The City as Text

Authors

Reades, Jonathan; Centre for Advanced Spatial Analysis (CASA), UCL, UK*

Hu, Yingjie; Department of Geography, University at Buffalo, USA

Tranos, Emmanouil; School of Geographical Sciences, University of Bristol, UK

Delmelle, Elizabeth; Department of City and Regional Planning, University of Pennsylvania, USA

Preface

Urban researchers now have access to vast amounts of textual data—from social media and news to planning documents and property listings. These textual data provide important information about the activities of people and organizations in urban environments. Meanwhile, recent advancements in computational tools, including large language models, have expanded our ability to analyze textual data. This article explores how these tools are reshaping the ways we analyse, understand, and theorise the city through text. By outlining key developments, applications, and challenges, it argues that text is no longer a 'fringe resource' but a central component in urban analytics with the potential to connect quantitative and qualitative researchers.

Introduction

Cities through structured big data

The study of cities through a quantitative lens has, for the most part, been a story of abstraction and systematisation. While secondary data have long been used to systematically apprehend — and, in particular, manage — territory, improvements over the past two decades in our ability to obtain and analyse urban data at scale [1,2] seem to hold out the hope of a 'science of cities' [3]. Indeed, Kitchin (2014) [4] connected the advent of 'big data' to paradigm shifts in the wider social sciences, signposting the clear potential for 'grounded data' *from* cities — here in a quantitative form — to produce new grounded theories *of* cities. Together with a parallel set of developments in digital platforms, computational power, and low-cost hardware we have seen a further acceleration in the already-existing process of big data production (e.g. Arribas-Bel & Reades, 2018 [5]).

Perhaps the most notable feature of these new forms of big data, however, is their 'accidental' nature [1]: they come from the 'data exhaust' [6] of everyday urban life. For example, telecommunications and smart card travel data (e.g., Long & Thill, 2015 [7]) differ from previous sources of urban data because they are not collected and organised by enumerators, nor is the data in any sense quality-controlled. However, these data are still fundamentally structured, and they *could*, resources permitting, have been harnessed at any point in the history of quantitative and computational social science [8].

This does not apply though to unstructured data — principally, text and images — which presented an array of challenges — including of noise and complexity — that meant their analysis 'at scale' was largely undertaken by specialist groups or, more likely, corporate researchers. Specifically for text, the

This is an informal preprint. The formal version is at: https://doi.org/10.1038/s44284-025-00314-x

hard work of grappling with *unstructured* textual data from cities for research purposes was usually allocated to qualitative researchers. However, recent advancements in computational tools for handling textual data, including large language models (LLMs), offer exciting new opportunities for studying cities through the lens of text [9].

Technological advancements that enable the large-scale analysis of text data offer a new way of understanding cities beyond numerical analysis — through the words available from web pages, newspapers, social media, government and planning documents, real estate advertisements, and many others. This previously untapped resource offers an important complement to the prevailing quantitative modes that currently dominate urban analytics and city science, and can complement qualitative approaches to comprehending cities.

This Perspective seeks to ground this opportunity in an introduction to the kinds of text and tools available to researchers, providing examples of the state-of-the-art in urban research while contextualising these applications in the broader framework within which this interest in textual data evolved. Such an approach cannot be exhaustive, but we have sought to critically — if briefly — examine these cases in a structured way, before turning to a sketching out of challenges and future directions.

The mainstreaming of text in urban research

While quantitative approaches to text in urban research are not new, the intersection of the availability of textual data 'at scale' with radical improvements over the past decade in the computational tools needed to analyse such data is an opportunity to transform our approaches to urban analysis. Indeed, it should be no surprise that Lazer et al.'s [10] updated conceptualisation of computational social science now "encompasses language, location and movement, networks, images, and video…" despite the majority of these sources being largely inaccessible to academe much more than a decade ago.

However, with some important exceptions, we would primarily attribute the initial interest in text and cities to the expansion in geo-tagged user-generated data accessible to researchers; and it did not take long for geographers to begin exploring how such social media enabled novel approaches to everything from the geography of beer [11] to disaster management [12]. Crampton et al. (2013) [13] , however, noted the need to look 'beyond the geotag' and Johnson et al. (2016) [14] showed how different forms of 'localness' manifested in geo-tagged content.

These kinds of reflections helped to foreground a more general critique, rooted in parallel critiques of 'big data' in general, of the essential (urban) biases of user-generated text and its perpetuation of the digital divide [15,16]. Hristova et al. (2016) [17] are amongst the few to make a virtue of this bias in their research by drawing on the socio-economic profile of Twitter and Foursquare users to map gentrification in London.

We are not suggesting that quantitative social scientists more widely had, until then, been ignorant of the potential of text: in addition to the above, 'science mapping' for policy and research purposes has long-engaged with text in sophisticated ways (e.g. Borner 2010 [18]), often with conceptual and interpretive input from geography (e.g. Skupin, 2004 [19]). Rather, our point is that working with a substantial corpus — that is, a collection of texts — using quantitative methods was just really, really *hard*: a supercomputer was required to train and apply the Self-Organising Maps algorithm to a keyword corpus drawn from two million abstracts [20].

