Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Review article

Natural language processing for urban research: A systematic review

Meng Cai [*]

School of Planning, Design and Construction, Michigan State University, East Lansing, Michigan, 48824, United States

ARTICLE INFO

Keywords:
Natural language processing
Urban research
Urban big data
Text mining

ABSTRACT

Natural language processing (NLP) has shown potential as a promising tool to exploit under-utilized urban data sources. This paper presents a systematic review of urban studies published in peer-reviewed journals and conference proceedings that adopted NLP. The review suggests that the application of NLP in studying cities is still in its infancy. Current applications fell into five areas: urban governance and management, public health, land use and functional zones, mobility, and urban design. NLP demonstrates the advantages of improving the usability of urban big data sources, expanding study scales, and reducing research costs. On the other hand, to take advantage of NLP, urban researchers face challenges of raising good research questions, overcoming data incompleteness, inaccessibility, and non-representativeness, immature NLP techniques, and computational skill requirements. This review is among the first efforts intended to provide an overview of existing applications and challenges for advancing urban research through the adoption of NLP.

1. Introduction

The advancement of technologies is not just changing cities (Urban Land Institute, 2019); it is also transforming the way urban researchers are able to study cities. Gray's notion of the fourth paradigm of science pointed out that the wide availability of data changes the practice of science (Hey et al., 2009). Abundant urban big data is being generated and stored at unprecedented speed and scale; researchers nowadays are able to ask and answer questions in ways that were impossible in the past.

The paradigm shift of scientific research highlights the need for a new generation of scientific tools and methods. Among all existing data, 95% are in unstructured form, which lacks an identifiable tabular organization required by traditional data analysis methods (Gandomi and Haider, 2015). Unstructured data, such as Web pages, emails, and mobile phone records, may contain numerical information (e.g. dates) but is usually text-heavy. Unlike numbers, textual data are inherently inaccurate and vague. According to a conservative estimate by Britton (1978), at least 32% of the words used in English text are lexically ambiguous. The messy reality of textual data makes it challenging for researchers to take advantage of urban big data.

On the other hand, the large quantity of textual data provides new opportunities for urban researchers to examine people's perceptions, attitudes, and behaviors, so as to advance the knowledge and understanding of urban dynamics. For example, Jang and Kim (2019) have proved that crowd-sourced text data gathered from social media can effectively represent the collective identity of urban space. Conventional data gathering techniques, such as surveys, focus groups, and interviews, are oftentimes expensive and time-consuming. If used wisely, organic text data without pre-specified purposes could be incredibly powerful and complement purposefully designed data collection.

Natural language processing (NLP) has demonstrated tremendous capabilities of harvesting the abundance of textual data. As a form of artificial intelligence, it uses computational algorithms to learn, understand, and produce human language content (Hirschberg and Manning, 2015). It is interrelated with machine learning and deep learning. Basic NLP procedures include processing text data, converting text to features, and identifying semantic relationships (Ghosh and Gunning, 2019). In addition to its ability to structure large volumes of unstructured data, NLP can improve the accuracy of text processing and analysis because it follows rules and criteria in a consistent way. NLP has proven to be useful in many fields. For example, in medical research, Guetterman et al. (2018) conducted an experiment to compare the results from an NLP analysis and a traditional text analysis. They reported that NLP was able to identify major themes that were manually summarized by traditional text analysis.

Here, a comprehensive review of the ways that researchers have utilized NLP in urban studies is presented. This work is among the first efforts intended to provide a synthesis of opportunities and challenges for advancing urban research through the adoption of NLP.

[*] Corresponding author.
E-mail address: caimeng2@msu.edu.

https://doi.org/10.1016/j.heliyon.2021.e06322
Received 8 November 2020; Received in revised form 2 February 2021; Accepted 16 February 2021

Table 1. Literature search criteria.

Database	Search term	Search field	Subject area	Source/document type	Other filter
EBSCO Urban Studies Abstracts	“natural language processing”	“Title, Abstract, or Keywords”	N/A	N/A	N/A
Scopus	“natural language processing” AND (city OR urban)	“Title, Abstract, or Keywords:	“Social sciences”	“Journals OR conference proceedings”	N/A
ProQuest	“natural language processing” AND (city OR urban)	“Anywhere except full text”	N/A	“Conference Papers & Proceedings OR Scholarly Journals”	“Peer reviewed”
Web of Science	“natural language processing” AND (city OR urban)	“Topic” (i.e. title, abstract, author keywords, and Keywords Plus)	N/A	“Article”	N/A

2. Methodology

2.1. Literature search

The aim of this literature search was to gather all scientific publications in urban studies that utilized the method of NLP. To serve this aim, journal articles and conference papers were searched in four online databases: EBSCO Urban Studies Abstracts, Scopus, ProQuest, and Web of Science. Due to the fact that each database has different searchable fields and filtering options, slightly different search criteria were adopted depending on the different databases used (see Table 1). Besides the criteria listed in Table 1, the language of publications in all four database searches was also constrained so the results only included literature in English. The search timeframe was “all years,” which means the results contained all publications to date (November 2019).

