nearest neighbor by mean	MeanNN10
	The average distance from the 1st nearest neighbor by median	MedianNN1
	The average distance from the 5th nearest neighbor by median	MedianNN5
	The average distance from the 10th nearest neighbor by median	MedianNN10
	The average pairwise distance by mean	MeanPW
	The average pairwise distance by median	MedianPW
	The ratio of maximum distance from the 1st nearest neighbor to MeanNN1	PeaktoAverage

Spatial Homogeneity	The Ripley's L function value at 50 m	RL1
	The Ripley's L function value at 150 m	RL2
	The Ripley's L function value at 250 m	RL3
	The Ripley's L function value at 350 m	RL4
	The Ripley's L function value at 450 m	RL5
	The Ripley's L function value at 550 m	RL6
	The Ripley's L function value at 650 m	RL7
	The Ripley's L function value at 750 m	RL8
	The Ripley's L function value at 850 m	RL9
	The Ripley's L function value at 950 m	RL10
	The skewness of Ripley's L function	Skewness
	The kurtosis of Ripley's L function	Kurtosis
Others	The area of the convex hull of the point set	ConvexHullArea
	The ratio of the number of points to ConvexhullArea	ConvexHullDensity
	The ratio of the rotating bounding rectangle's width to its length	WLR

Conflicts of Interest: The authors declare no conflicts of interest.

Data and codes availability statement: The data and code supporting this study's findings are available at https://figshare.com/s/217847d533e31055865c .

References

Abbasi, O. R., & Alesheikh, A. A. (2023). A Place Recommendation Approach Using Word Embeddings in Conceptual Spaces. *IEEE Access, 11*, 11871-11879. doi:10.1109/ACCESS.2023.3241806

Abdollah Pour, M. M., & Momtazi, S. (2022). Comparative study of text representation and learning for Persian named entity recognition. *ETRI Journal, 44*(5), 794-804. doi:10.4218/etrij.2021-0269

Acheson, E., De Sabbata, S., & Purves, R. S. (2017). A quantitative analysis of global gazetteers: Patterns of coverage for common feature types. *Computers, Environment and Urban Systems, 64*, 309-320. doi:10.1016/j.compenvurbsys.2017.03.007

Acheson, E., & Purves, R. S. (2021). Extracting and modeling geographic information from scientific articles. *PloS one, 16*(1), e0244918. doi:10.1371/journal.pone.0244918

Adams, B. (2017). Wāhi, a discrete global grid gazetteer built using linked open data. *International journal of digital earth, 10*(5), 490-503. doi:10.1080/17538947.2016.1229819

Ahmadi, F., & Moradi, H. (2015). *A hybrid method for Persian named entity recognition.* Paper presented at the 2015 7th conference on information and knowledge technology (IKT).

Aldstadt, J. (2010). Spatial clustering. In *Handbook of applied spatial analysis* (pp. 279-300): Springer.

Asgari-Bidhendi, M., Janfada, B., Roshani Talab, O., & Minaei-Bidgoli, B. (2021). ParsNER-Social: A Corpus for Named Entity Recognition in Persian Social Media Texts. *Journal of AI and Data Mining, 9*(2), 181-192.

Avvenuti, M., Cresci, S., Del Vigna, F., Fagni, T., & Tesconi, M. (2018). CrisMap: a big data crisis mapping system based on damage detection and geoparsing. *Information Systems Frontiers, 20*, 993-1011. doi:10.1007/s10796-018-9833-z

Ayala-Izurieta, J. E., Márquez, C. O., García, V. J., Recalde-Moreno, C. G., Rodríguez-Llerena, M. V., & Damián-Carrión, D. A. (2017). Land cover classification in an ecuadorian mountain geosystem using a random forest classifier, spectral vegetation indices, and ancillary geographic data. *Geosciences, 7*(2), 34. doi:10.3390/geosciences7020034