The surge in practical applications of text can therefore be linked to two recent developments: first, the discovery of novel text-modelling techniques that were expensive to *train*, but cheap to *use*; and second, the ability to perform *useful* calculations using the models trained in this manner. NLP techniques have evolved from simple keyword and syntax matching in the early days to sophisticated neural networks that 'learn' statistical patterns in documents — for words and, more recently, parts of words — to create 'embeddings' that represent words as (very) long numerical vectors.

Modern language models, including LLMs, are therefore typically trained on enormous corpora that make them relatively costly to develop (to say nothing of the copyright issues they thereby raise); however, once trained, these models can be used repeatedly and relatively inexpensively. Crucially, the learned relationships can then be used for analytically useful calculations: Mikolov et al. [21], for instance, showed that word embeddings enable the following types of calculation:

King - Man + Woman ≈ Queen

The reality, of course, is rather more complicated [22], and we're not suggesting that a language model can understand Shakespeare or the latest national urban redevelopment policy, nor that it replaces the need for close reading and careful interpretation. But the release of ChatGPT in late 2022 showed just how far — and how quickly — the embedding approach has come in just a few years: unlike earlier models trained for specific tasks, 'foundation models' could not only perform a huge range of analytical tasks using a natural language interface, they could now *generate* new text in response to user input.

The potential *advantages* of text over 'traditional' sources of quantitative data are far from being limited to their tractability in LLMs. Remotely sensed imagery, Census surveys, and traffic counts, have all been widely used in urban analysis, but they generally provide *reach* rather than *richness*. Previously, when there was a need for *richness*, researchers would necessarily turn to qualitative analysis, which inevitably impacted *reach*. The integration of modern computational methods and big textual data has the potential to overcome this duality [23]**.** Text is also, in a sense, closer to being 'raw data' and this allows for more flexible research approaches and even for new questions to be asked of old data. For example, shifts in word usage or meaning that might only be visible in retrospect can still be identified and investigated later, whether that is a few years (e.g. Würschinger & McGillivray, 2024 [24]) or several centuries (e.g. Taylor & Gregory, 2022 [25]) later. To be clear, scale alone does not address — and never has — the substantive critique of critical data studies that we must pay close attention to *whose* speech is recorded, but it does move quantitative research closer to the 'raw' data upon which it has always relied.

Understanding Cities through Text

So whether 'born digital' [26] or digitised later, textual data from Web pages, newspapers, social media, government documents, housing advertisements, and many other sources open up new windows for studying cities at scale through practices of 'distant reading' [27]. Clearly, these variegated sources provide insight into the activities of people and organizations in urban environments and, critically, they potentially offer insights into the *why* questions (e.g., why people are happy or not happy about their living environments) that usually cannot be answered through remote sensing images or Census statistics. In this section, we seek to provide a sense of the studies that have already leveraged textual data to examine aspects of cities. This sample is not, and cannot be, exhaustive but it shows how textual data has an important role to play in advancing our understanding of cities. Readers interested in more systematic reviews may wish to read Fu (2024) [28] and Hu (2018) [29].

Sentiments toward urban places

The sentiments of people toward urban places was one of the earliest applications of modern text-mining methods. Clearly, understanding whether people perceive restaurants, parks, neighbourhoods, or even entire cities, positively or negatively — as well as why such sentiments exist — has applications in urban planning and city management [30,31,28]. While surveys and interviews have been traditionally used to study these aspects of urban life [32–34], they are often limited by the relatively small sample sizes and the costly data collection process. The availability of social media and online review corpora provides alternative means of understanding sentiments from large numbers of individuals in a timely manner.

Sentiment analysis seeks to automatically classify input text as, for example, positive, neutral, or negative. There are two main types of sentiment analysis methods: lexicon-based and Machine Learning-based (ML). The former uses a kind of dictionary of pre-defined sentiments, such as Affective Norms for English Words (ANEW), and does not require labelled training data. Considering the ways in which expressive norms differ between, for example, formal letters and social media highlights the difficulty of applying sentiment lexicons across domains [35]. ML methods can better reflect domain-level differences, but require labelled training data that can be hard to obtain. It is also possible to combine the lexicon and ML methods in hybrid approaches, e.g., by using a sentiment lexicon as one of the input features to a ML model [36].

A variety of studies have used textual data and lexicon-based sentiment analysis to understand cities. Hu et al. (2019) [37] used both lexicon-based and machine learning-based approaches in an analysis of online neighbourhood reviews to understand the perceptions of people toward their living environments. Zou et al. (2019) [38] studied sentiment disparities on Twitter around Hurricane Harvey. Huang et al. (2023) [39] also analyzed Twitter data to re-examine urban vitality and Jacobs' urban form theory. Fu et al. (2024) [40] leveraged ChatGPT and textual feedback data to examine the sentiments of citizens in response to a proposed local plan change in Hamilton City, New Zealand.

We caution, however, that the accuracy of sentiment analysis is often affected by 'technical' limitations such as failing to detect sarcasm. While recent LLMs have mitigated those limitations to some extent, they also introduce new issues such as biased results due to unbalanced training data. Nevertheless, sentiment analysis can still be a useful approach when augmented by other data sources and human feedback to improve accuracy.

Place names in cities

Place names, including the names of local restaurants, parks, streets, and neighbourhoods, reflect local cultures, histories, influential families, natural and manmade landmarks, are all forms of text embedded in the cityscape [2,41]. Place names have historically been studied in human geography via case-by-case examinations and qualitative approaches [42,43], but automated extraction and geolocation from larger corpora opens up new research avenues and extends existing ones to larger spatial and temporal extents.