The initial search returned 271 publications: 6 from EBSCO Urban Studies Abstracts, 69 from Scopus, 125 from ProQuest, and 71 from Web of Science. After removing 73 duplicates, the titles and abstracts of the remaining articles were reviewed. The publications were further narrowed down by determining that 152 were of irrelevant topics to urban research, such as travel planning, regional linguistic variations, or corpus development; 18 studies did not use the method of NLP; and four articles were without full-text access. The above mentioned 174 articles were removed and the remaining 24 studies were reviewed in full texts. Two articles identified from citations of the included studies were added for review. As a result of reviewing the publications found based on criteria of relevance and full-text access, this study included a total of 26 publications for detailed analysis.

2.2. Limitations

While the strategy used during the literature search was meant to be a comprehensive and systematic approach, it had several limitations. First, the search had a language bias because it only included studies published in English. Articles in non-English languages with English abstracts were not included either. Second, the method for retrieving publications may have excluded studies that used NLP techniques but had been labeled with other terminology. For example, studies that used latent Dirichlet allocation (LDA), a statistical model in NLP, listed LDA as a keyword rather than NLP and therefore did not match the literature search criteria. Third, this review only included peer-reviewed journal articles and conference papers, which eliminated possible NLP applications documented in dissertations, theses, reports, and working papers. This was a tradeoff between literature quantity and quality.

3. Literature search results

3.1. Amount of publications

The systematic literature search returned a total of 26 urban studies that used NLP, of which 21 were journal articles and five were conference papers. All of those appeared from 2012 onwards and more than half

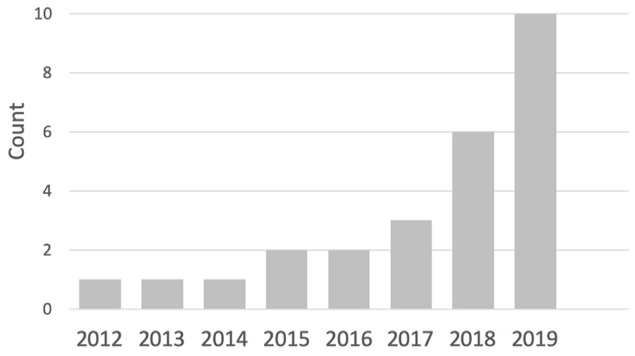

Figure 1. Amount of urban studies using NLP by year.

Table 2. Summary of included literatures.

Study	Topic	NLP application	Data	Study area
Abali, Karaarslan, Hurriyetoglu and Dalkilic (2018)	Detecting citizen problems and their locations	Urban governance and management	100 tweets	Aegean Region, Turkey
Bardhan, Sunikka-Blank and Haque (2019)	Gender mainstreaming in slum rehabilitation housing management in Mumbai, India	Urban governance and management	12 interviews and 2 focus groups	Mumbai, India
Estévez-Ortiz, García-Jiménez and Glösekötter (2016)	An application of people's sentiment from social media to smart cities	Urban governance and management	200,000 tweets	New York City, NY, US
Fu, McKenzie, Frias-Martinez and Stewart (2018)	Identifying spatiotemporal urban activities	Public health	8,098,864 tweets	Baltimore, MD, Washington D.C., and New York City, NY, US
Helderop, Huff, Morstatter, Grubesic and Wallace (2019)	Detecting urban prostitution activities	Urban governance and management	3,387 policy department prostitution arrest records and 10 years' hotel reviews and price data	Phoenix, AZ, US
Hong, Fu, Wu and Frias-Martinez (2018)	Information needs and communication gaps between citizens and local governments online during natural disasters	Urban governance and management	96,423 tweets	Maryland, US
Hu et al. (2019a)	Understanding the perceptions of people toward their living environments based on online neighborhood reviews	Public health	7,673 neighborhood reviews on Niche	New York City, NY, US
Hu et al. (2019b)	A framework for harvesting local place names from geotagged housing advertisements	Urban governance and management	35,852 housing advertisements from Craigslist	New York City, NY, Los Angeles, CA, Chicago, IL, Richmond, VI, Boise, ID, and Spokane, WA, US
Huang et al. (2018)	Quantifying the spatiotemporal dynamics of industrial land uses	Land use and functional zones	POIs data from Gaode Map and Google Earth images	Mega Hangzhou Bay Region, China
Iaconesi (2015)	Emotional landmarks in cities	Urban design	61,516,961 posts from Facebook, Twitter, Instagram, Foursquare, and Yelp	Rome, Milan, and Turin, Italy; Berlin, Germany; Sao Paulo, Brazil; Montreal and Toronto, Canada; New York, NY and New Haven, CT, US; Hong Kong; Cairo, Egypt; and Istanbul, Turkey
Imran, Elbassuoni, Castillo, Diaz and Meier (2013)	Extraction of disaster-relevant information from social media	Urban governance and management	346,764 tweets	Joplin, MO and New York City, NY, US
Jang and Kim (2019)	A crowd-sourced cognitive map to display people's cognitive perception of urban space	Urban design	1,785,768 posts from Instagram	Bundang, Dongtan, Ilsan, and Songdo, Korea
Lai and Kontokosta (2019)	Thematic structure and spatial-temporal patterns of building renovation and adaptive reuse in cities	Urban governance and management	2,500,000 building permits	New York City, NY, Los Angeles, CA, Chicago, IL, Austin, TX, San Francisco, CA, Seattle, WA, and Boston, MA, US
Li, Fei and Zhang (2019)	A regionalization method for clustering and partitioning trajectories	Land use and functional zones	27,000,000 trajectories of call records from mobile phones	Beijing, China
Liu et al. (2017)	Measuring traffic interactions in urban road system from massive travel routes	Mobility	Taxi GPS trajectories	Beijing, China
Liu, Gao and Lu (2019)	Identifying spatial interaction patterns of vehicle movements	Mobility	Taxi trajectories collected by the Global Navigation Satellite System	Beijing, China
Markou, Kaiser and Pereira (2019)	Predicting taxi demand hotspots	Mobility	Taxi pickup/drop-off data and event data from event listing sites	New York City, NY, US

