Reading the city through its neighbourhoods: Deep text embeddings of Yelp reviews as a basis for determining similarity and change

Alexander W Olson^{a,d,c}, Fernando Calderón-Figueroa^{b,d}, Olimpia Bidian^{b,d}, and
Daniel Silver^{b,d}, Scott Sanner^{a,d}

^a*Department of Mechanical and Industrial Engineering*

^b*Department of Sociology*

^c*Corresponding Author—aolson@mie.utoronto.ca*

40 St George St, Toronto, ON M5S 2E4

^d*University of Toronto, Canada*

Abstract

This paper develops novel methods for using Yelp reviews as a window into the collective representations of a city and its neighbourhoods. Basing analysis on social media data such as Yelp is a challenging task because review data is highly sparse and direct analysis may fail to uncover hidden trends. To this end, we propose a deep autoencoder approach for embedding the language of neighbourhood-based business reviews into a reduced dimensional space that facilitates similarity comparison of neighbourhoods and their change over time. Our model improves performance in distinguishing real and fake neighbourhood descriptions derived from real reviews, increasing performance in the task from an average accuracy of 0.46 to 0.77. This improvement in performance indicates that this novel application of embedded language analysis permits us to uncover comparative trends in neighbourhood change through the lens of their venues' reviews, providing a computational methodology for reading a city through its neighbourhoods. The resulting toolkit makes it possible to examine a city's current sociological trends in terms of its neighbourhoods' collective identities.

Keywords: Computational Social Science, Urban Informatics, Neighbourhood Analysis, Machine Learning, Text Embedding

1. Introduction

Until recently, people learned about cities mostly through personal experience, conversations, travel guides, and other printed material in newspapers and
Preprint submitted to Cities

December 3, 2020

4 books. The advent of the Internet has tremendously expanded the number of
5 others from whom we can learn about what a city has to offer, and how easily we
6 can access this information. Indeed, we now routinely receive recommendations
7 from millions of people through consumer review websites such as Yelp.com.
8 These websites are designed to provide consumers with information about the
9 quality of various amenities they may wish to try, while also offering them a
10 chance to participate in an ongoing conversation about those amenities.

11 The millions of reviews accumulated in the process also provide researchers
12 with rich data for uncovering how people interpret and evaluate the context
13 around them — provided we can devise the tools necessary to translate this in-
14 formation into a form by which we too can ‘read’ [1, 2, 3] the city. In this paper,
15 we demonstrate via a case study of Toronto, Canada, that it is possible to charac-
16 terize neighbourhoods using Yelp review data and examine their relation to each
17 other. We propose novel computational methods for reading the city through
18 the massive corpus generated and compiled by users of these websites and show
19 that the resulting representation captures the evolving discourse by which users
20 represent to one another what various parts of a city mean to them. In so doing,
21 we combine concepts traditionally associated with qualitative and phenomeno-
22 logical approaches to urban life [4, 5, 6]—legibility, interpretation, discourse,
23 meaning—with novel computational methods that allow us to investigate them
24 systematically, comparatively, and in a way that is reproducible by others. By
25 no means do we believe that this “thin description” can or should replace the
26 traditional tools of “thick description.” However, we do suggest that the two can
27 work in tandem and that a significant contribution of this paper is to illustrate
28 that the qualities and meanings of urban life can be examined quantitatively.

29 To meet this goal, we introduce an embedding-based computational approach
30 to analyzing the discursive transformation of a city over time, chiefly at the
31 neighbourhood level. We work with a dataset comprising all Yelp reviews for the
32 city of Toronto over ten years. To draw a quantitative analysis of the change in
33 neighbourhoods from these reviews, we introduce an autoencoder-based method-
34 ology, which learns a compact embedded representation of review data from rel-
35 atively sparse word usage at the surface review level. We then go on to demon-
36 strate that this compact representation, when applied to aggregated reviews at the
37 neighbourhood level, contains critical information about each neighbourhood,
38 which can be used to investigate how neighbourhoods change over time. By
39 separating reviews by the year in which they were written, we employ our em-
40 bedding space to examine how neighbourhoods change relative to each other.
41 We identify a set of very stable neighbourhoods and use the Most Stable Neigh-

42 bourhood (MSN) to find categories of change.

43 These methodological innovations are the central contribution of the paper.
44 However, they point toward a significant substantive contribution as well: the
45 capacity to identify stable and dynamic neighbourhood collective identities us-
46 ing the actual textual artifacts of neighbourhood users that could be scaled up
47 more broadly. If economists have used similar data as a way to monitor eco-
48 nomic trends in a way that goes beyond what standard official indicators permit
49 (e.g. [7, 8, 9]), our methods offer techniques for using Yelp data to monitor soci-
50 ological trends in a large-scale way, even where sources such as survey data are
51 unavailable.

52 To our knowledge, no previous study on neighbourhood change has exploited
53 the richness of data offered by Yelp reviews — i.e., both the attributes of geolo-
54 cated venues and the underlying structure of review texts. Capitalizing on these
55 features, we aim to answer two questions:

- 56 1. Do Yelp review data allow us to identify neighbourhood boundaries using
57 text analysis and machine learning techniques?
- 58 2. By aggregating Yelp reviews by neighbourhood and year, is it possible to
59 identify neighbourhood change at the discursive level? If so, is it possible
60 to identify which neighbourhoods become more similar or more different
61 to one another over time?

62 Section 2 reviews prior attempts to “read” the city, existing research using
63 Yelp and other social media data, and relevant work on computational meth-
64 ods for large-scale latent text analysis. Sections 3 and 4 present our analytical
65 approach, methods, and results. We conclude with a discussion of the broader
66 significance of this research, its limits, and future directions.

67 **2. Literature Review**

68 *Reading the City’s Neighbourhoods*

69 The neighbourhood is one of the most distinctive features of a city. Neigh-
70 bourhoods are not only crucial for people’s cognitive maps and their sense of
71 identity, but they are increasingly central targets for planning and social pol-
72 icy [10]. Therefore, a proper account of the boundaries of a city’s neighbour-
73 hoods is not just a matter of scholarly interest but has significant social and politi-
74 cal consequences. Neighbourhood boundaries are generally difficult to establish.
75 Spatial features alone have proven to be ineffective in identifying boundaries
76 that local people can distinguish accurately [11, 12]. While official neighbour-
77 hood boundaries drawn by municipal authorities do not always reflect existing

78 communities [13], they are often used for policy, real estate, and marketing pur-
79 poses [14]. Official boundaries accordingly offer a useful baseline to test how
80 accurately ‘big data’ can identify neighbourhoods in a city.