Again, little of this is necessarily 'new', since geography researchers have long worked on methods to extract and disambiguate place names from texts more accurately [44]. The long-standing approach — which remains relevant because place names extracted from documents are still only character strings and do not magically gain point, line, or polygon features — is the gazetteer, which is a large, dictionary-like database of place names, place types, and their associated spatial footprints [45].

However, gazetteers are necessarily only as comprehensive as their creators, and Named Entity Recognition (NER) tools — such as the Stanford NER tool — offer a more flexible way to extract spatial features and relationships. Examples of the latter include ensembles of NER tools [46,47], spatial pattern-based place name disambiguation [48,49], and methods based on deep neural networks such as Bidirectional Encoder Representations from Transformers [50,51] as well as more recent LLM-based methods [52,53].

Applications of these techniques have explored the influence of local cultures and surrounding geographic features on place naming in cities [54]; the identification of colloquial place names which might be relevant in disaster contexts [55]; the "nearby" exaggeration of places in real estate in cities [56]; information diffusion among cities [57]; and the evolution of place names in historical archives [58].

Neighbourhood changes and housing

The emergence of location-based, user-generated, and social media data sources has been particularly promising for the study of neighbourhoods, since it is more timely than traditional census or administrative survey data, and paints a complementary picture of neighbourhood perception and marketing [59]. Initial research provided alternative indicators of the popularity of a neighbourhood

through counts of visitors Tweeting from a neighbourhood [60], changes in the number or type of businesses captured from web-scraped Yelp data [61], or 'check-ins' on social media sites like Foursquare [62]. These applications demonstrated the potential to detect socioeconomic change in a neighbourhood; and the contemporaneousness of the data held out the possibility of 'nowcasting' [61] changes *as* they were occurring, rather than waiting months or years for retrospective administrative data.

More recently, researchers have moved from merely measuring where activity is occurring to analysing the text so as to better apprehend the discourses surrounding neighbourhoods. For example, text analyses of Airbnb advertisements [63], Yelp restaurant reviews [64,65], and real estate advertisements [66,67], all highlight how neighbourhoods are marketed in ways that shape the flow of consumers, tourists, and residents [64]. This research shows differences in the language of marketing materials by neighbourhood racial composition, with a disproportionate mentioning of cultural consumption amenities, like restaurants, and walkability in primarily White neighbourhoods [68] or by White AirBnB hosts in Black neighbourhoods [63].

In addition to providing important contextual insight into the ways that language may shape or reflect neighbourhood perception and change, research has shown that including words describing properties and neighbourhoods improves the performance of house price models [69]. NLP enables additional 'features' — in the sense of property attributes — to be extracted from descriptive text, allowing connections to be made to the value attached to off-street parking, river views, or a large and well-landscaped garden [70]. More sophisticated models also allow us to group attributes and explore commonalities across models even where word choices differ [71], as well as to (crudely) distinguish between subjective and factual statements about a property listing [72]. When grounded in geography, these approaches show enormous promise in developing a more nuanced understanding of what people in different places and times value in a home.

Neighbourhood change can also be examined through the lens of the built environment. Housing renovations, demolitions, and construction all point to flows of capital into a neighbourhood, a key indicator of gentrification. Historically, obtaining this type of building information has been arduous, especially for multiple cities. Lai and Kontokosta (2019) [73] demonstrate the application of NLP and ML to read and classify construction and building permit documents to create a multi-city database of these types of projects. Similar efforts have used NLP, including LLMs, to read and analyze zoning data [74,75]. Brinkley and Stahmer (2021) [76] called for a more coherent planning framework after identifying that housing topics were missing from hundreds of planning documents in California by using NLP; and Brinkley and Wagner (2022) [77] further assessed environmental justice using these documents. As Fu (2024) [28] highlights, information extraction and summarisation of policy documents is the most common usage of NLP in urban planning research. These examples demonstrate the potential of NLP to fill in key data gaps for indicators that are often collected inconsistently across cities, and that may not be available in a spatial data format.

Institutions and economic activities in cities

Although all texts are, ultimately, written by individuals (*pace* the advocates of LLMs), there can be an institutional layer to authorship which is also of interest to urban researchers. Documents from businesses, government, and Third sector organisations — including from websites, briefings, and published accounts — provide opportunities to understand cities from institutional perspectives. Local policy documents, political speeches, and court proceedings carry an institutional weight that should make them a priority for interrogating power itself, and this is an area almost completely lacking from quantitative research (see D'ignazio & Klein, 2023 [78] for a notable exception). Thomas et al. (2024) [79] used NLP to 'read' court records in order to map and analyse evictions; their case study in the State of Washington, USA revealed substantive racial and gender disparities. Similarly, Gromis et al. (2022) [80] constructed a US-wide eviction database to assess the state likelihood for a household to experience eviction, though this too foregrounds issues of sample bias in what is, and is not, recorded (see, for example, issues raised by Nelson et al. 2021 [81] and Summers and Steil 2024 [82]).