Bahrehdar, A. R., Adams, B., & Purves, R. S. (2020). Streets of London: Using Flickr and OpenStreetMap to build an interactive image of the city. *Computers, Environment and Urban Systems, 84*, 101524. doi:10.1016/j.compenvurbsys.2020.101524

Berragan, C., Singleton, A., Calafiore, A., & Morley, J. (2023). Transformer based named entity recognition for place name extraction from unstructured text. *International Journal of Geographical Information Science, 37*(4), 747-766. doi:10.1080/13658816.2022.2133125

Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test, 25*(2), 197-227. doi:10.1007/s11749-016-0481-7

Blanchi, A., Fusco, G., Emsellem, K., & Cadorel, L. (2022). *Studying Urban Space from Textual Data: Toward a Methodological Protocol to Extract Geographic Knowledge from Real Estate Ads.* Paper presented at the International Conference on Computational Science and Its Applications.

Bokaei, M. H., Nouri, M., & Sepahvand, A. (2021). Improving Persian Named Entity Recognition Through Multi Task Learning. *International Journal of Information and Communication Technology Research, 13*(2), 39-48. doi:10.52547/itrc.13.2.39

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5-32. doi:10.1023/A:1010933404324

Brindley, P., Goulding, J., & Wilson, M. L. (2018). Generating vague neighbourhoods through data mining of passive web data. *International Journal of Geographical Information Science, 32*(3), 498-523. doi:10.1080/13658816.2017.1400549

Cadorel, L., Blanchi, A., & Tettamanzi, A. G. (2021). *Geospatial Knowledge in Housing Advertisements: Capturing and Extracting Spatial Information from Text.* Paper presented at the Proceedings of the 11th on Knowledge Capture Conference.

Cardoso, S. D., Amanqui, F. K., Serique, K. J., dos Santos, J. L., & Moreira, D. A. (2016). SWI: a semantic web interactive gazetteer to support linked open data. *Future Generation Computer Systems, 54*, 389-398. doi:10.1016/j.future.2015.05.006

Chang, S., Wang, Z., Mao, D., Guan, K., Jia, M., & Chen, C. (2020). Mapping the Essential Urban Land Use in Changchun by Applying Random Forest and Multi-Source Geospatial Data. *Remote Sensing, 12*(15), 2488. doi:10.3390/rs12152488

Chen, H., Vasardani, M., Winter, S., & Tomko, M. (2018). A graph database model for knowledge extracted from place descriptions. *ISPRS International Journal of Geo-Information, 7*(6), 221. doi:10.3390/ijgi7060221

Chen, X., Vo, H., Wang, Y., & Wang, F. (2018). A framework for annotating OpenStreetMap objects using geo-tagged tweets. *GeoInformatica, 22*, 589-613. doi:10.1007/s10707-018-0323-8

Cheng, C., Zhang, T., Su, K., Gao, P., & Shen, S. (2019). Assessing the intensity of the population affected by a complex natural disaster using social media data. *ISPRS International Journal of Geo-Information, 8*(8), 358. doi:10.3390/ijgi8080358

Chopin, J., & Caneppele, S. (2019). Geocoding child sexual abuse: An explorative analysis on journey to crime and to victimization from French police data. *Child Abuse & Neglect, 91*, 116-130. doi:10.1016/j.chiabu.2019.03.001

Clementini, E., Di Felice, P., & Van Oosterom, P. (1993). *A small set of formal topological relationships suitable for end-user interaction.* Paper presented at the Advances in Spatial Databases: Third International Symposium, SSD'93 Singapore, June 23–25, 1993 Proceedings 3.

Cruz, P., Vanneschi, L., Painho, M., & Rita, P. (2021). Automatic Identification of Addresses: A Systematic Literature Review. *ISPRS International Journal of Geo-Information, 11*(1), 11. doi:10.3390/ijgi11010011

de Bruijn, J. A., de Moel, H., Jongman, B., de Ruiter, M. C., Wagemaker, J., & Aerts, J. C. (2019). A global database of historic and real-time flood events based on social media. *Scientific data, 6*(1), 311. doi:10.1038/s41597-019-0326-9

De Longueville, B., Ostländer, N., & Keskitalo, C. (2010). Addressing vagueness in Volunteered Geographic Information (VGI)–A case study. *International Journal of Spatial Data Infrastructures Research, 5*, 1725-0463.