(*continued on next page*)

Table 2 (*continued*)

Study	Topic	NLP application	Data	Study area
Rahimi, Mottahedi and Liu (2018)	The geography of taste and urban culture	Public health	4,100,000 restaurant reviews on Yelp	Boston, MA, Charlotte, NC, Cleveland, OH, Washington D.C., Detroit, MI, Las Vegas, NV, Philadelphia, PA, Phoenix, AZ, and Pittsburgh, PA, US and Toronto, Canada
Riga and Karatzas (2014)	Real-time observations for urban air quality and public health	Public health	17,560 tweets and weather data from the European Centre for Medium-Range Weather Forecasts	Europe
Serna, Gerrikagoitia, Bernabé and Ruiz (2017)	Sustainability analysis of urban mobility	Mobility	43,251 comments from Minube	Bilbao, Valencia, and Madrid, Spain
Souza et al. (2016)	A platform to analyze social streams in smart city initiatives	Urban governance and management	530,000 policy department emergency phone call records and 126 tweets	Natal, Brazil
Vargas-Calderón and Camargo (2019)	Characterization of citizens	Urban governance and management	2,634,176 tweets	Bogotá, Colombia
Wakamiya, Kawai and Aramaki (2018)	Influenza detection	Public health	Tweets spanning 5 years and influenza diagnostic records from the Infectious Disease Surveillance Center	Japan
Yao et al. (2017)	Sensing the spatial distribution of urban land use	Land use and functional zones	High spatial resolution remote-sensing images	Guangzhou, Guangdong, China
Yuan et al. (2015)	Discovering urban functional zones	Land use and functional zones	GPS trajectory data generated by 12,000 taxis and public transit records of 1,500,000 trips from 300,000 card holders	Beijing, China
Yuan, Zheng and Xie (2012)	Discovering regions of different functions in a city	Land use and functional zones	2 POIs datasets and 2 3-month GPS trajectory datasets generated by over 12,000 taxis	Beijing, China

(62%) from 2018 onwards (Figure 1). The exponential increase in the number of publications reflects the growing interest in NLP among urban researchers.

3.2. NLP application

Urban researchers have explored diverse topics using NLP as summarized in Table 2. In general, researchers have applied NLP in five areas: urban governance and management, public health, land use and functional zones, mobility, and urban design. Urban governance and management is the most dominant topic (39% of all literatures), which includes discussions on citizen engagement, disaster response, crime detection, and construction management. Researchers also have used NLP to study urban health (19% of all literatures), such as urban epidemics prediction, air quality monitoring, and assessment of living environments. Land use and functional zones is another popular area of research (19% of all literatures), in which researchers used NLP to model urban spatiotemporal dynamics. Besides, though only a limited number, researchers have adopted NLP in mobility (15% of all literatures) and urban design research (8% of all literatures).

3.3. Data

The majority of studies involved in this review used social media as their data source, including Twitter, Instagram, Facebook, Foursquare, Craigslist, Minube, and Yelp (Table 2). Researchers typically extract and analyze the text along with geolocation information embedded in social media posts. However, the data source is not limited to social media, researchers used NLP to process information gathered from interviews, focus groups, phone call records, building permits, online hotel reviews, event listings, and neighborhood reviews. The data size could be big (e.g. millions of tweets) or small (e.g. a dozen of interviews). Additionally, researchers have extended the usage of NLP from analyzing textual data to non-textual data such as points of interest (POIs) data in maps and GPS trajectories generated by cell phones and taxis. It is worth mentioning that when a study objective was predictive modeling, it was common to check the validity of NLP results with records from official sources.