All such work relies on knowing where to source the data, and Cai et al. (2023) [83] built and released a valuable database containing the websites of *all* 19,518 municipalities in the US to serve as a foundation for systematic textual analysis of administrative 'speech' and self-description. Other types of organisational self-description offer an equally valuable means of capturing the content and the micro-urban geographies of economic activities: Stich et al. (2023) [23] used an archive of company websites to infer detailed economic activities — and their evolution over time — within London's Shoreditch start-up area. These findings matched those from extensive, in-depth qualitative studies, and Occhini (2024) [84] employed a similar approach to map and model the clustering of digital economy companies across all of London. Adjacent disciplines, such as regional economics, have been making use of patents [85], newspapers [86], and websites [87] to develop geographies of innovation.

Beyond Text

Much of what we have described above in terms of how text is approached by computational models has parallels in spatial analytic frameworks using text as an analogy. We are not the first to note the potential for correspondences between cities and texts: from words to points-of-interest, and from paragraphs to neighbourhoods . This slippage was exploited by algorithms such as Place2Vec [88] and Loc2Vec [89] which learn 'place embeddings' as a novel route to categorising places 'by the company they keep'. Similar work has drawn on the predictive aspect of embeddings (given *this* word, what is the most probable *next* word in this sentence?) to improve trajectory prediction (given *this* origin, what is the most probable *next* location in this journey?) in urban environments (e.g. Woźniak & Szymański, 2021 [90] and Du et al., 2018 [91]). LLMs appear to perform particularly well in this regard, highlighting the extent to which the uses of text modelling methods are not limited to textual data *per se* and the potential tractability of phenomena that can be (re)conceptualized — or analogised — through text.

Challenges and Possible Future Directions

While we are profoundly excited by the potential range and scope of transformative applications for NLP and text in quantitative urban research, this potential does not come without risks. Because so much of NLP now builds on neural-network architectures, questions about the processing and training of data, its suitability and representativeness, and the underlying 'black box' nature of such approaches raise significant challenges for open, trustworthy, and reproducible research.

The challenge of training data speaks to the lack of appropriately labelled data to train text models for more specialised applications. Because textual data was not initially intended for urban research, it is often difficult to find relevant, 'gold standard' manually-annotated data; but without human feedback, the trained models are much more vulnerable to learning the 'wrong' lessons from the data. Urban researchers must thus ask: *how can we train text models effectively in the absence of labeled reference data (or the money to annotate our own)?*

One obvious route is to increase our ability to create labeled datasets more efficiently and, for example, Sun et al., (2025) [92] developed a data annotation tool to help annotators label location descriptions in text much more efficiently. Another approach is to use weakly-supervised learning methods such that, instead of labelling the entire training data set, researchers initialise classes with seed words which help the model to iteratively learn synonyms and refine classes [93,94]. So-called transfer learning might then help these models to be redeployed across urban contexts; however, this again raises thorny issues of context and appropriateness that have no ready answer.

The challenge of circularity highlights the fact that, while it is well-understood that research is rarely linear, there is nonetheless a clear progression from inputs to outputs with the distinction between them being fairly obvious. Generative AI and LLMs have erased this boundary as the outputs of one model can then be re-used as the input to another [95]. The errors and biases of LLMs can thereby propagate in unpredictable ways; worse, the automation of 'user-generated content' through LLMs and chatbots further muddies the waters as to what, exactly, we are even studying.

This uncertainty prompts the following research question: *can we recognise the fingerprints of LLMs in user feedback and control for the kinds of bias they might introduce to urban analysis undertaken with LLMs?* The answer to such a question is clearly likely to depend on the nature of a particular study, but future research will certainly need to explore how to detect text generated by LLMs [96]. These issues are particularly germain where questions of policy-formation or popular preference are in play in contested urban contexts.

The challenge of noise stresses the complex effects of scale in textual analysis: traditional topic modelling seeks to categorise documents on the basis of word co-occurrence, but some — or, indeed, many — of these occurrences are simply a function of data volumes. As big data enthusiasts well know, given enough samples almost any relationship can appear statistically significant! Again, LLMs are typically more robust to this kind of noise because of the way they are trained, but that does not mean that they are immune. This prompts research on the question of *how to ensure that attention in text models focuses on the elements of a document that are important to the study and mitigates the impact of noise?*

Real estate listings, for instance, are surprisingly complex documents because they are highly compressed both literally — in the sense of making intensive use of abbreviations and other standardised forms that mean a good deal more than they say — and figuratively — in the coded use of language to signal particular features (or their absence) to potential buyers or renters. Moreover, the signal may well vary geographically: the adjective 'cozy', for instance, carries different valences for detached properties in a scenic area and urban rental units. Future research will need to explore how human feedback can augment the model training process (perhaps in conjunction with other cues such as the metadata or tags linked to online documents) to help text models focus on, and distinguish between, the things that we know are important.

This brings us to **the challenge of no-human-in-the-loop** in the use of large, pre-trained models to perform a variety of research tasks cheaply. LLMs are enormously promising for annotating and classifying large amounts of textual data without the need for expensive labeled data (zero-shot models). Recent work suggests that LLMs can almost match results obtained using experienced annotators [97], and Park et al. (2023) [98] showed that they can even help to generate plausible behaviours for Agent Based Models that interact with one another using natural language and use 'memories' of past choices and preferences to 'plan' future activities.

These are clearly potentially 'game changing' applications of LLMs, but they beg the question: *how can we even begin to understand and mitigate the risks of no-human-in-the-loop data analysis?* We should seek to preserve the uncanniness of these outputs lest we develop a false sense of security in the validity of their 'findings'. Future research will need both to find effective ways to validate LLM-based data, code, and results so as to mitigate the very real risks (e.g. hallucinations), and to keep bringing humans back into the loop when it might seem quicker and easier to exclude them [99].