Deidda, M., Pala, A., & Vacca, G. (2013). An example of a tourist location-based service (LBS) with open-source software. *Applied Geomatics, 5*, 73-86. doi:10.1007/s12518-012-0097-x

Delmelle, E. C., & Nilsson, I. (2021). The language of neighborhoods: A predictive-analytical framework based on property advertisement text and mortgage lending data. *Computers, Environment and Urban Systems, 88*, 101658. doi:10.1016/j.compenvurbsys.2021.101658

Delmelle, E. M., Marsh, D. M., Dony, C., & Delamater, P. L. (2021). Travel impedance agreement among online road network data providers. In *Uncertainty and Context in GIScience and Geography* (pp. 122-140): Routledge.

Derungs, C., & Purves, R. S. (2016). Mining nearness relations from an n-grams Web corpus in geographical space. *Spatial Cognition & Computation, 16*(4), 301-322. doi:10.1080/13875868.2016.1246553

Dewandaru, A., Widyantoro, D. H., & Akbar, S. (2020). Event geoparser with pseudo-location entity identification and numerical argument extraction implementation and evaluation in Indonesian news domain. *ISPRS International Journal of Geo-Information, 9*(12), 712. doi:10.3390/ijgi9120712

Etemad, S., Mosayebi, R., Khodavirdian, T. A., Dastan, E., Telmadarreh, A. S., Jafari, M., & Rafiei, S. (2023). Clustering of Urban Traffic Patterns by K-Means and Dynamic Time Warping: Case Study. *arXiv preprint arXiv:2309.09830*. doi:10.48550/arXiv.2309.09830

Farahani, M., Gharachorloo, M., Farahani, M., & Manthouri, M. (2021). ParsBERT: Transformer-based Model for Persian Language Understanding. *Neural Processing Letters, 53*(6), 3831-3847. doi:10.1007/s11063-021-10528-4

Gao, S., Janowicz, K., Montello, D. R., Hu, Y., Yang, J.-A., McKenzie, G., . . . Yan, B. (2017). A data-synthesis-driven method for detecting and extracting vague cognitive regions. *International Journal of Geographical Information Science, 31*(6), 1245-1271. doi:10.1080/13658816.2016.1273357

Gao, S., Li, L., Li, W., Janowicz, K., & Zhang, Y. (2017). Constructing gazetteers from volunteered big geo-data based on Hadoop. *Computers, Environment and Urban Systems, 61*, 172-186. doi:10.1016/j.compenvurbsys.2014.02.004

Ghayoomi, M., & Momtazi, S. (2009). *Challenges in developing Persian corpora from online resources.* Paper presented at the International Conference on Asian Language Processing.

Goodchild, M. F., & Hill, L. L. (2008). Introduction to digital gazetteer research. *International Journal of Geographical Information Science, 22*(10), 1039-1044. doi:10.1080/13658810701850497

Grace, R. (2021). Toponym usage in social media in emergencies. *International Journal of Disaster Risk Reduction, 52*, 101923. doi:10.1016/j.ijdrr.2020.101923

Gritta, M. (2019). *Where are you talking about? advances and challenges of geographic analysis of text with application to disease monitoring.* University of Cambridge,

Gritta, M., Pilehvar, M. T., Limsopatham, N., & Collier, N. (2018). What's missing in geographical parsing? *Language Resources and Evaluation, 52*(2), 603-623. doi:10.1007/s10579-017-9385-8

Grossner, K., Janowicz, K., & Keßler, C. (2016). Place, period, and setting for linked data gazetteers. *Placing names: Enriching and integrating gazetteers*, 80-96.