3.4. Study area

Studies using NLP have covered a wide range of geographic locations (Table 2). Most studies focused on major cities with large populations, such as New York City, US and Beijing, China. Some examined multiple cities for comparison. The scale of analysis ranges from a single city to a continent.

4. Applications of NLP in urban research

Using NLP has the advantages of improving the usability of urban big data sources, expanding study areas and scales, and reducing research costs. In this section, the opportunities shown in the current applications of NLP are discussed in five areas: urban governance and management, public health, land use and functional zones, mobility, and urban design.

4.1. Urban governance and management

NLP adds new opportunities to citizen engagement, which is the most dominant topic among studies in urban governance challenges (Cruz et al., 2019). NLP techniques combined with online crowd-sourced data opens up a communication channel between city managers and the general public. From 2001 to 2004, the Electronic Democracy European Network (EDEN) project launched a real-life pilot to test if a particular NLP approach could improve communication between citizens and public administrators (European Commission, 2015). Though the EDEN project run into multiple obstacles, the project managers and engineers concluded that “it seems reasonable to approach e-democracy by seeking a democratic approach to software solutions” (Carenini, Whyte, Bertorello and Vanocchi, 2007, p. 27). Computer scientists also developed NLP applications to function as a citizen feedback gathering tool (Estévez-Ortiz et al., 2016), a citizen concern detector (Abali et al., 2018), and an urban community identifier (Vargas-Calderón and Camargo, 2019), and all showed promising results. In addition, combining NLP with interviews and focus groups, Bardhan et al. (2019) discovered gender inequality in Indian slum rehabilitation housing management, which suggested a need for a more systematic participatory approach to improve well-being among the rehabilitated occupants.

Additionally, NLP shows potential to support natural disaster responses. According to the US Congress's think tank, there are two ways that government agencies could use social media in emergency and disaster management: 1) as an outlet for information dissemination, and 2) as a systematic tool for emergency communication, victim assistance, situation monitoring, and damage estimation (Lindsay, 2011). The second category is where NLP has a direct role. An early work by Imran et al. (2013) trained a model that extracts disaster-relevant information from tweets and achieved 40%–80% correctness. More recently, Hong et al. (2018) built an unsupervised NLP topic model that requires minimal human efforts in text collecting and analyzing, which could help government agencies to identify citizens' needs and prioritize tasks during natural disasters. Additionally, with the integration of NLP and geospatial clustering methods, Hu et al. (2019a, b) collected local place names from housing advertisements, which has implications in disaster response, because these names may not exist in official gazetteers and could lead to miscommunication between local residents and disaster responders.

Furthermore, researchers have completed proof-of-concept studies for the method of using NLP, machine learning, and spatial analysis to spot urban crime. In Brazil, Souza et al. (2016) trained a classification model by emergency phone call records from the state police department and their model was able to analyze real-time tweets for crime detection. In the US, Helderop et al. (2019) were successful in detecting prostitution activities by examining hotel reviews, locations, and prices. Though the generalizability of the methods used in these studies needs further verification, with future improvements, they could eventually contribute to improving urban security.

Finally, scholars demonstrated the power of NLP in the research of construction management. Lai and Kontokosta (2019) conducted an exploratory study analyzing building permit records to uncover building renovation and adaptive reuse patterns in seven major cities in the US. The method they developed may benefit the monitoring of building alterations in urban areas.

4.2. Public health

Urban public health has drawn growing attention among researchers in recent years. Studies have revealed that various economic, social, and environmental factors, including the spread of infectious diseases, poor living conditions, unhealthy lifestyles, and pollution, could negatively affect public health in urban areas (Moscato and Poscia, 2015).

NLP is essential to large-scale application of social media as sensors to predict epidemic outbreaks. Traditional epidemic monitors rely on clinical reports gathered by public health authorities (Vaughan et al., 1989). For instance, health care providers in the US depend on the information provided by the Centers for Disease Control and Prevention (CDC) to learn about disease outbreaks (CDC, 2018). However, the time lag between the date that a disease starts and the date that clinical cases are reported to authorities is a major drawback of official surveillance systems (CDC, 2018). For this reason, many researchers have developed data processing and modeling techniques to use social media as a data source to conduct real-time epidemic analysis (Al-garadi et al., 2016). Though manually filtering and classifying relevant messages eliminates false positive and negative errors, the tradeoff is a slow analysis process (Nagar et al., 2014). NLP classification, on the other hand, can process data relatively fast with reasonable accuracy, supporting early detection of a disease. For example, in a Japanese nationwide study, Wakamiya et al. (2018) used an NLP module to effectively estimate when and where influenza outbreaks were happening.