Finally, **the challenge of theory** points to the fact that many applications of text-mining to urban problems lack an engagement with theory. Do we let the text 'speak for itself' or do we need to contextualise textual modelling within an urban theoretical framework? Clearly, we think it's the latter, not only because of the well-established social science arguments about the importance of theory (e.g. Kitchin, 2014) [4], but also because of the myriad challenges detailed above.

The kinds of ethically and statistically inappropriate, theory-free applications proposed for structured 'big data' do not disappear with the use of 'big text', they are likely to be much worse. We therefore ask: *how can we best leverage and advance theory in text-based urban studies?* To this end, future research will need to explore how best to integrate computational and qualitative textual methods, perhaps using the former to apprehend the outlines of a phenomena 'at scale' and the latter to dig deeper at sites selected for their representativeness or divergence from 'the norm'. We hope such integration might enable significant advances in urban theory without appearing to imply the death of close reading.

Conclusion

This Perspective was motivated not by the idea that quantitative researchers had been entirely ignorant of the value of text or existence of large corpora, but by the fact that many of the preexisting barriers to their widespread use have fallen away under the impact of, first, a trickle of techniques such as topic modelling and named-entity recognition and, now, the deluge of LLMs. It was not our purpose to provide a systematic review of 'urban text' or NLP methods, rather we wished to share our excitement about the emerging opportunities while stressing the ongoing importance of domain knowledge grounded in theory when researchers might be tempted to 'see what the (textual) data say'.

We have explored both the types and, importantly, the attributes of textual data for cities, and discussed the applications—what we can learn about cities through text—and outstanding conceptual challenges in dealing with such data. We juxtaposed this strand of empirical urban research to well-rooted epistemological discussions about the richness of qualitative studies vis-à-vis the reach of quantitative research based on more traditional data sources.

What became clear is that analysing text at scale to answer urban questions stretches disciplinary boundaries. Urban researchers have always been active interpreters and 'choosers' of data, but the complex source and nature of urban textual archives foregrounds the importance of this consideration. As (primarily) geographers and urban planners in interdisciplinary research groups, we are well-versed in the debates surrounding extensive and intensive methods, and the combining of sources of data traditionally used in one discipline with methods traditionally used in another; however, we would argue that this dynamic stresses the subjectivity of *all* data analysis in ways that should empower researchers to learn from one another.

Looking forward, text analytics is fast-becoming a staple in our urban methodological toolkit, and AI developments will reinforce this trajectory by enhancing our capacity to extract new (urban) knowledge from textual data. At the same time, generative AI has already become disruptive and, if one thing is certain, it is that urban researchers will not be able to address such challenges on their own. Instead cross-fertilisation with cognate disciplines is a necessity.

References

1. Arribas-Bel, D. Accidental, open and everywhere: Emerging data sources for the understanding of cities. *Appl. Geogr.* **49**, 45–53 (2014).
2. Oto-Peralías, D. What do street names tell us? The 'city-text' as socio-cultural data. *J. Econ. Geogr.* **18**, 187–211 (2018).
3. Batty, M. *The New Science of Cities*. (MIT press, 2017).
4. Kitchin, R. Big Data, new epistemologies and paradigm shifts. *Big Data Soc.* **1**, 2053951714528481 (2014).
5. Arribas-Bel, D. & Reades, J. Geography and computers: Past, present, and future. *Geogr. Compass* **12**, e12403 (2018).
6. Harford, T. Big data: are we making a big mistake? *Significance* **11**, 14–19 (2014).
7. Long, Y. & Thill, J.-C. Combining smart card data and household travel survey to analyze jobs–housing relationships in Beijing. *Comput. Environ. Urban Syst.* **53**, 19–35 (2015).
8. Lazer, D. *et al.* Computational social science. *Science* **323**, 721–723 (2009).
9. Lore, M., Harten, J. G. & Boeing, G. A hybrid deep learning method for identifying topics in large-scale urban text data: Benefits and trade-offs. *Comput. Environ. Urban Syst.* **111**, 102131 (2024).
10. Lazer, D. M. J. *et al.* Computational social science: Obstacles and opportunities. *Science* **369**, 1060–1062 (2020).
11. Zook, M. & Poorthuis, A. Offline Brews and Online Views: Exploring the Geography of Beer Tweets. in *The Geography of Beer* (eds. Patterson, M. & Hoalst-Pullen, N.) 201–209 (Springer Netherlands, Dordrecht, 2014). doi:10.1007/978-94-007-7787-3_17.