Haberman, C. P., Hatten, D., Carter, J. G., & Piza, E. L. (2021). The sensitivity of repeat and near repeat analysis to geocoding algorithms. *Journal of criminal justice, 73*, 101721. doi:10.1016/j.jcrimjus.2020.101721

Habib, M. K. (2021). The Challenges of Persian User-generated Textual Content: A Machine Learning-Based Approach. *arXiv preprint arXiv:2101.08087*. doi:10.48550/arXiv.2101.08087

Haris, E., & Gan, K. H. (2017). Mining graphs from travel blogs: a review in the context of tour planning. *Information Technology & Tourism, 17*, 429-453.

Hastings, J., & Hill, L. (2002). Treatment of duplicates in the alexandria digital library gazetteer. *Proceedings of GeoScience*.

Hill, L. L. (2000). *Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints.* Paper presented at the Research and Advanced Technology for Digital Libraries, Berlin, Heidelberg.

Hu, X., Al-Olimat, H. S., Kersten, J., Wiegmann, M., Klan, F., Sun, Y., & Fan, H. (2022). GazPNE: annotation-free deep learning for place name extraction from microblogs leveraging gazetteer and synthetic data by rules. *International Journal of Geographical Information Science, 36*(2), 310-337. doi:10.1080/13658816.2021.1947507

Hu, Y. (2018). Geo-text data and data-driven geospatial semantics. *Geography Compass, 12*(11), e12404. doi:10.1111/gec3.12404

Hu, Y., & Adams, B. (2021). Harvesting big geospatial data from natural language texts. In *Handbook of Big Geospatial Data* (pp. 487-507): Springer.

Hu, Y., Gao, S., Janowicz, K., Yu, B., Li, W., & Prasad, S. (2015). Extracting and understanding urban areas of interest using geotagged photos. *Computers, Environment and Urban Systems, 54*, 240-254. doi:10.1016/j.compenvurbsys.2015.09.001

Hu, Y., Mao, H., & McKenzie, G. (2019). A natural language processing and geospatial clustering framework for harvesting local place names from geotagged housing advertisements. *International Journal of Geographical Information Science, 33*(4), 714-738. doi:10.1080/13658816.2018.1458986

Hu, Y., & Wang, J. (2020). *How Do People Describe Locations During a Natural Disaster: An Analysis of Tweets from Hurricane Harvey.* Paper presented at the 11th International Conference on Geographic Information Science (GIScience 2021)-Part I.

Huck, J., Whyatt, J., & Coulton, P. (2014). Spraycan: A PPGIS for capturing imprecise notions of place. *Applied geography, 55*, 229-237. doi:10.1016/j.apgeog.2014.09.007

Idakwo, P. O., Adekanmbi, O., Soronnadi, A., & David, A. (2025). Geo-parsing and analysis of road traffic crash incidents for data-driven emergency response planning. *Heliyon, 11*(4). doi:10.1016/j.heliyon.2024.e41067

Imani, M. B., Chandra, S., Ma, S., Khan, L., & Thuraisingham, B. (2017). *Focus location extraction from political news reports with bias correction.* Paper presented at the 2017 IEEE International Conference on Big Data (Big Data).

Janowicz, K., & Keßler, C. (2008). The role of ontology in improving gazetteer interaction. *International Journal of Geographical Information Science, 22*(10), 1129-1157. doi:10.1080/13658810701851461

Javidaneh, A., Karimipour, F., & Alinaghi, N. (2020). How Much Do We Learn from Addresses? On the Syntax, Semantics and Pragmatics of Addressing Systems. *ISPRS International Journal of Geo-Information, 9*(5), 317. doi:10.3390/ijgi9050317

Jones, C. B., Purves, R. S., Clough, P. D., & Joho, H. (2008). Modelling vague places with knowledge from the Web. *International Journal of Geographical Information Science, 22*(10), 1045-1065. doi:10.1080/13658810701850547

Karimzadeh, M., Pezanowski, S., MacEachren, A. M., & Wallgrün, J. O. (2019). GeoTxt: A scalable geoparsing system for unstructured text geolocation. *Transactions in GIS, 23*(1), 118-136. doi:10.1111/tgis.12510

Keßler, C., Janowicz, K., & Bishr, M. (2009). *An agenda for the next generation gazetteer: Geographic information contribution and retrieval.* Paper presented at the Proceedings of the 17th ACM SIGSPATIAL international conference on advances in Geographic Information Systems.