In a similar sense, researchers view social media users as soft sensors to measure urban air quality. Riga and Karatzas (2014) adopted an NLP bag-of-words model to process social media posts and concluded that users' reports of their surrounding environmental conditions on social media platforms are highly correlated with the actual observations obtained from official monitoring sites.

Analyzing urban residents' perceptions of living environments and evaluating urban communities' lifestyles is another sphere of public health research in which NLP appears to be useful. Hu et al. (2019a) used NLP to process online neighborhood reviews to assess New Yorkers' satisfaction with their living conditions and their perceived quality of life. Also using NLP, Fu et al. (2018) derived urban citizens' activities through their linguistic patterns. Additionally, Rahimi, Mottahedi, and Liu (2018) were able to examine different communities' food consumption behaviors and lifestyles in ten major cities in North America by a bag-of-words model. Findings from these studies could serve as a valuable reference for city policymakers as they provide multifaceted health-related information complementary to conventional Census.

4.3. Land use and functional zones

Looking back into the history of urban planning, collecting information concerning land use functions is a critical step before laying out urban plans (Breheny and Batey, 1981). Traditional approaches to examine structures and changes in urban land use include analyzing aerial photographs (Philipson, 1997), field survey (Pissourios, 2019), and remote sensing (Bowden, 1975).

More recently, researchers have extended the usage of NLP from analyzing textual data to non-textual data, and applied it to urban land use and functionality studies. NLP typically detects underlying correlations between words according to their context. To capture urban spatial structures, researchers consider a region as a text document, a function as a topic, and research entities as words (Li, Fei and Zhang, 2019; Yuan et al., 2015). In this way, various NLP modeling methods allow researchers to determine contextual relationships between urban functional regions or different land use types based on the similarities among entities (i.e. geographic space interactions). This method makes use of urban data generated by sensors, vehicle geolocation tracking systems, and location-based services.

Yuan et al. (2015) explained the concept of mobility semantics by arguing that people's socioeconomic activities in a region are strongly correlated with the spatiotemporal patterns of those who visit the region (i.e. mobility semantics). Another key concept, location semantics, refers to urban road networks and the allocation of POIs (Yuan et al., 2015). By leveraging mobility and location semantics, Yuan et al. (2012, 2015) identified urban functional zones (e.g. residential, business, and educational areas) through topic modeling. Similarly, Yao et al. (2017) classified urban land use at the level of irregular land parcels by integrating a semantic model and deep learning. Huang et al. (2018) quantified industrial land use changes in a bay area in China using POIs data. Based on a Word2Vec model, Li et al. (2019) proposed a regionalization method to cluster similar spatial units in an area and inspect the clusters' socioeconomic patterns by analyzing all mobility trajectories of people in that area.

Demonstrating the advantages of being efficient and capable of handling large volumes of data, these researchers' approaches of NLP modeling to classify land use and functional zones show potential as great tools to monitor urban landscape dynamics and provide calibrations and for urban planning.

4.4. Mobility

Urban mobility researchers have begun to leverage NLP in their studies as well. Serna, Gerrikagoitia, Bernabé, and Ruiz (2017) demonstrated the feasibility of using NLP to automatically identify sustainable mobility issues from social media data, which could enrich the data of traditional travel surveys. Markou, Kaiser, and Pereira (2019) were able to predict taxi demand hotspots for special events by a tool they developed that scans the internet for time-series data.

Similar to the previously explained usage of NLP in land use and functional zones, researchers also adopted NLP for non-textual data analysis on urban mobility. By analyzing taxi moving paths recorded by the Global Navigation Satellite System, researchers measured spatio-temporal relationships among roads (Liu et al., 2017) and identified the interaction pattern of vehicle movements on road networks (Liu et al., 2019), which, they argued, could be useful in understanding and managing urban traffic.

4.5. Urban design

NLP can facilitate urban design with imageability analysis. “Urban design is the process of understanding people and place in an urban context, leading to a strategy for the improvement of urban life and the evolution of the built environment...” (Building Design Partnership, 1991, p. 14). Introduced by Lynch (1960), imageability is an important concept in urban design that is still being discussed today. It involves subjective urban identity, emphasizing the quality of the built environment perceived and assessed by observers (Lynch, 1960). NLP enables researchers to evaluate the emotional responses evoked by urban places and visually map urban identity at various scales and times.

Researchers have already used NLP to process hashtags from Instagram photos; using this information together with photos' geolocations, they created a cognitive map of the Seoul metropolitan area in Korea to represent its residents' collective perceptions of the place (Jang and Kim, 2019). To explore the emotional dynamics of urban space, Iaconesi (2015) used NLP to connect geographical locations within a city and emotions expressed in social media, through which established urban emotional landmarks. The observation of urban identity and emotional landmarks helps urban designers and planners make interventions and shape urban spaces into more positive and imageable places.