81 Building on Suttles [6], a stream of urban research approaches the neighbour-
82 hood as a *creative imposition* that sits in the intersection between the physical
83 environment, both internal and external to its boundaries, and people’s interac-
84 tions on it. People’s activities in the city are closely connected to their ‘cognitive
85 maps’ — their internal representations of the space of the city. People develop
86 cognitive maps beyond their neighbourhoods of residence as they move around
87 the city either on their way to work or for other purposes [15]. Thus, the map
88 of a single neighbourhood is the result of an ongoing conversation among resi-
89 dents and external agents, such as occasional visitors, local authorities, and de-
90 velopers. As a result, the “common knowledge” about what a neighbourhood’s
91 identity consists of is a *collective representation* that transcends the interpreta-
92 tion of any single individual or group. This is a pressing issue in contempo-
93 rary neighbourhood scholarship. In her recent book, Talen [16] identifies several
94 crucial features of neighbourhood efficacy. One of these is a coherent shared
95 identity that residents and others can identify and rally around, as well as vari-
96 ous meeting places and venues that affirm and support this meaning. However,
97 a systematic reconstruction of cognitive maps, particularly the neighbourhood
98 boundaries within them, poses several methodological challenges. Some survey
99 and interview-based studies rely on mapping exercises administered to residents.
100 The data obtained is later processed to identify either the “consensus map,” the
101 average, or the largest resident-defined boundaries [17]. These techniques are
102 expensive and difficult to validate [18, 19].

103 Among the earliest attempts to draw cognitive maps from interview-based
104 mapping evidence, Lynch’s [5] work is one of the most influential to this day
105 (cf. [12, 20]). Lynch’s point of departure is that cities are legible—human be-
106 ings are capable of recognizing different parts of cities and organizing them
107 as coherent patterns. One type of pattern Lynch discusses is the district: each
108 two-dimensional area of the city whose boundaries are recognizable through a
109 “common, identifying character.” Although districts are not necessarily neigh-
110 bourhoods, they allow us to think about neighbourhoods beyond people and res-
111 idential infrastructure, and to pay attention to the configuration of businesses
112 and public places that can give a sense of coherence to a specific portion of the
113 city. Recently, Filomena and colleagues [12] tried to reproduce Lynch’s cogni-
114 tive map of Boston using GIScience techniques. To identify district boundaries,
115 they relied on modularity optimization to analyze the road layout of contempo-

116 rary Boston. Their findings were modest in relation to district identification due
117 to the lack of data about people’s activities in the space.

118 Studies that rely mostly on physical features to characterize neighbourhood
119 boundaries do not aptly capture the meanings attached to such features. Lynch’s [5]
120 metaphorical link between the city and language has been extended by more re-
121 cent studies that aim to read the city as text, which typically implies some form of
122 semiotic practice—i.e., the practice involved in meaning-making. For instance,
123 for Barthes [4] analyzing the discourse around the topics of interest in a city is an
124 essential complement to the survey and interview data traditionally used. Other
125 authors [21, 22] look at the language of the city in a manner similar to reading
126 works of literature. Still others employ the paratextual meanings of advertis-
127 ing [23], interpretations of art and architecture [24], or landscapes [25]. Nonethe-
128 less, the nature of the data used, combined with the interpretation-intensive meth-
129 ods proposed, limits the scope of these studies substantially.

130 *Big Data and Social Media as Tools for Reading the City*

131 Contemporary big data has the potential to identify linguistic patterns in the
132 city in a more systematic fashion, even if this possibility has not yet been fully
133 realized. Twitter data for example has been used in many geographical analy-
134 ses [26, 27], for instance, to explore the degree of geographic concentration of
135 virtual social networks [28], to predict intrinsic population characteristics (in-
136 cluding the geographical location of the authors) from food-related tweets [29],
137 and to map linguistic communities in a cosmopolitan city like Milan [30] and
138 “city communities” through machine learning techniques in Bogotá [31]. In a
139 recent study, researchers used geo-located tweets to measure the social integra-
140 tion of U.S. cities based on the everyday travels of people across neighbour-
141 hoods [15]. It is worth noting that Twitter data, even when georeferenced, is
142 limited in its geographic scope since there is no guarantee that a tweet posted
143 from a given location is about that location. Nonetheless, as with earlier stud-
144 ies [15], our method is applicable to Twitter just as it is applicable to Yelp. The
145 method works with any large-scale text corpus, we have simply selected Yelp as
146 the resource to apply it on in this paper.

147 Other location-based services have provided an avalanche of data for urban
148 and regional scholarship, but these too are not always ideal for uncovering pat-
149 terns of urban discourse. Foursquare data for example has been used to locate
150 groups of users in the city by analyzing the places they visit [32], to distinguish
151 cultural boundaries by looking at food and drink habits [33, 34], to identify the
152 sources of urban “buzz” [35], to identify neighbourhood boundaries through pat-
153 terns of collective activities and social proximity [11, 36, 33], and in urban plan-

ning to allocate retail stores in the city better [37]. In relation to neighbourhood attributes and change, Foursquare data has been compared to traditional spatial data (e.g., property maps) to study the relationship between the presence of specific physical venues, such as tanning salons or flea markets, and the neighbourhood socioeconomic deprivation [38, 28]. Other studies have shown that Foursquare—and similar location-based services—accurately predict changes in local socioeconomic attributes, which is especially useful for regions with unreliable or delayed official data sources [8, 9]. Urban patterns have been extracted from other types of georeferenced data [39]. A notable example is the Streetscore method, which employs deep learning to extract information about the perceived quality of an area using Google StreetView data [40].

Nevertheless, Foursquare and similar services are limited for the purpose of reading the city due to the lack of user-generated text. Additionally, Foursquare’s categories are not fully structured—the categorization of venues is not always reliable. Overall, while neighbourhood change is a recurrent topic of analysis that has received new energy as novel data and techniques allow long-term trajectories to be analyzed systemically, the discursive meanings associated with neighbourhood change are difficult to capture beyond case studies. There is abundant research on gentrification and displacement focused on changes in the demographic composition of neighbourhoods, and changes in their built environment — e.g, the spread of cafés. Other studies that explore the experiences of gentrification by residents are usually limited in geographic scope.

Yelp reviews, by contrast, provide a large dataset on how people talk about neighbourhoods. Yelp offers multiple distinctive data attributes that have been exploited in previous research, though this neighbourhood dimension has not been featured to our knowledge. For example, studies have used natural language analysis tools to test the relationship between language choices and the rating given to a venue on Yelp. For instance, Jurafsky and colleagues extract the narrative framings more commonly associated with positive and negative reviews of restaurants [41, 42]. Other studies show how the use of expressions of authenticity in reviews is usually associated with higher ratings [43]. The rich categorization of Yelp venues, particularly restaurants, allows us to categorize its users according to consumption patterns and explore their likelihood to value “pure” versus “transgressive” cultural objects [44].

Yelp data has been used for neighbourhood change and gentrification research, albeit not to its full potential. Yelp’s record of active venues has been used to monitor changes in the local economy of neighbourhoods, complementing the lagged official data such as County Business Patterns in the U.S. [7].

Glaeser and colleagues [45] also used Yelp reviews of coffee shops as proxies of change in the local economy of neighbourhoods associated with gentrification. The findings suggest that Yelp data can provide real-time and up-to-date insights into the changes in the neighbourhood economy and gentrification before the official statistics are released. Studies on neighbourhood change that do engage with review texts have limitations due to the nature of the methods they use. Zukin, Lindeman and Hurson [46] identify racial bias in comparing the Yelp reviews of restaurants in two gentrifying neighbourhoods—one predominantly Black and another predominantly White. Their study, however, is limited to 7046 reviews, since they use manual coding to capture terms that show the implicit racial bias.