12. Crooks, A., Croitoru, A., Stefanidis, A. & Radzikowski, J. # Earthquake: Twitter as a distributed sensor system. *Trans. GIS* **17**, 124–147 (2013).
13. Crampton, J. W. *et al.* Beyond the geotag: situating 'big data' and leveraging the potential of the geoweb. *Cartogr. Geogr. Inf. Sci.* **40**, 130–139 (2013).
14. Johnson, I. L., Sengupta, S., Schöning, J. & Hecht, B. The Geography and Importance of Localness in Geotagged Social Media. in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* 515–526 (ACM, San Jose California USA, 2016). doi:10.1145/2858036.2858122.
15. Hecht, B. & Stephens, M. A tale of cities: Urban biases in volunteered geographic information. in *proceedings of the international AAAI conference on web and social media* vol. 8 197–205 (2014).
16. Wang, Z., Lam, N. S., Obradovich, N. & Ye, X. Are vulnerable communities digitally left behind in social responses to natural disasters? An evidence from Hurricane Sandy with Twitter data. *Appl. Geogr.* **108**, 1–8 (2019).
17. Hristova, D., Williams, M. J., Musolesi, M., Panzarasa, P. & Mascolo, C. Measuring urban social diversity using interconnected geo-social networks. in *Proceedings of the 25th international conference on world wide web* 21–30 (2016).
18. Borner, K. *Atlas of Science: Visualizing What We Know*. (Mit Press, 2010).
19. Skupin, A. The world of geography: Visualizing a knowledge domain with cartographic means. *Proc. Natl. Acad. Sci.* **101**, 5274–5278 (2004).
20. Boyack, K. W. *et al.* Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches. *PloS One* **6**, e18029 (2011).
21. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient estimation of word representations in vector space. in *Proceedings of the 1st International Conference on Learning Representations (ICLR)* 1–12 (Scottsdale, Arizona, USA, 2013).
22. Nissim, M., van Noord, R. & Van Der Goot, R. Fair is better than sensational: Man is to doctor as woman is to doctor. *Comput. Linguist.* **46**, 487–497 (2020).
23. Stich, C., Tranos, E. & Nathan, M. Modeling clusters from the ground up: A web data approach. *Environ. Plan. B Urban Anal. City Sci.* **50**, 244–267 (2023).
24. Würschinger, Q. & McGillivray, B. Semantic change and socio-semantic variation: the case of COVID-related neologisms on Reddit. *Linguist. Vanguard* (2024).
25. Taylor, J. E. & Gregory, I. N. *Deep Mapping the Literary Lake District: A Geographical Text Analysis*. (Rutgers University Press, 2022).
26. National Archives. Born-digital records and metadata. https://www.nationalarchives.gov.uk/information-management/manage-information/digital-records-transfer/what-are-born-digital-records/ (2024).
27. Moretti, F. *Distant Reading*. (Verso Books, 2013).
28. Fu, X. Natural Language Processing in Urban Planning: A Research Agenda. *J. Plan. Lit.* 08854122241229571 (2024) doi:10.1177/08854122241229571.
29. Hu, Y. Geo-text data and data-driven geospatial semantics. *Geogr. Compass* **12**, e12404 (2018).
30. Ahmed, K. B., Radenski, A., Bouhorma, M. & Ahmed, M. B. Sentiment analysis for smart cities: state of the art and opportunities. in *Proceedings on the international conference on internet computing (ICOMP)* 55 (The Steering Committee of The World Congress in Computer Science, Computer …, 2016).
31. Kovacs-Gyori, A., Ristea, A., Havas, C., Resch, B. & Cabrera-Barona, P. # London2012: Towards citizen-contributed urban planning through sentiment analysis of twitter data. *Urban Plan.* **3**, 75–99 (2018).
32. Ceccato, V. & Snickars, F. Objective and subjective indicators to evaluate quality of life (QOL) in two districts in the Stockholm region. in *Urban Ecology* 273–277 (Springer, 1998).
33. Das, D. Urban quality of life: A case study of Guwahati. *Soc. Indic. Res.* **88**, 297–310 (2008).
34. Eby, J., Kitchen, P. & Williams, A. Perceptions of quality life in Hamilton's neighbourhood hubs: A qualitative analysis. *Soc. Indic. Res.* **108**, 299–315 (2012).
35. Khoo, C. S. & Johnkhan, S. B. Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. *J. Inf. Sci.* **44**, 491–511 (2018).