5. NLP challenges in urban research

Ultimately, a method or technique alone will never solve any problem. NLP opens up an exciting new direction, but at the same time it brings about more challenges. In this section, four aspects of potential challenges that apply to urban research are discussed: research questions, data, the method itself, and researchers.

The challenge of research questions lies in identifying novel issues that could not be well solved by traditional techniques. NLP holds great promise for the quest to untangle the complex relationships among urban systems, however, what questions it enables researchers to answer are still waiting to be explored. In fact, the study of cities involves a variety of disciplines in urban contexts (Ramadier, 2004). The existing urban studies have complemented or, sometimes, completely replaced traditional methods with NLP to solve problems in various urban-related fields. While conventional text analysis methods have high accuracy, NLP has advantages in dealing with a massive amount of data at a large scale with fine resolution. In a reality of limited time and resources, NLP could provide insight into questions that are impossible to answer with traditional methods. Looking ahead, more urban studies using NLP to answer questions that could not be otherwise answered will sure to emerge.

The data challenge for NLP goes hand-in-hand with the characteristics of urban big data. As pointed out by Salganik (2018), big data's characteristics of incompleteness, inaccessibility, and non-representativeness are generally problematic for academic research. Data incompleteness refers to the fact that no matter the size, urban big data are not purposeful designed structural data and are very likely to miss some valuable information to research such as demographic factors. In addition, inaccessibility means that data owned by private companies or government agencies are not always accessible to urban researchers due to legal or ethical barriers. Moreover, big data usually could not represent a certain urban population. As a result, studies that use NLP to process big data are not likely to yield generalizable results and face the risk of overlooking certain populations.

Though there have been revolutionary advances in NLP, its mainstream application is still very limited (Hirschberg and Manning, 2015). While the goal of NLP is that algorithms will ultimately be able to determine the relationships between words and grammar in human language and organize meaning by computer logic, the current techniques do not have the exact same capabilities of resolving natural language as humans do. NLP still needs improvements in “deal [ing] with irony, humor and other linguistic, psychological, anthropologic and cultural issues” (Iaconesi, 2015, p. 16), which is a difficult task for human analysts as well. Most recently, significant progress has been made in the field of NLP with increasing ease of implementing pre-trained models such as ULMFiT (Howard and Ruder, 2018) and BERT (Devlin et al., 2019). While this may trigger more adoption of NLP among researchers, it is important to validate NLP analysis with results reached by traditional methods.

Furthermore, people who study cities usually do not have professional training or a background in computer science. As a result, the complexity of detecting patterns, fitting models, and training classifiers limits urban researchers’ ability to take full advantage of NLP. This further hinders transferring knowledge into practice for urban planners and designers. In order to harness the new opportunities offered by NLP, urban researchers face the challenge to expand their skill set. On the other hand, computer scientists who wish to conduct urban research face the challenge to comprehend sophisticated social concepts and theories. Robust collaboration among researchers in different fields is likely to drive NLP applications in the study of cities.

While some of these challenges are common to studies using NLP in all fields, some others are more prominent for urban studies. Almost every researcher adopting NLP faces the challenges of acquiring good data and immature NLP techniques. For example, addressing implicit bias is still a daunting task when building NLP models. The success of NLP application in urban studies and other domains depends highly on the quality of data and modeling. Asking good research questions and skill requirements are more specific challenges facing urban researchers who intend to use NLP to facilitate their work. The spatial aspect of urban studies further compounds the challenge of adopting NLP. For instance, while NLP is effective in harvesting location data in texts, urban researchers need to be mindful of the massiveness and messiness of such data and assess the accuracy of uncovered geospatial information.

6. Conclusion

This systematic literature review suggests that there have been only a limited number of urban studies that adopted the approach of NLP. Current applications fell into five areas of study: urban governance and management, public health, land use and functional zones, mobility, and urban design. Using NLP in urban research demonstrates the advantages of improving the usability of urban big data sources, expanding study areas and scales, and reducing research costs. While recognizing this new opportunity is exciting, it is important for urban researchers not to overestimate what NLP is capable of accomplishing and acknowledge its limitations. To take advantage of NLP, urban researchers face challenges of raising good research questions, overcoming data incompleteness, inaccessibility, and non-representativeness, immature NLP techniques, and computational skill requirements.

Declarations

Author contribution statement

M. Cai: Developed and the wrote this article.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability statement

Data included in article/supplementary material/referenced in article.

Declaration of interests statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

Acknowledgements

I am very grateful to Dr. Mark Wilson, Dr. Jason Zhao, and the two anonymous reviewers for their insightful suggestions and comments on this paper.

References

Abali, G., Karaarslan, E., Hurriyetoglu, A., Dalkilic, F., 2018. Detecting citizen problems and their locations using twitter data. In: 2018 6th International Istanbul Smart Grids and Cities Congress and Fair (ICSG), 30–33.