In sum, our brief survey indicates that Yelp data has the potential to make the abstract notion of a neighbourhood’s collective representation by myriad individuals empirically tractable. Yelp reviews offer a set of advantages for the identification of geographic and linguistic patterns given that all information is explicitly about a specific location in the city, known by both the user and the researcher. Just as important, this dataset captures changes over time. We can therefore identify when certain discourses emerge and how they spread, using the tools of modern computational text analysis. Ultimately, neighbourhood changes matter the most because they affect peoples’ experiences in the city. We aim to show how new data and methods can incorporate these experiences into quantitative studies of neighbourhood change.

Computational Methods for Large-scale Latent Text Analysis

To uncover the discursive structures associated with neighbourhoods in the Yelp data, we apply and extend computational techniques using large-scale social media data to understand city dynamics [11]. This literature has addressed the issue of recognizing patterns of discursive stability and change in cities through longitudinal analysis of social media text. While a full review is beyond the scope of this paper, we focus specifically on latent analysis methods that aim to reduce large volumes of textual data to low dimensional representations that preserve meaning and facilitate measurement of semantic similarity.

A first class of latent text analysis methods relate to clustering and topic modeling methods that seek to assign explicitly interpretable latent labels to “documents”. In the context of large-scale social media data in an urban setting, a document may consist of an aggregation of all text content relating to a neighbourhood, an individual, a business, a time period, or any other observable factors (or combination thereof). Clustering text documents using K -means has a long history [47] and assumes that every document belongs to one of K latent

classes (or types); by analyzing the word frequencies of text with the same latent class assignment, it is often possible to understand the meaning of clusters through their most frequent words. In a related vein, others use unsupervised topic modelling methods such as Latent Dirichlet Allocation [48], increasingly common in social sciences [49, 50], to examine semantic trends in urban spaces (e.g. [51]). Unlike clustering, topic modeling allows a single document to have multiple labels to varying degrees. One goal of topic modeling in this context is to ascribe each neighbourhood to one or more topic labels that can be understood according to the vocabulary distribution learned for each topic.

A second type of study focuses on machine learning and deep learning methods to encode documents or words into low dimensional latent spaces that lack explicit interpretation, but which facilitate semantic similarity comparison of embedded content. Two of the standard models for word embeddings, GloVe [52] and word2vec [53], transform words from English to a vector-based encoding. This encoding captures the meaning of each word by positioning it in embedding space, such that the distance between any two words corresponds to their similarity. These approaches, then, create an embedded space for words to exist within, such that the exact position of the word captures essential information about its meaning. These methods can also be extended to documents as in doc2vec [54].

Our paper builds on this second type of embedding approach through a novel application of an autoencoder model on geolocated textual data. While the architecture of word and document embedding models diverge somewhat from a standard autoencoder model, it is well established that the latter is strongly suited to extracting low-dimensional embedded representations of text that retain the underlying meaning and facilitate semantic similarity analysis (e.g. [55, 56, 57]). This allows us to identify geographically-anchored semantic changes over time through the similarity computations enabled by the use of embeddings.

3. Data and Methods

Data

Our study uses data from Yelp.com. Yelp is a popular consumer review website. For instance, it offers approximately 163 million reviews, in 18 different languages, and the average number of unique website visitors is around 150 million [58]. Over 92% of Yelp users are from the United States, with Japan, India, and Canada as its next largest, but at a modest 1% per country. In addition to reviews, Yelp provides information and reminders about local events, lists of local businesses and their offers, can make restaurant reservations, and allows those who have an account to connect with (i.e. follow) other ‘Yelpers’.

267 As with many social media data sources [59], Yelp represents an unbalanced
268 sample of the total population of the cities it covers.¹ Yelp nevertheless tracks
269 differences between real neighbourhoods, as we show below. It should also be
270 noted that Yelp data cannot be construed as representative of a particular geo-
271 graphically defined population, which may pose concerns in terms of external
272 validity and generalizability. For this study, the location of reviewers is irrele-
273 vant, as the main goal is to analyze changes in the reviewers’ language, to under-
274 stand better how city neighbourhoods are represented and change over time.

275 Yelp is attractive as a data source for many reasons. In particular, it contains
276 rich information about (a) points of interest (POI) and (b) qualitative assessments
277 of these POI (reviews). POIs are mainly businesses, but also may include public
278 spaces such as parks or hiking trails. For each POI listed, there is a large amount
279 of data available describing its location, opening hours, and other minutiae. Each
280 POI is classified within a hierarchical list of approximately 1200 categories and
281 by ambience (e.g. ‘family-friendly’ or ‘romantic’). These characteristics are
282 valuable in learning about the type of features available in different areas, yet
283 the reviews add more: qualitative and evaluative assessments. These combine a
284 star rating (out of five) with a review written by the user. The reviews are date
285 stamped and correspond to a discrete location, enabling geospatial and temporal
286 analysis.

287 We have collected Yelp data for the city of Toronto, which we use as a case
288 study for developing methods that should be applicable elsewhere. The dataset
289 includes approximately 24,000 businesses and 500,000 reviews. This number
290 was validated against results from [7], which found that across the US, Yelp had
291 18.7% of the listings that the official business patterns dataset did. Our Toronto
292 dataset comprises a similar proportion of all businesses officially registered in
293 Toronto, which includes many non-consumer facing businesses that would not
294 be expected to have Yelp pages. In addition to the count of establishments, our
295 dataset also includes information about each POI’s category (e.g. ‘sushi restau-
296 rant’ or ‘rock climbing’) and all of its reviews. Each review is also indexed to the

¹In terms of demographics, Yelp reviewers are relatively equally represented within the 18-34 and 35-54 age groups (i.e., standing at approximately 35%), with the 55+ age group representing about 30%. Furthermore, over 60% are college-educated, and an additional 23% have a graduate degree. Concerning income, almost half make over \$100,000 annually, and another 25% are in the \$60,000-99,000 range. These numbers suggest that while Yelp reviews capture a broad swathe of the population, as with other Internet data, they skew toward the younger and more educated (e.g. [60, 61]). Information about the gender distribution, as well as other attributes among Yelp users, has not been made available by the company. However, independent studies indicate that Yelp’s user base is composed slightly more of women than of men [62, 63].

reviewer who wrote it. Finally, we have a Business ID, a User ID, and a location (lat/lon) associated with each review.

Analytical Approach

In order to suitably represent neighbourhoods, we adopt a standard term-frequency bag-of-words (BoW) representation (as seen in e.g. [64, 65, 66]). This method indicates only the presence of words and discards their order. We could compute similarity directly on these vectors, but as they are sparse, this would not capture the different ways linguistic expressions may refer to the same things. To address this, Deerwester [67] established Latent Semantic Analysis (LSA) as a means to apply Principal Component Analysis to linearly project sparse vector-based data to a lower-dimensional dense representation where semantically similar expressions are located nearby in vector space. In the modern era of deep learning, we can now learn non-linear embeddings from large datasets that can better exploit the non-linear relationship between word co-occurrence to produce improved low-dimensional representations for similarity comparison. In order to compute the similarity metric over neighbourhoods, we employ Cosine Similarity, which is the standard metric for any vector-based representation [47].