Keßler, C., Maué, P., Heuer, J. T., & Bartoschek, T. (2009). *Bottom-up gazetteers: Learning from the implicit semantics of geotags.* Paper presented at the International conference on geospatial semantics.

Kim, J., Vasardani, M., & Winter, S. (2015). *Harvesting large corpora for generating place graphs.* Paper presented at the International Workshop on Cognitive Engineering for Spatial Information Processes.

Kordjamshidi, P., Van Otterlo, M., & Moens, M.-F. (2011). Spatial role labeling: Towards extraction of spatial relations from natural language. *ACM Transactions on Speech and Language Processing (TSLP), 8*(3), 1-36. doi:10.1145/2050104.2050105

Laurini, R. (2015). Geographic ontologies, gazetteers and multilingualism. *Future Internet, 7*(1), 1-23. doi:10.3390/fi7010001

Leppämäki, T., Toivonen, T., & Hiippala, T. (2024). Geographical and linguistic perspectives on developing geoparsers with generic resources. *International Journal of Geographical Information Science, 38*(10), 2039-2060. doi:10.1080/13658816.2024.2369539

Levy, D., & Lee, C. K. (2011). Neighbourhood identities and household location choice: estate agents' perspectives. *Journal of Place Management and Development, 4*(3), 243-263. doi:10.1108/17538331111176066

Li, D., Cova, T. J., & Dennison, P. E. (2017). Using reverse geocoding to identify prominent wildfire evacuation trigger points. *Applied geography, 87*, 14-27. doi:10.1016/j.apgeog.2017.05.008

Li, J., Liu, R., & Xiong, R. (2017). *A Chinese Geographic Knowledge Base for GIR.* Paper presented at the IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC).

Li, L., Wang, W., He, B., & Zhang, Y. (2018). A hybrid method for Chinese address segmentation. *International Journal of Geographical Information Science, 32*(1), 30-48. doi:10.1080/13658816.2017.1379084

Lim, J., Nitta, N., Nakamura, K., & Babaguchi, N. (2019). Constructing Geographic Dictionary from Streaming Geotagged Tweets. *ISPRS International Journal of Geo-Information, 8*(5). doi:10.3390/ijgi8050216

Liu, Y., Li, R., Chen, K., Yuan, Y., Huang, L., & Yu, H. (2009). *KIDGS: A geographical knowledge-informed digital gazetteer service.* Paper presented at the 2009 17th International Conference on Geoinformatics.

Lynch, K. (1960). *The image of the city*: MIT Press.

Machado, I. M. R., de Alencar, R. O., de Oliveira Campos, R., & Davis, C. A. (2011). An ontological gazetteer and its application for place name disambiguation in text. *Journal of the Brazilian Computer Society, 17*(4), 267-279. doi:10.1007/s13173-011-0044-4

Madhulatha, T. S. (2012). An overview on clustering methods. *arXiv preprint arXiv:1205.1117*. doi:10.48550/arXiv.1205.1117

Mai, G., Janowicz, K., Prasad, S., Shi, M., Cai, L., Zhu, R., . . . Lao, N. (2020). Semantically-enriched search engine for Geoportals: A case study with ArcGIS online. *AGILE: GIScience Series, 1*, 13. doi:10.5194/agile-giss-1-13-2020

Manguinhas, H., Martins, B., Borbinha, J., & Siabato Vaca, W. L. (2009). The DIGMAP geo-temporal Web gazetteer service. *e-Perimetron, 4*(1), 9-24.