Al-garadi, M.A., Khan, M.S., Varathan, K.D., Mujtaba, G., Al-Kabsi, A.M., 2016. Using online social networks to track a pandemic: a systematic review. J. Biomed. Inf. 62, 1–11.

Bardhan, R., Sunikka-Blank, M., Haque, A.N., 2019. Sentiment analysis as tool for gender mainstreaming in slum rehabilitation housing management in Mumbai, India. Habitat Int. 92, 102040.

Bowden, L.W., 1975. Urban environments: Inventory and analysis. In: Manual of Remote Sensing, 12. American Society of Photogrammetry, pp. 1815–1880.

Breheny, M.J., Batey, P.W.J., 1981. The history of planning methodology: a preliminary sketch. Built. Environ. (1978) 7 (2), 109–120. JSTOR.

Britton, B.K., 1978. Lexical ambiguity of words used in English text. Behav. Res. Methods Instrum. 10 (1), 1–7.

Building Design Partnership, 1991. Urban design in practice. Urban Design Quarterly 40.

Carenini, M., Whyte, A., Bertorello, L., Vanocchi, M., 2007. Improving communication in E-democracy using natural language processing. IEEE Intell. Syst. 22 (1), 20–27.

CDC, 2018. November 16). *Interpretation Of Epidemic (Epi) Curves During Ongoing Outbreak Investigations*. Centers for Disease Control and Prevention. https://www.cdc.gov/foodsafety/outbreaks/investigating-outbreaks/epi-curves.html.

Cruz, N. F. da, Rode, P., McQuarrie, M., 2019. New urban governance: a review of current themes and future priorities. J. Urban Aff. 41 (1), 1–19.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171–4186.

Estévez-Ortiz, F.-J., García-Jiménez, A., Glösekötter, P., 2016. An application of people's sentiment from social media to smart cities. El Prof. Inf. 25 (6), 851.

European Commission, 2015 June 13. Electronic Democracy European Network | EDEN Project. CORDIS | European Commission. https://cordis.europa.eu/project/rcn/57135/factsheet/en.

Fu, C., McKenzie, G., Frias-Martinez, V., Stewart, K., 2018. Identifying spatiotemporal urban activities through linguistic signatures. Comput. Environ. Urban Syst. 72, 25–37.

Gandomi, A., Haider, M., 2015. Beyond the hype: big data concepts, methods, and analytics. Int. J. Inf. Manag. 35 (2), 137–144.

Ghosh, S., Gunning, D., 2019. Natural Language Processing Fundamentals: Build Intelligent Applications that Can Interpret the Human Language to Deliver Impactful Results. Packt Publishing Ltd.

Guetterman, T.C., Chang, T., DeJonckheere, M., Basu, T., Scruggs, E., Vydiswaran, V.V., 2018. Augmenting qualitative text analysis with natural language processing: methodological study. J. Med. Internet Res. 20 (6).

Helderop, E., Huff, J., Morstatter, F., Grubesic, A., Wallace, D., 2019. Hidden in plain sight: a machine learning approach for detecting prostitution activity in phoenix, Arizona. Appl. Spat. Analy. Pol. 12 (4), 941–963.

Hey, T., Tansley, S., Tolle, K., 2009. The Fourth Paradigm: Data-Intensive Scientific Discovery. https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/.

Hirschberg, J., Manning, C.D., 2015. Advances in natural language processing. Science 349 (6245), 261–266.

Hong, L., Fu, C., Wu, J., Frias-Martinez, V., 2018. Information needs and communication gaps between citizens and local governments online during natural disasters. Inf. Syst. Front. New York 20 (5), 1027–1039.

Howard, J., Ruder, S., 2018. Universal language model fine-tuning for text classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 328–339.

Hu, Y., Deng, C., Zhou, Z., 2019a. A semantic and sentiment analysis on online neighborhood reviews for understanding the perceptions of people toward their living environments. Ann. Assoc. Am. Geogr. 109 (4), 1052–1073.

Hu, Y., Mao, H., McKenzie, G., 2019b. A natural language processing and geospatial clustering framework for harvesting local place names from geotagged housing advertisements. Int. J. Geogr. Inf. Sci. 33 (4), 714–738.

Huang, L., Wu, Y., Zheng, Q., Zheng, Q., Zheng, X., Gan, M., Wang, K., Shahtahmassebi, A., Deng, J., Wang, J., Zhang, J., 2018. Quantifying the spatiotemporal dynamics of industrial land uses through mining free access social datasets in the mega hangzhou bay region, China. Sustainability 10 (10), 3463.

Iaconesi, S., 2015. Emotional landmarks in cities. Sociologica 9 (3), 22.

Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., Meier, P., 2013. Practical extraction of disaster-relevant information from social media. In: Proceedings of the 22nd International Conference on World Wide Web - WWW '13 Companion, 1021–1024.