To answer our first research question — do Yelp review data allow us to identify neighbourhood boundaries using text analysis and machine learning techniques? — we introduce the concept of the autoencoder, and describe how our data is formatted to train the model. We then explain the underlying concept of *latent embedding space*. Using these concepts, we show that Yelp review texts accurately identify the language patterns characteristic of the City of Toronto’s 140 official neighbourhoods (Figure 1), despite having not been trained on data directly about them. In order to demonstrate the validity of this result, we conduct a “fake neighbourhood experiment,” in which our model is tasked with distinguishing “real” and “fake” neighbourhoods. We find it does so with a high degree of accuracy, especially in contrast to alternatives.

To answer our second research question — is it possible to identify neighbourhood change at the discursive level? If so is it possible to identify which neighbourhoods become more similar or more different to one another over time? — we aggregate the Yelp data into vectors representing individual reviews (R), businesses (B), users (U), and neighbourhoods (N). We then break up these vectors by year to create a platform for temporal analysis. We embed these units in the same space as the atemporal units so that we can track the motion of a single object — the linguistic patterns of a single neighbourhood — over time. By embedding the temporal neighbourhood objects in a single space, we can construct a concept of stable vs unstable neighbourhoods. Stable neighbourhoods

Figure 1: The 140 officially-defined neighbourhoods of Toronto.

do not move much in the embedding space year-on-year, while unstable neighbourhoods move rapidly. We treat the stable neighbourhoods as anchors and examine how other neighbourhoods change relative to them. Finally, we employ K-Means clustering to identify different groups of motion through the resulting proximity space, such as those neighbourhoods which are rapidly changing, or well-established downtown neighbourhoods which, despite being home to some of the city’s most dynamic cultural scenes, change little in terms of their collective meaning. This approach allows us to read the city in terms of how neighbourhoods function as templates or models that others become more or less similar to over time and to classify urban text by its patterns of evolution.

Figure 2 summarizes the encoding process. The concrete steps taken, from cleaning and vectorizing the raw data to performing neighbourhood recognition experiments, are detailed in the following subsections.

Figure 2: Methodology pipeline from raw review data to the result of the first experiment. For the baseline comparison, the autoencoder step is omitted and the full 1024-dimension vectors are used in the Support Vector Machine (SVM) step.

Cleaning and Vectorization

The standardization of review text is initially noisy and somewhat unreliable. We employ a natural language processing tool, spaCy [68], to extract standard word tokens from reviews. Using spaCy, we can transform the raw text into a set of standard words. We then remove accents, as well as stop words, which are non-informative words such as “the,” “of” or “and.” We also omit any words which occur in less than 100 distinct reviews.

We then convert each review into a 1024-long vector of integers. Each value i in this vector corresponds to the i^{th} most commonly used word, or pairs of words, in the dataset, and its value corresponds to the number of times that word is observed in the review. Our selection of 1024 words is not arbitrary - it has been shown [69] that in any large corpus, a tiny proportion of the words *available* can account for a considerable proportion of the words *used*. In our case, the 1024 most frequently used words account for approximately 70% of all words used in reviews. We consider pairs of words as well, in order to include some common phrases in our dataset. Of the 1024 words used, 54 are word pairs. Word pairs include phrases such as ‘ice cream’, ‘pad thai’ or ‘years ago’.

From our 1024-dimension vectors representing each review, we then construct vectors representing users (U), individual reviews (R), businesses (B), and neighbourhoods (N). For each of these, we simply add all of the review vectors corresponding to that subject (e.g. a single business). We normalize these vectors to sum to 1, so that the size of the vector does not correspond to review length.

I had a good experience at this store. I was going to order online but instead thought I'd ask a few questions before making the purchase. I received good and knowledgeable responses and advice, and elected to make the order through the store rather than online. It was an easy process. (I ultimately picked up the mattress instead of opting for free delivery since delivery was booking out farther than we would have liked.)The actual mattress we chose - The Bloom Cloud - is super comfy and we are pleased with the quality. The mattress comes compressed in the box and you can easily set it up on your own. It is also feels really light once open, \xa0which makes moving it around and making the bed easy. I would shop here again and would recommend this store.

Figure 3: The raw review text.

good experience store going order
online instead thought ask
questions making purchase received
good knowledgeable advice make
order store online easy process
ultimately picked instead free
delivery delivery booking actual
chose super comfy pleased quality
comes box easily set feels really
light open makes moving making
bed easy shop recommend store

Figure 4: The same review text, after the entire cleaning and vectorization process.

371 *Dimensionality Reduction*

372 After obtaining the review vectors for all categories of objects we wish to
373 analyze, the final stage is reducing the dimensionality of the vectors. This step
374 allows us to 'combine' words in the vector which have similar meanings, or al-
375 most always co-occur. Examples of such words are entries such as 'making' and
376 'makes'. We select a dimensionality reduction to 2^4 , i.e. 16 dimensions. In order
377 to accomplish this task, we employ a standard fully-connected autoencoder, the
378 structure of which is shown in Figure 5.

379 *Distinguishing Real and Fake neighbourhoods*

380 In order to demonstrate the validity of the resulting 16-dimensional vec-
381 tor representations of neighbourhoods, we devised an experiment that requires
382 meaningful information to be included in the vectors. This experiment requires a

Figure 5: The structure of the Autoencoder model used. The first 3 dense layers, together referred to as the Encoder, transform the full 1024-dimensional vector into a compact, 16-dimensional representation. The remaining 3 dense layers, known as the Decoder, then use this 16-dimensional representation to reconstruct the original data as accurately as possible. The 'None' value paired with the dimensionality at each layer is due to the fact that any number of 1024-dimensional vectors can be given to the model at a time - each one is still operated on independently of the others.

383 classification model to correctly distinguish between the 140 real Toronto neigh-
384 bourhoods, and a collection of ‘fake’ neighbourhoods constructed from real data.

385 To generate our fake neighbourhoods, we elected not only to draw reviews
386 from the original review sets, but also to select reviews from a contiguous region
387 of the city. In this way, the ‘fake’ neighbourhoods are nearly identical to real
388 neighbourhoods — they comprise a specific region of Toronto and include all of
389 the reviews within this region. The only distinguishing feature is that the fake
390 neighbourhoods are not officially recognized as neighbourhoods (Figure 5).

391 We use the following constraints to design our methodology for fake neigh-
392 bourhood generation.² A total of 140 fake neighbourhoods are generated, so that
393 randomly guessing the veracity of all neighbourhoods would generate an accu-
394 racy of approximately 50%, the lowest effective score on this task. We selected
395 a Support Vector Machine (SVM) model for this task.³

- 396 1. Fake neighbourhoods have roughly the same area distribution as real neigh-
397 bourhoods.
- 398 2. Fake neighbourhood boundaries are drawn as polygons, with a number of
399 sides corresponding to that of real neighbourhoods.
- 400 3. A fake neighbourhood should encapsulate at least one business since there
401 is no neighbourhood that does not contain any businesses.