36. Wankhade, M., Rao, A. C. S. & Kulkarni, C. A survey on sentiment analysis methods, applications, and challenges. *Artif. Intell. Rev.* **55**, 5731–5780 (2022).
37. Hu, Y., Deng, C. & Zhou, Z. A semantic and sentiment analysis on online neighborhood reviews for understanding the perceptions of people toward their living environments. *Ann. Am. Assoc. Geogr.* **109**, 1052–1073 (2019).
38. Zou, L. *et al.* Social and geographical disparities in Twitter use during Hurricane Harvey. *Int. J. Digit. Earth* **12**, 1300–1318 (2019).
39. Huang, J. *et al.* Re-examining Jane Jacobs' doctrine using new urban data in Hong Kong. *Environ. Plan. B Urban Anal. City Sci.* **50**, 76–93 (2023).
40. Fu, X., Sanchez, T. W., Li, C. & Reu Junqueira, J. Deciphering public voices in the digital era: benchmarking ChatGPT for analyzing citizen feedback in Hamilton, New Zealand. *J. Am. Plann. Assoc.* **90**, 728–741 (2024).
41. Azaryahu, M. Renaming the Past: Changes in" City Text" in Germany and Austria, 1945-1947. *Hist. Mem.* **2**, 32–53 (1990).
42. Zelinsky, W. Along the frontiers of name geography. *Prof. Geogr.* **49**, 465–466 (1997).
43. Rose-Redwood, R., Alderman, D. & Azaryahu, M. Geographies of toponymic inscription: new directions in critical place-name studies. *Prog. Hum. Geogr.* **34**, 453–470 (2010).
44. Purves, R. S., Clough, P., Jones, C. B., Hall, M. H. & Murdock, V. Geographic information retrieval: Progress and challenges in spatial search of text. *Found. Trends® Inf. Retr.* **12**, 164–318 (2018).
45. Goodchild, M. F. & Hill, L. L. Introduction to digital gazetteer research. *Int. J. Geogr. Inf. Sci.* **22**, 1039–1044 (2008).
46. Alex, B., Byrne, K., Grover, C. & Tobin, R. Adapting the Edinburgh geoparser for historical georeferencing. *Int. J. Humanit. Arts Comput.* **9**, 15–35 (2015).
47. Karimzadeh, M., Pezanowski, S., MacEachren, A. M. & Wallgrün, J. O. GeoTxt: A scalable geoparsing system for unstructured text geolocation. *Trans. GIS* **23**, 118–136 (2019).
48. DeLozier, G., Baldridge, J. & London, L. Gazetteer-independent toponym resolution using geographic word profiles. in *Twenty-Ninth AAAI conference on artificial intelligence* (2015).
49. Gritta, M., Pilehvar, M. T. & Collier, N. Which melbourne? augmenting geocoding with maps. in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 1285–1296 (2018).
50. Wang, J., Hu, Y. & Joseph, K. NeuroTPR: A neuro-net toponym recognition model for extracting locations from social media messages. *Trans. GIS* **24**, 719–735 (2020).
51. Zhou, B., Zou, L., Hu, Y., Qiang, Y. & Goldberg, D. TopoBERT: a plug and play toponym recognition module harnessing fine-tuned BERT. *Int. J. Digit. Earth* **16**, 3045–3063 (2023).
52. Hu, Y. *et al.* Geo-knowledge-guided GPT models improve the extraction of location descriptions from disaster-related social media messages. *Int. J. Geogr. Inf. Sci.* **37**, 2289–2318 (2023).
53. Hu, X., Kersten, J., Klan, F. & Farzana, S. M. Toponym resolution leveraging lightweight and open-source large language models and geo-knowledge. *Int. J. Geogr. Inf. Sci.* 1–28 (2024).
54. Hu, Y. & Janowicz, K. An Empirical Study on the Names of Points of Interest and Their Changes with Geographic Distance. in *Proceedings of the 10th International Conference on Geographic Information Science* (Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018).
55. Hu, Y., Mao, H. & McKenzie, G. A natural language processing and geospatial clustering framework for harvesting local place names from geotagged housing advertisements. *Int. J. Geogr. Inf. Sci.* **33**, 714–738 (2019).
56. McKenzie, G. & Hu, Y. The "Nearby" exaggeration in real estate. in *Proceedings of the Workshop on Cognitive Scales of Spatial Information* 1–4 (2017).
57. Peris, A., Meijers, E. & Van Ham, M. Information diffusion between Dutch cities: Revisiting Zipf and Pred using a computational social science approach. *Comput. Environ. Urban Syst.* **85**, 101565 (2021).
58. Southall, H., Mostern, R. & Berman, M. L. On historical gazetteers. *Int. J. Humanit. Arts Comput.* **5**, 127–145 (2011).
59. Delmelle, E. C. GIScience and neighborhood change: Toward an understanding of processes of change. *Trans. GIS* **26**, 567–584 (2022).
60. Chapple, K., Poorthuis, A., Zook, M. & Phillips, E. Monitoring streets through tweets: Using user-