Martins, B. (2011). *A supervised machine learning approach for duplicate detection over gazetteer records.* Paper presented at the International Conference on GeoSpatial Semantics.

McKenzie, G., & Janowicz, K. (2015). Where is also about time: A location-distortion model to improve reverse geocoding using behavior-driven temporal semantic signatures. *Computers, Environment and Urban Systems, 54*, 1-13. doi:10.1016/j.compenvurbsys.2015.05.003

McKenzie, G., Liu, Z., Hu, Y., & Lee, M. (2018). Identifying urban neighborhood names through user-contributed online property listings. *ISPRS International Journal of Geo-Information, 7*(10). doi:10.3390/ijgi7100388

Mehta, A., Kim, D., Allo, N., Odusola, A. O., Malolan, C., & Nwariaku, F. E. (2023). Using parallel geocoding to analyse the spatial characteristics of road traffic injury occurrences across Lagos, Nigeria. *BMJ global health, 8*(5), e012315. doi:10.1136/bmjgh-2023-012315

Merschdorf, H., & Blaschke, T. (2018). Revisiting the role of place in geographic information science. *ISPRS International Journal of Geo-Information, 7*(9), 364. doi:10.3390/ijgi7090364

Middleton, S. E., Kordopatis-Zilos, G., Papadopoulos, S., & Kompatsiaris, Y. (2018). Location extraction from social media: Geoparsing, location disambiguation, and geotagging. *ACM Transactions on Information Systems (TOIS), 36*(4), 1-27. doi:10.1145/3202662

Middleton, S. E., & Krivcovs, V. (2016). Geoparsing and geosemantics for social media: Spatiotemporal grounding of content propagating rumors to support trust and veracity analysis during breaking news. *ACM Transactions on Information Systems (TOIS), 34*(3), 1-26. doi:10.1145/2842604

Mohseni, M., & Tebbifakhr, A. (2019). *MorphoBERT: A Persian NER system with BERT and morphological analysis.* Paper presented at the Proceedings of The First International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2019) co-located with ICNLSP 2019-Short Papers.

Moura, T. H., Davis Jr, C. A., & Fonseca, F. T. (2017). Reference data enhancement for geographic information retrieval using linked data. *Transactions in GIS, 21*(4), 683-700. doi:10.1111/tgis.12238

Nallur, V., Elgammal, A., & Clarke, S. (2015). *Smart route planning using open data and participatory sensing.* Paper presented at the Open Source Systems: Adoption and Impact: 11th IFIP WG 2.13 International Conference, OSS 2015, Florence, Italy, May 16-17, 2015, Proceedings 11.

Oliveira, M. G. d., Campelo, C. E., Baptista, C. d. S., & Bertolotto, M. (2016). Gazetteer enrichment for addressing urban areas: a case study. *Journal of Location Based Services, 10*(2), 142-159. doi:10.1080/17489725.2016.1196755

Oto-Peralías, D. (2018). What do street names tell us? The 'city-text'as socio-cultural data. *Journal of Economic Geography, 18*(1), 187-211. doi:10.1093/jeg/lbx030

Poostchi, H., Borzeshi, E. Z., Abdous, M., & Piccardi, M. (2016). *PersoNER: Persian named-entity recognition.* Paper presented at the COLING 2016-26th International Conference on Computational Linguistics, Proceedings of COLING 2016: Technical Papers.

Poostchi, H., Borzeshi, E. Z., & Piccardi, M. (2018). *Bilstm-crf for persian named-entity recognition armanpersonercorpus: the first entity-annotated persian dataset.* Paper presented at the Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

Popescu, A., Grefenstette, G., & Moëllic, P. A. (2008). *Gazetiki: automatic creation of a geographical gazetteer.* Paper presented at the Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries.