Jang, K.M., Kim, Y., 2019. Crowd-sourced cognitive mapping: a new way of displaying people's cognitive perception of urban space. PloS One 14 (6).

Lai, Y., Kontokosta, C.E., 2019. Topic modeling to discover the thematic structure and spatial-temporal patterns of building renovation and adaptive reuse in cities. Comput. Environ. Urban Syst. 78, 101383.

Li, Y., Fei, T., Zhang, F., 2019. A regionalization method for clustering and partitioning based on trajectories from NLP perspective. Int. J. Geogr. Inf. Sci. 33 (12), 2385–2405.

Lindsay, B.R., 2011. Social Media and Disasters: Current Uses, Future Options, and Policy Considerations, p. 13.

Liu, K., Gao, S., Lu, F., 2019. Identifying spatial interaction patterns of vehicle movements on urban road networks by topic modelling. Comput. Environ. Urban Syst. 74, 50–61.

Liu, K., Gao, S., Qiu, P., Liu, X., Yan, B., Lu, F., 2017. Road2Vec: measuring traffic interactions in urban road system from massive travel routes. ISPRS Int. J. Geo Inf. 6 (11), 321.

Lynch, K., 1960. The Image of the City. MIT Press.

Markou, I., Kaiser, K., Pereira, F.C., 2019. Predicting taxi demand hotspots using automated Internet Search Queries. Transport. Res. C Emerg. Technol. 102, 73–86.

Moscato, U., Poscia, A., 2015. Urban public health. In: Boccia, S., Villari, P., Ricciardi, W. (Eds.), A Systematic Review of Key Issues in Public Health. Springer International Publishing, pp. 223–247.

Nagar, R., Yuan, Q., Freifeld, C.C., Santillana, M., Nojima, A., Chunara, R., Brownstein, J.S., 2014. A case study of the New York city 2012-2013 influenza season with daily geocoded twitter data from temporal and spatiotemporal perspectives. J. Med. Internet Res. 16 (10).

Philipson, W.R., 1997. Urban analysis and planning. In: Manual of Photographic Interpretation, 2. American Society of Photogrammetry and Remote Sensing, pp. 517–554.

Pissourios, I.A., 2019. Survey methodologies of urban land uses: an oddment of the past, or a gap in contemporary planning theory? Land Use Pol. 83, 403–411.

Rahimi, S., Mottahedi, S., Liu, X., 2018. The geography of taste: using Yelp to study urban culture. ISPRS Int. J. Geo Inf. Basel 7 (9).

Ramadier, T., 2004. Transdisciplinarity and its challenges: the case of urban studies. Futures 36 (4), 423–439.

Riga, M., Karatzas, K., 2014. Investigating the relationship between social media content and real-time observations for urban air quality and public health. In: Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14) - WIMS '14, 1–7.

Salganik, M., 2018. *Bit By Bit: Social Research in the Digital Age* (Open Review Edition). Princeton University Press. https://www.bitbybitbook.com/en/preface/.

Serna, A., Gerrikagoitia, J.K., Bernabé, U., Ruiz, T., 2017. Sustainability analysis on urban mobility based on social media content. Transp. Res. Proc. 24, 1–8.

Souza, A., Figueredo, M., Cacho, N., Araujo, D., Coelho, J., Prolo, C.A., 2016. Social smart city: a platform to analyze social streams in smart city initiatives. In: 2016 IEEE International Smart Cities Conference (ISC2), 1–6.

Urban Land Institute, 2019. Urban Technology Framework. https://ulidigitalmarketing.blob.core.windows.net/ulidcnc/2019/05/ULI-Urban-Technology-Framework-2019.pdf.

Vargas-Calderón, V., Camargo, J.E., 2019. Characterization of citizens using word2vec and latent topic analysis in a large set of tweets. Cities 92, 187–196.

Vaughan, J.P., Morrow, R.H., Organization, W.H., 1989. Manual of Epidemiology for District Health Management. World Health Organization. http://apps.who.int/iris/handle/10665/37032.

Wakamiya, S., Kawai, Y., Aramaki, E., 2018. Twitter-based influenza detection after flu peak via tweets with indirect information: text mining study. JMIR Publ. Health Surv. 4 (3), e65.

Yao, Y., Li, X., Liu, X., Liu, P., Liang, Z., Zhang, J., Mai, K., 2017. Sensing spatial distribution of urban land use by integrating points-of-interest and Google Word2Vec model. Int. J. Geogr. Inf. Sci. 31 (4), 825–848.

Yuan, J., Zheng, Y., Xie, X., 2012. Discovering regions of different functions in a city using human mobility and POIs. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD, 12, p. 186.

Yuan, N.J., Zheng, Y., Xie, X., Wang, Y., Zheng, K., Xiong, H., 2015. Discovering urban functional zones using latent activity trajectories. IEEE Trans. Knowl. Data Eng. 27 (3), 712–725.