generated geographic information to predict gentrification and displacement. *Environ. Plan. B Urban Anal. City Sci.* **49**, 704–721 (2022).
61. Glaeser, E. L., Kim, H. & Luca, M. Nowcasting gentrification: using yelp data to quantify neighborhood change. in *AEA Papers and Proceedings* vol. 108 77–82 (American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, 2018).
62. Zhou, X. & Zhang, L. Crowdsourcing functions of the living city from Twitter and Foursquare data. *Cartogr. Geogr. Inf. Sci.* **43**, 393–404 (2016).
63. Törnberg, P. & Chiappini, L. Selling black places on Airbnb: Colonial discourse and the marketing of black communities in New York City. *Environ. Plan. Econ. Space* **52**, 553–572 (2020).
64. Zukin, S., Lindeman, S. & Hurson, L. The omnivore's neighborhood? Online restaurant reviews, race, and gentrification. *J. Consum. Cult.* **17**, 459–479 (2017).
65. Olson, A. W., Calderón-Figueroa, F., Bidian, O., Silver, D. & Sanner, S. Reading the city through its neighbourhoods: Deep text embeddings of Yelp reviews as a basis for determining similarity and change. *Cities* **110**, 103045 (2021).
66. Delmelle, E. C. & Nilsson, I. The language of neighborhoods: A predictive-analytical framework based on property advertisement text and mortgage lending data. *Comput. Environ. Urban Syst.* **88**, 101658 (2021).
67. Kennedy, I., Hess, C., Paullada, A. & Chasins, S. Racialized discourse in Seattle rental ad texts. *Soc. Forces* **99**, 1432–1456 (2021).
68. Nilsson, I. & Delmelle, E. C. Smart growth as a luxury amenity? Exploring the relationship between the marketing of smart growth characteristics and neighborhood racial and income change. *J. Transp. Geogr.* **106**, 103522 (2023).
69. Zhang, H., Li, Y. & Branco, P. Describe the house and I will tell you the price: House price prediction with textual description data. *Nat. Lang. Eng.* 1–35 (2023).
70. Huang, Z. How languages used in property listing descriptions vary and affect its price geographically across the UK? (University College London, 2020).
71. Jiang, Y. Housing Price Prediction in London: A Predictive Analysis Based on Property Advertisement Texts. (University College London, 2022).
72. Wang, W. How do textual information and sentiment analysis improve house price estimation? (University College London, 2022).
73. Lai, Y. & Kontokosta, C. E. 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 (2019).
74. Mleczko, M. & Desmond, M. Using natural language processing to construct a National Zoning and Land Use Database. *Urban Stud.* **60**, 2564–2584 (2023).
75. Xu, W., Markley, S., Bronin, S. C. & Drogaris, D. A National Zoning Atlas to Inform Housing Research, Policy, and Public Participation. *Cityscape* **25**, 55–72 (2023).
76. Brinkley, C. & Stahmer, C. What Is in a Plan? Using Natural Language Processing to Read 461 California City General Plans. *J. Plan. Educ. Res.* **44**, 632–648 (2021).
77. Brinkley, C. & Wagner, J. Who Is Planning for Environmental Justice—and How? *J. Am. Plann. Assoc.* **90**, 63–76 (2022).
78. D'ignazio, C. & Klein, L. F. *Data Feminism*. (MIT press, 2023).
79. Thomas, T., Ramiller, A., Ren, C. & Toomet, O. Toward a National Eviction Data Collection Strategy Using Natural Language Processing. *Cityscape* **26**, 241–260 (2024).
80. Gromis, A. *et al.* Estimating eviction prevalence across the United States. *Proc. Natl. Acad. Sci.* **119**, e2116169119 (2022).
81. Nelson, K., Garboden, P., McCabe, B. J. & Rosen, E. Evictions: The Comparative Analysis Problem. *Hous. Policy Debate* **31**, 696–716 (2021).
82. Summers, N. & Steil, J. Pathways to Eviction. *Law Soc. Inq.* 1–41 (2024).
83. Cai, M., Huang, H. & Decaminada, T. Local data at a national scale: Introducing a dataset of official municipal websites in the United States for text-based analytics. *Environ. Plan. B Urban Anal. City Sci.* **50**, 1988–1993 (2023).
84. Occhini, G. Who, What and Where. (University of Bristol, 2024).
85. Arts, S., Hou, J. & Gomez, J. C. Natural language processing to identify the creation and impact of

new technologies in patent text: Code, data, and new measures. *Res. Policy* **50**, 104144 (2021).
86. Ozgun, B. & Broekel, T. The geography of innovation and technology news - An empirical study of the German news media. *Technol. Forecast. Soc. Change* **167**, 120692 (2021).
87. Axenbeck, J. & Breithaupt, P. Innovation indicators based on firm websites—Which website characteristics predict firm-level innovation activity? *PLOS ONE* **16**, e0249583 (2021).
88. Yan, B., Janowicz, K., Mai, G. & Gao, S. From ITDL to Place2Vec–Reasoning About Place Type Similarity and Relatedness by Learning Embeddings From Augmented Spatial Contexts. in *Proceedings of SIGSPATIAL* vol. 35 1–10 (2017).
89. Spruyt, V. Loc2vec: Learning location embeddings with triplet-loss networks. *Sentiance* https://sentiance.com/loc2vec-learning-location-embeddings-w-triplet-loss-networks (2018).
90. Woźniak, S. & Szymański, P. Hex2vec: Context-aware embedding h3 hexagons with openstreetmap tags. in *Proceedings of the 4th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery* 61–71 (2021).
91. Du, J., Chen, Y., Wang, Y. & Pu, J. Zone2vec: Distributed representation learning of urban zones. in *2018 24th International Conference on Pattern Recognition (ICPR)* 880–885 (IEEE, 2018).
92. Sun, K., Hu, Y., Joseph, K. & Zhou, R. Z. GALLOC: a GeoAnnotator for Labeling LOCation descriptions from disaster-related text messages. *Int. J. Geogr. Inf. Sci.* **39**, 1623–1653 (2025).
93. Mekala, D. & Shang, J. Contextualized weak supervision for text classification. in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* 323–333 (2020).
94. Occhini, G., Tranos, E. & Wolf, L. *Measuring a Country's Digital Industrial Structure: Commercial Websites and Weakly Supervised Classification to the Rescue*. (2023).
95. Singleton, A. D. & Spielman, S. Segmentation using large language models: A new typology of American neighborhoods. *EPJ Data Sci.* **13**, 1–21 (2024).
96. Wu, J. *et al.* A survey on LLM-generated text detection: Necessity, methods, and future directions. *Comput. Linguist.* 1–66 (2025).
97. Mellon, J. *et al.* Do AIs know what the most important issue is? Using language models to code open-text social survey responses at scale. *Res. Polit.* **11**, 20531680241231468 (2024).
98. Park, J. S. *et al.* Generative Agents: Interactive Simulacra of Human Behavior. in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* 1–22 (ACM, San Francisco CA USA, 2023). doi:10.1145/3586183.3606763.
99. Zheng, Z. & Sieber, R. Putting humans back in the loop of machine learning in Canadian smart cities. *Trans. GIS* **26**, 8–24 (2022).