Pustejovsky, J. (2017). ISO-Space: Annotating static and dynamic spatial information. *Handbook of linguistic annotation*, 989-1024. doi:10.1007/978-94-024-0881-2_37

Qazi, U., Imran, M., & Ofli, F. (2020). GeoCoV19: a dataset of hundreds of millions of multilingual COVID-19 tweets with location information. *SIGSPATIAL Special, 12*(1), 6-15. doi:10.1145/3404820.3404823

Rattenbury, T., Good, N., & Naaman, M. (2007). *Towards automatic extraction of event and place semantics from flickr tags.* Paper presented at the Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval.

Ringland, N., Dai, X., Hachey, B., Karimi, S., Paris, C., & Curran, J. R. (2019). NNE: A dataset for nested named entity recognition in English newswire. *arXiv preprint arXiv:1906.01359*. doi:10.48550/arXiv.1906.01359

Sabzali Yameqani, A., & Alesheikh, A. A. (2025). Enhancing Reverse Geocoding With Weather Data: Modeling Human Check-In Behavior in California and New York for Smart Cities. *Transactions in GIS, 29*(3), e70059. doi:10.1111/tgis.70059

Samarin, M., & Sharma, M. (2020). A Neighborhood-level Perspective of Real Estate Determinants in Three US Cities. *International Journal of Geospatial and Environmental Research, 7*(3), 2.

Schiersch, M., Mironova, V., Schmitt, M., Thomas, P., Gabryszak, A., & Hennig, L. (2020). A german corpus for fine-grained named entity recognition and relation extraction of traffic and industry events. *arXiv preprint arXiv:2004.03283*. doi:10.48550/arXiv.2004.03283

Schubert, E., Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (2017). DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems (TODS), 42*(3), 1-21. doi:10.1145/3068335

Shahshahani, M. S., Mohseni, M., Shakery, A., & Faili, H. (2018). PEYMA: A tagged corpus for Persian named entities. *arXiv preprint arXiv:1801.09936*. doi:10.48550/arXiv.1801.09936

Shamsfard, M. (2011). *Challenges and open problems in Persian text processing.* Paper presented at the Proceedings of LTC.

Shamsfard, M. M., Puneh-Sadat. (2009). *Named Entity Recognition in Persian texts*. Paper presented at the 15th International Conference of Irainian Computer Community.

Stock, K., Jones, C. B., Russell, S., Radke, M., Das, P., & Aflaki, N. (2022). Detecting geospatial location descriptions in natural language text. *International Journal of Geographical Information Science, 36*(3), 547-584. doi:10.1080/13658816.2021.1987441

Stock, K., & Yousaf, J. (2018). Context-aware automated interpretation of elaborate natural language descriptions of location through learning from empirical data. *International Journal of Geographical Information Science, 32*(6), 1087-1116. doi:10.1080/13658816.2018.1432861

Suat-Rojas, N., Gutierrez-Osorio, C., & Pedraza, C. (2022). Extraction and analysis of social networks data to detect traffic accidents. *Information, 13*(1), 26. doi:10.3390/info13010026

Surano, F. V., Porfiri, M., & Rizzo, A. (2022). Analysis of lockdown perception in the United States during the COVID-19 pandemic. *The European Physical Journal Special Topics, 231*(9), 1625-1633. doi:10.1140/epjs/s11734-021-00265-z

Suwaileh, R., Elsayed, T., Imran, M., & Sajjad, H. (2022). When a disaster happens, we are ready: Location mention recognition from crisis tweets. *International Journal of Disaster Risk Reduction, 78*, 103107. doi:10.1016/j.ijdrr.2022.103107

Taher, E., Hoseini, S. A., & Shamsfard, M. (2020). Beheshti-NER: Persian named entity recognition Using BERT. *arXiv preprint arXiv:2003.08875*. doi:10.48550/arXiv.2003.08875

Talen, E., & Jeong, H. (2019). What is the value of 'main street'? Framing and testing the arguments. *Cities, 92*, 208-218. doi:10.1016/j.cities.2019.03.023