402 Once the fake neighbourhoods have been generated, all neighbourhoods are
403 transformed into the low-dimensional space, shuffled, and randomly assigned to
404 the training and testing sets. 67% of the vectors are assigned to the training set
405 and the remaining third to the testing set. When comparing the similarity of two
406 neighbourhoods in the embedding space, we must select a suitable comparison
407 metric. For this, we use the Cosine Similarity metric that is the standard for
408 text-based analysis [47], which is defined for two vectors A and B as follows:

$$\cos(\theta) = \frac{\sum_{i=1}^N A_i B_i}{\sqrt{\sum_{i=1}^N A_i^2} \sqrt{\sum_{i=1}^N B_i^2}} \quad (1)$$

²Our full code for this process can be found in the codebase hosted on GitHub.

³ SVM models identify the decision boundary which best separates the two classes of data in the reduced dimensionality space. A core benefit of selecting this type of model is that it does not require a significant amount of training data. This is important as we only have a total of 280 samples for the real and fake neighbourhoods. Our SVM was trained using the default parameters provided in the Scikit-Learn package [70], i.e. a regularization constant of 1.0 and a kernel coefficient of 1.76, and achieved high accuracy during training and validation with a mean training accuracy of 0.87 across 100 runs. This indicated that the model was tuned correctly.

Figure 6: 140 fake neighbourhoods generated using our procedure.

Intuitively, the cosine similarity measures the angle between two vectors, ignoring the length of those vectors. Two vectors pointing in the same direction are identical, and so would have a similarity of 1, while two vectors pointing in opposite directions would have a similarity of -1. Orthogonal, or unrelated, vectors have a cosine similarity of 0.

4. Results

Autoencoder Reconstruction Error

The autoencoder model that was designed for this task was trained on a combination of the vectors representing individual reviews (R), the vectors representing businesses (B), and the vectors representing users (U). As stated earlier, we hold out the neighbourhood vectors N so that, during the fake neighbourhood detection task, the model cannot “remember” that it has seen some of the neighbourhoods before.

As a result of this design, however, it is essential to verify that the model can accurately embed N despite never having been presented with it during training.

424 To test this, we pass N through the autoencoder network and record the cosine
 425 similarity between the original vector and the reconstructed vector (Table 1).

Data	Cosine Similarity (μ)	Cosine Similarity (σ^2)
R	0.5853	0.1331
B	0.7027	0.1525
U	0.5941	0.1294
N	0.9179	0.0538

Table 1: Cosine Similarity between original and reconstructed vectors for R, B, U, N . For each result, the worst possible score - indicating total dissimilarity between the original vector and the reconstruction - would be -1, while the best possible score would be 1, indicating identity.

426 *SVM Classification Performance*

Format	Testing Accuracy (μ)	Testing Accuracy (σ)
Full Dimensionality	0.4630	0.0278
Embedding Space	0.7719	0.0341

Table 2: SVM performance in distinguishing real and fake neighbourhoods, over 100 different fake neighbourhood generations. The worst possible score is 0, however in practice the worst possible performance is around 50% as this indicates random guesses by the model. The best possible performance is 1, indicating perfect accuracy.

427 140 real and 140 fake neighbourhoods were generated, and randomly split
 428 into training and testing sets. The SVM model was able to distinguish these real
 429 and fake neighbourhoods in the low-dimensional review space with an accuracy
 430 of 0.7719 (Table 2).

431 In the full, 1024-dimensional space, however, accuracy was approximately
 432 0.46, which indicates random assignment by the classifier. This result is signif-
 433 icant for two reasons. First, and most critical, the fact that the real and fake
 434 neighbourhoods are indistinguishable in this high-dimensional space demon-
 435 strates that the fake neighbourhoods generated are sufficiently ‘realistic’. It is
 436 vital to be sure that the fake neighbourhoods are not ‘too fake’ such that any clas-
 437 sifier would be able to tell them apart from the real ones. The poor performance
 438 of the classifier in this task instead shows that the fake neighbourhoods appear
 439 real without proper analysis. Secondly, this result suggests that the structure

Figure 7: Pipeline for review reconstruction using the autoencoder. The raw review is first encoded into a vector (stage 2). This 1024-dimensional word frequency representation is then passed into the encoder to produce the embedding shown in stage 3. Finally, the decoder can be used to reconstruct the original vector approximately.

critical to understanding the difference between real and fake neighbourhoods is non-linear and occurs across multiple words in the full space. This inference can be made due to the non-linearity of the Autoencoder model, and the fact that dimensionality reduction is required to distinguish the neighbourhoods successfully.

In sum, the ‘fake neighbourhood’ experiment indicates that Yelp reviews can be used to identify real neighbourhoods with a high degree of accuracy. In turn, it suggests that Toronto’s 140 official neighbourhoods do exhibit relatively distinct styles, and that the official neighbourhood boundaries map to some degree onto underlying differences in how the city is represented. That fake neighbourhoods can be consistently distinguished from real in the latent space, but not in the full word frequency space, suggests that the Autoencoder model is critical to extract the latent information included in the review text.

Interpreting Neighborhood Embeddings

The embedding space produced by the Autoencoder model is a challenge to analyze. Though we can be sure that the representation learned by the model is a compact representation of the full data, we cannot directly tell *what* meaning is contained in each of the dimensions in the low-dimensional space. To gain some insight as to the meaning of these dimensions, we provide the second stage of the Autoencoder – the decoder – with a vector in which one dimension is fully maximized, and all other dimensions minimized. By passing such a vector to the decoder, we can learn which words in the full 1024-dimensional space are most strongly associated with each dimension in the encoded representation. For each of the twelve dimensions, we show here the three most strongly associated words (Figure 3). This type of analysis can be referred to as *Sensitivity Analysis*, as it allows us to examine what aspects of the input data each part of our neural network is sensitive to.

In sum, each dimension of our latent representation captures distinct aspects of what typically defines a business. In depicting the top three words for each dimension here, we demonstrate how our encoder has learned which words essentially mean the same or highly similar things when used to describe a business. In this way, the encoder has deduced that their distinct meanings can be compressed into a single feature. While articulating the full meaning of these dimensions would require going beyond these three-word summaries, in the present context, the simple point is that they do carry distinct meanings that a richer interpretation could seek to unpack.

Dimension	1st Word	2nd Word	3rd Word
0	order	delivery	service
1	burger	park	station
2	store	good service	service
3	company	called	phone
4	shop	guys	service review
5	salon	terrible	dress
6	good	place	food
7	tea	cake	cheesecake
8	dress	wedding	store
9	did	went	hair
10	pad thai	nail	nails
11	salon	dress	nails
12	terrible	waited	order
13	bakery	park	cake
14	delicious	atmosphere	cocktails
15	sushi	rice	pho

Table 3: The three words with the highest value when a 1x16 vector containing a single 1 and 15 zeros is passed. Each word triple corresponds to the K^{th} dimension being given the single 1.

476 *Temporal Analysis: Intra-neighbourhood Change*

477 Having answered our first research question in the affirmative, we now turn to
 478 the second. By separating each neighbourhood’s reviews by year, and then em-
 479 bedding all reviews into the same embedding space, we can examine the change
 480 over time in a neighbourhood’s embedded review language. In the following
 481 sections, we examine how neighbourhoods change in relation to each other, but
 482 first, we study how each neighbourhood changes relative to itself over time.