Tao, L., Xie, Z., Xu, D., Ma, K., Qiu, Q., Pan, S., & Huang, B. (2022). Geographic Named Entity Recognition by Employing Natural Language Processing and an Improved BERT Model. *ISPRS International Journal of Geo-Information, 11*(12), 598. doi:10.3390/ijgi11120598

Tewari, S., & Beynon, D. (2018). Changing neighbourhood character in Melbourne: Point Cook a case study. *Journal of urban design, 23*(3), 456-464. doi:10.1080/13574809.2017.1383152

Tian, Q., Ren, F., Hu, T., Liu, J., Li, R., & Du, Q. (2016). Using an optimized Chinese address matching method to develop a geocoding service: a case study of Shenzhen, China. *ISPRS International Journal of Geo-Information, 5*(5), 65. doi:10.3390/ijgi5050065

Twaroch, F. A., & Jones, C. B. (2010). *A web platform for the evaluation of vernacular place names in automatically constructed gazetteers.* Paper presented at the Proceedings of the 6th Workshop on Geographic Information Retrieval.

W3Techs. (2021). Usage statistics of content languages for websites. Retrieved from https://w3techs.com/technologies/history_overview/content_language/ms/y

Wallgrün, J. O., Klippel, A., & Baldwin, T. (2014). *Building a corpus of spatial relational expressions extracted from web documents.* Paper presented at the Proceedings of the 8th workshop on geographic information retrieval.

Wang, W., & Stewart, K. (2015). Spatiotemporal and semantic information extraction from Web news reports about natural hazards. *Computers, Environment and Urban Systems, 50*, 30-40. doi:10.1016/j.compenvurbsys.2014.11.001

Windfuhr, G. (2009). *The Iranian Languages*. London: Routledge.

Won, M., Murrieta-Flores, P., & Martins, B. (2018). Ensemble Named Entity Recognition (NER): Evaluating NER Tools in the Identification of Place Names in Historical Corpora. *Frontiers in Digital Humanities, 5*. doi:10.3389/fdigh.2018.00002

Yameqani, A. S., & Alesheikh, A. A. (2019). Evaluating a location distortion model to improve reverse geocoding through temporal semantic signatures. *Computers, Environment and Urban Systems, 77*, 101349. doi:10.1016/j.compenvurbsys.2019.101349

Yang, J., Yu, M., Qin, H., Lu, M., & Yang, C. (2019). A twitter data credibility framework—Hurricane harvey as a use case. *ISPRS International Journal of Geo-Information, 8*(3). doi:10.3390/ijgi8030111

Zhang, C., He, B., Guo, R., & Ma, D. (2023). A graph-based approach for representing addresses in geocoding. *Computers, Environment and Urban Systems, 100*, 101937. doi:10.1016/j.compenvurbsys.2022.101937

Zhang, T., Shen, S., Cheng, C., Su, K., & Zhang, X. (2021). A topic model based framework for identifying the distribution of demand for relief supplies using social media data. *International Journal of Geographical Information Science, 35*(11), 2216-2237. doi:10.1080/13658816.2020.1869746

Zhang, W., & Gelernter, J. (2014). Geocoding location expressions in Twitter messages: A preference learning method. *Journal of Spatial Information Science, 2014*(9), 37-70. doi:10.5311/JOSIS.2014.9.170

Zhang, Y., Ma, Q., Chiang, Y.-Y., Knoblock, C., Zhang, X., Yang, P., . . . Hu, X. (2019). Extracting geographic features from the Internet: A geographic information mining framework. *Knowledge-Based Systems, 174*, 57-72. doi:10.1016/j.knosys.2019.02.031

Zhang, Y., Wu, W., Wang, Q., & Su, F. (2017). A geo-event-based geospatial information service: A case study of typhoon hazard. *Sustainability, 9*(4), 534. doi:10.3390/su9040534