483 We define stability as the following for some neighbourhood N :

$$stability(N) = \prod_{i=2}^{|N|} cosine_similarity(N_{i-1}, N_i) \quad (2)$$

484 For e.g. four years, can be expanded out to the form:

$$\frac{A \cdot B \times B \cdot C \times C \cdot D}{\|A\| \times \|B\|^2 \times \|C\|^2 \times \|D\|} \quad (3)$$

485 We employ this metric, as opposed to an average, in order to represent the
 486 “through-line” of semantic similarity. For some neighbourhood with a high inter-
 487 year similarity, save for two years in which the words used are entirely different,
 488 then the average cosine similarity would still be high, losing the insight that at
 489 one point the word usage entirely changed. With our metric, however, this single
 490 low similarity would affect the entire result, ensuring that no neighbourhood with
 491 a high stability value can have *any* years of dissimilarity within.

492 For a small set of neighbourhoods, we identified stability below 0.0. These
 493 neighbourhoods are noted in Figure 8 as white, and we remove them from further
 494 analysis as we consider their motion to be too erratic to analyze. These neigh-
 495 bourhoods all have a very low quantity of reviews associated with them - while
 496 the average neighbourhood contains review text amounting to approximately 2.2
 497 million words, none of these neighbourhoods surpassed 300 thousand words,
 498 with the average among them being just 80 thousand.

499 Figure 7 shows in bold the neighbourhoods with the highest stability. These
 500 are clustered in the western part of the downtown area, with the Annex neigh-
 501 bourhood showing the highest stability overall (0.9821). Home to Jane Jacobs
 502 for the last decades of her life and nearby the University of Toronto, this area was
 503 in the past a centre of Toronto’s alternative scenes and continues to have one its
 504 highest concentrations of artists and young people – even as home prices have
 505 risen dramatically in recent years [71, 72, 73]. While it and most of the other
 506 bolded neighbourhoods downtown are generally recognized as some of the most
 507 dynamic parts of Canada’s creative economy, this result indicates that despite

508 this – or perhaps because of it – their *meaning* in the collective imaginary of the
 509 city has remained highly stable throughout the period under study.

510 This stability result overlaps other research using more conventional data
 511 sources [74]. Silver’s approach for example uses harmonized census data to
 512 examine neighbourhood change in Toronto. This technique finds a similar pat-
 513 tern to ours, with the most “creative” neighbourhoods in terms of population —
 514 young, artistic, creative class workers — remaining highly stable. By contrast,
 515 inner suburban areas showed a higher probability of transitioning to alternative
 516 neighbourhood types. Combined, these findings suggest that the underlying se-
 517 mantic structure of the downtown core of the city remains mostly stable when
 518 compared to its further reaches.

519 *Temporal Analysis: Inter-neighbourhood Change*

520 We can use the same metrics defined in the previous section to compare the
 521 change in different neighbourhoods year on year. This allows us to measure how,
 522 for any pair of neighbourhoods, they become more or less semantically similar
 523 over time. Figure 9 provides an example of this, showing two neighbourhoods
 524 – Pleasant View and North St. James Town – which become distinctly more
 525 similar over time.

Figure 9: Cosine similarity between embeddings for Pleasant View and North St. James Town over 9 years. Blue line represents original cosine values, orange line represents linear fit.

526 Unfortunately, when comparing two neighbourhoods in this way, it is dif-
 527 ficult to easily ascertain which neighbourhood is responsible for what part of

Intra-neighborhood similarity (2009-2018)

Figure 8: neighbourhood Consistency (2009-2018). A value of 1.0 represents no motion at all through embedding space, while a value of -1 would correspond to total orthogonality to the initial vector (Note that all values here are positive, indicating some minimal similarity with the original vector is always present in this dataset). Bold neighbourhoods are the top 10 most stable.

Figure 10: Identifying the optimal clustering K for neighbourhood motion using the Calinski-Harabasz criterion.

the motion. Two neighbourhoods moving slowly together may appear the same as one neighbourhood moving quickly towards another. Additionally, with this metric of comparison, it is difficult to plot more than one pair of neighbourhoods on a single graph.

As we wish to compare all neighbourhoods simultaneously, we need to introduce a “ground truth” to which all neighbourhoods can be compared. For this purpose, we select the most stable neighbourhood from the previous section – the Annex. We consider this a strong candidate for the “ground truth” as while this neighbourhood does contain its own information, its extremely low motion through the embedding space allows us to examine the motion of other neighbourhoods relative to this very stable Annex embedding.

Temporal Analysis: Clustering neighbourhood Change

In order to identify underlying trends in the motion graphs, we utilize K-Means Clustering [75], to group together similar motion. In order to do so, K-Means iteratively assigns each motion graph to a cluster with the most similar average motion, recalculating this average motion and reassigning motion graphs until convergence. In order to select an appropriate number of clusters, K , we employ the Calinski-Harabasz criterion [76], which is defined as:

$$\frac{SSE_B}{SSE_W} \times \frac{N - k}{k - 1} \quad (4)$$

Where N is the number of data points (here 139), k the number of clusters, and SSE_B and SSE_W respectively the between- and within-cluster Sum Squared Error. The Calinski-Harabasz criterion is well established as a suitable metric with which to identify an appropriate value for K in clustering tasks (e.g. [77, 78, 79]). Figure 10 shows the Calinski-Harabasz score for potential values of K , and we select the value for K at the knee point, defined as the point of maximum curvature in a line [80]. This gives us a value of $K = 12$, which we then use to plot motion graphs separately. These different types of motion are then presented in Figure 11, and the clusters are shown on a map in Figure 12.

Figure 4 shows that neighbourhoods cluster into a few core types of motion, relative to the stable meaning of the Annex. Each line represents a neighbourhood, and points closer to the left side of the graph are more similar to the Annex. Accordingly, Cluster 4 is a set of neighbourhoods that are both relatively stable and represented similarly to the Annex throughout. For these neighbourhoods, the 2009 and 2018 vectors are so similar that representing the 2018 neighbourhood with the vector from 2009, would lose almost no information. Cluster 0 neighbourhoods are also stable but are not as similar to the Annex. Cluster 8 neighbourhoods – a small group on the outskirts of the city – start highly dissimilar to the Annex but have grown steadily more similar to it over time. Major immigrant landing areas, these neighbourhoods correspond to locations in which significant change has been ongoing. Cluster 2 areas also became more similar to the Annex, though they have retained a more distinct representation. Cluster 9 neighbourhoods have become less similar to the Annex in recent years. Cluster 1 and Cluster 7 – geographically near each other – each exhibit momentary ”shocks” in which they temporarily and drastically take on very distinct meanings, first in 2011 (in Cluster 1) and second in 2012 (Cluster 7).

Fully articulating and interpreting these different trajectories in relation to their underlying business, review, and user features, as well as other demographic information, goes beyond the scope of this paper. Nevertheless, Figure 7 constitutes a strong affirmative answer to our second research question. Yelp data do provide a powerful window into the temporal evolution of the meaning of urban neighbourhoods.

Figure 11: The motion clusters. The cluster indices correspond to those in Figure 12.

Figure 12: The clusters selected by K-Means clustering on motion graphs. Each color represents a distinct cluster, and the colorings chosen are not related to each other. The cluster indices correspond to those in Figure 11.

5. Discussion

In this paper, we present a novel methodology for uncovering linguistic structures of cities via social media review texts. In this section, we discuss the implications of this contribution, as well as our experiments.

Real vs Fake neighbourhoods

The high degree of accuracy that the classifier was able to obtain in the low dimensional space demonstrates the validity of using this review space to examine neighbourhoods. It is important to emphasize that the ‘fake’ neighbourhoods are indeed real, contiguous sections of the city, meaning that the model was able to distinguish at a high conceptual level the differences between these two sets. Reviews reliably map onto the real neighbourhoods, and the official neighbourhoods exhibit coherent meanings.

Nevertheless, not all neighbourhoods were accurately classified. There are several possible explanations for the incorrectly classified neighbourhoods. For one, due to the random nature of the fake neighbourhood generation, it is entirely possible to accidentally generate a ‘fake’ neighbourhood, which in reality overlaps significantly with a real neighbourhood. In this situation, it is reasonable for the model to misclassify the fake neighbourhood as real.

In the other direction, regarding real neighbourhoods that are misclassified as fake, such neighbourhoods mainly occur in suburban areas of the city. It is possible that these areas, which often have fewer businesses and reviews when compared to the downtown core of Toronto, are noisier and so appear less ‘real’ to the model. This is an area for further research.

The broader significance of this analysis emerges when we compare real and fake neighbourhoods in the full-dimensional space, vs that of the embedding, as in Figure 13. This graphic depicts how distinguishing real and fake neighbourhoods in the embedding space is far easier, even upon visual inspection, than it is in the full-dimensional space. This reinforces that the type of analysis we have developed is not possible in the original word frequency space, and that the learned embedding does indeed extract latent information critical to our analyses.

Stable neighbourhoods

The stable neighbourhood result (Figure 8) demonstrates that especially the west-end downtown centre of Toronto has the most coherent identity compared to the more volatile surrounding areas. The exception to this is Miliken, the only top MSN not in central Toronto, which is located in the far Northeast part of

Figure 13: Comparison of the average values for each dimension in the full, 1024-dimensional space, as well as the 16-dimensional embedding space, for real and fake neighbourhoods. The distinction between values in the embedding space is much easier to identify, and these distinctions can be significantly larger, which allows the model to identify fake neighbourhoods far better in the embedding space than it can in the full space.

the city in bold in Figure 8. This neighbourhood contains a significant number of business parks, built during the 1990s, which are likely responsible for the neighbourhood’s stability in the 2010s.

This ability to identify areas with coherent and stable meaning on the basis of review texts is a central contribution of this paper. Most research uses qualitative and survey methods to identify shared local meaning, often suggesting linkages between neighbourhood identity, solidarity, and local advocacy. Our techniques capture similar information about identity and a platform for pursuing such linkages, but in a way that is scalable, reliable, and more systematically comparable.

An additional benefit of our approach concerns the fact that even the most stable neighbourhoods do have *vocabulary* change, despite minimal semantic change. This is a key feature of the embedding space: the vocabulary itself is discarded, and meaning is retained. However, as a result of this, it is critical to note that individual words may easily change year-on-year, even in the most stable neighbourhood. The collective meaning of an area transcends changes in specific words, such that a neighbourhood like the Annex is consistently recognized as having a similar meaning within the discursive space of the city, even if specific words used to describe it change. This is an example of how the computational analysis pursued here carries forward key ideas from cultural research, namely about the holistic and relational character of meaning [49, 6].

Motion graph clustering

The motion graph clustering results, shown in Figures 11 and 12, clearly show that by reading the city, we can identify a range of different neighbourhood motions, as discussed above.

The confluence of geography and discursive meaning is particularly striking. Cluster 7, for example, contains a group of neighbourhoods tightly compacted together in the west end of Toronto (this is the Weston neighbourhood). This is despite the total lack of proximity information provided to the model. Patterns of meaning are not randomly distributed in space but tend to concentrate nearby and move along similar trajectories. Indeed, these neighbourhoods are all similar in the dissimilarity spike they all experience in 2012. This dissimilarity spike highlights the necessity of grouping motion graphs together. On their own, each motion graph might appear to have some noise causing the graph to change erratically. However, when clustered, it becomes apparent there was some activity in the west of Toronto in 2012, which caused this cluster of neighbourhoods to all rapidly change in the same way.⁴

⁴There are a number of clusters, most notably 5, 10 and 11, which contain a very small set

6. Conclusion

In this paper, we have introduced a methodology for overcoming some of the challenges associated with the empirical analysis of collective representations of urban spaces. These techniques might be considered a form of sociological observatory to complement recent uses of Yelp data for economic monitoring. The economic version uses social media data to identify patterns and trends in business activity before or beyond what official government data permit [7, 8, 9, 45]. Our sociological version uses similar data to identify patterns and trends in collective representations and identities. As a result of this, it expands the range of approaches available to researchers, allowing for more nuanced and effective research of neighbourhood change.

The methods presented here enable a range of novel quantitative analyses. For example, the fake neighbourhood experiment demonstrates that the embeddings capture *latent* information in the reviews, which are not easily accessible by reading directly (e.g. [81]). Further techniques reveal the relative temporal evolution of neighbourhoods' meaning within the city's discursive space. These techniques open up opportunities for investigating urban processes such as gentrification in a holistic way. For instance, a neighbourhood receiving increasing investments in recent years like Weston [82]—part of Cluster 7 in Figures 11 and 12—can be identified as becoming more like the Annex, an already gentrified neighbourhood. Findings like these allow us to ascribe tangible, interpretable physical neighbourhood change to these business review embeddings. Such validation shows that the work of the article goes beyond mathematical diversion, and instead is the genesis of a new computational methodology for identifying meaningful neighbourhood change through the lens of denizen-generated textual artifacts.

Developing techniques to uncover the causes and meanings of these and the other changes depicted in Figure 8 is an essential area of further research. Equally important is comparing these discursive trajectories to the sorts of trajectories revealed in the work of Delmelle [83, 84], which is based on demographic trends in, for instance, income, occupation, ethnicity, and age. The interplay between the two – demographic and discursive – raises questions for future research. For instance, we can ask which discursive patterns are most associated

of neighbourhoods. This is a valid result, as the motion graphs for these neighbourhoods do not closely correspond to those of any other neighbourhoods. It is important for our model to be able to correctly identify these neighbourhoods as outliers from general trends so that the trends which *do* exist are not ignored in favour of ensuring some minimum cluster size.

with which demographic patterns, which tend to shift before the other in various times and places, the extent to which one remains stable as the other changes, and more. The techniques developed in this paper put pursuing such questions squarely on the research agenda.

Other opportunities for future research involve further extending the use of modern computational tools. For example, even though the embedding process described here reliably identifies neighbourhoods and their trajectories, due to their highly compact form, their interpretability remains somewhat opaque. Developing tools for unpacking their meaning is a crucial task for future research. Likewise, we can pursue techniques that remove the need to use official neighbourhood boundaries as a baseline. Indeed, other research has suggested that social media data can be used to identify emergent neighbourhood boundaries that arise from out of activity patterns themselves [28, 35]. In order to scale up the techniques developed here to many more cities – another important future step – such methods will be crucial.

Perhaps the most exciting avenue involves moving from description to prediction. Can information about a neighbourhood in Year N, combined with its reviewers and constituent reviews, be used to estimate the direction the neighbourhood can be expected to change in the following year? As a first step toward realizing this goal, novel techniques will be required for identifying what changes in reviews, users, or businesses in the past lead to changes in neighbourhood identity in the future. Building on such techniques, we believe that it may indeed be possible to track motion through the embedding space into the future. Given the embeddings’ power to identify phenomena from the past, such as specific types of urban development processes, the applicability of this toolkit to the future may enable researchers to identify trends in neighbourhood change before they even complete. In this way, we can identify not only areas poised to change, but also the type and nature of the change they are likely to undergo. This would constitute not only a sociological observatory and “early warning system” for neighbourhood change, but a genuinely predictive model of urban evolution.

References

- [1] L. R. Ford, Reading the skylines of american cities, *Geographical Review* 82 (2) (1992) 180–200.
URL <http://www.jstor.org/stable/215431>
- [2] H. Wirth-Nesher, Impartial maps: reading and writing cities, *Handbook of urban studies* (2001) 52–66.
- [3] J. M. Leach, R. A. Mulhall, C. D. Rogers, J. R. Bryson, Reading cities: Developing an urban diagnostics approach for identifying integrated urban problems

- with application to the city of birmingham, uk, *Cities* 86 (2019) 136 – 144.
doi:<https://doi.org/10.1016/j.cities.2018.09.012>.
URL <http://www.sciencedirect.com/science/article/pii/S0264275118303093>
- [4] R. Barthes, *Semiology and the urban, The city and the sign: An introduction to urban semiotics* 8 (1986) 7–98.
- [5] K. Lynch, *The image of the city*, Vol. 11, MIT press, 1960.
- [6] G. D. Suttles, *The social construction of communities*, University of Chicago Press, 1972.
- [7] E. L. Glaeser, H. Kim, M. Luca, *Nowcasting the local economy: Using yelp data to measure economic activity*, Tech. rep., National Bureau of Economic Research (2017).
- [8] D. Arribas-Bel, J. Bakens, *Use and validation of location-based services in urban research: An example with Dutch restaurants*, *Urban Studies* 56 (5) (2019) 868–884.
- [9] L. Dong, C. Ratti, S. Zheng, *Predicting neighborhoods’ socioeconomic attributes using restaurant data*, *Proceedings of the National Academy of Sciences* 116 (31) (2019) 15447–15452.
- [10] C. N. Stone, R. P. Stoker, J. Betancur, S. E. Clarke, M. Dantico, M. Horak, K. Mossberger, J. Musso, J. M. Sellers, E. Shiau, et al., *Urban neighborhoods in a new era: Revitalization politics in the postindustrial city*, University of Chicago Press, 2015.
- [11] J. Cranshaw, R. Schwartz, J. Hong, N. Sadeh, *The livelihoods project: Utilizing social media to understand the dynamics of a city*, in: *Sixth International AAAI Conference on Weblogs and Social Media*, 2012.
- [12] G. Filomena, J. A. Verstegen, E. Manley, *A computational approach to the image of the city*, *Cities* 89 (2019) 14–25.
- [13] M. Horak, A. Moore, *Policy Shift without Institutional Change: The Precarious Place of Neighborhood Revitalization in Toronto*, in: C. N. Stone, R. P. Stoker (Eds.), *Urban Neighborhoods in a New Era: Revitalization Politics in the Postindustrial City*, 2015, pp. 182–208.
- [14] R. J. Sampson, *Great American City: Chicago and the Enduring Neighborhood Effect*, The University of Chicago Press, Chicago and London, 2012.
- [15] N. E. Phillips, B. L. Levy, R. J. Sampson, M. L. Small, R. Q. Wang, *The social integration of american cities: Network measures of connectedness based on everyday mobility across neighborhoods*, *Sociological Methods & Research* (2019) 0049124119852386.
- [16] E. Talen, *Neighborhood*, Oxford University Press, 2018.
- [17] J. Hwang, *The social construction of a gentrifying neighborhood: Reifying and redefining identity and boundaries in inequality*, *Urban Affairs Review* 52 (1) (2016) 98–128. doi: 10.1177/1078087415570643.
URL <https://doi.org/10.1177/1078087415570643>
- [18] C. J. Coulton, J. Korbin, T. Chan, M. Su, *Mapping residents’ perceptions of neighborhood boundaries: a methodological note*, *American journal of community psychology* 29 (2) (2001) 371–383.
- [19] M. Jenks, N. Dempsey, *Defining the neighbourhood: Challenges for empirical research*, *Town Planning Review* 78 (2) (2007) 153–177.
- [20] J. Portugali, *Complexity, cognition and the city*, Springer Science & Business Media, 2011.
- [21] J. Jones, *The city as text: Reading paris in rayuela*, *Revista Canadiense de Estudios Hispánicos* (1991) 223–234.
- [22] L. Pelaschiar, *The city as text. the case of belfast: from gothic horror-story to post-modern*

- 767 novel.
- 768 [23] G. Genette, M. Maclean, Introduction to the paratext, *New literary history* 22 (2) (1991)
769 261–272.
- 770 [24] R. B. Goh, B. S. Yeoh, *Theorizing the Southeast Asian City as Text: Urban Landscapes,*
771 *Cultural Documents, and Interpretive Experiences*, World Scientific, 2003.
- 772 [25] J. S. Duncan, *The city as text: the politics of landscape interpretation in the Kandyen*
773 *Kingdom*, Cambridge University Press, 2005.
- 774 [26] P. A. Longley, M. Adnan, Geo-temporal twitter demographics, *International Journal of*
775 *Geographical Information Science* 30 (2) (2016) 369–389.
- 776 [27] G. Lansley, P. A. Longley, The geography of twitter topics in london, *Computers, Environ-*
777 *ment and Urban Systems* 58 (2016) 85–96.
- 778 [28] D. Quercia, L. Capra, J. Crowcroft, The social world of twitter: Topics, geography, and
779 emotions, in: *Sixth International AAAI Conference on Weblogs and Social Media*, 2012.
- 780 [29] D. Fried, M. Surdeanu, S. Kobourov, M. Hingle, D. Bell, Analyzing the language of food
781 on social media, in: *2014 IEEE International Conference on Big Data (Big Data)*, IEEE,
782 2014, pp. 778–783.
- 783 [30] M. Arnaboldi, M. Brambilla, B. Cassottana, P. Ciuccarelli, S. Vantini, Urbanscope: A lens
784 to observe language mix in cities, *American Behavioral Scientist* 61 (7) (2017) 774–793.
- 785 [31] V. Vargas-Calderón, J. E. Camargo, Characterization of citizens using word2vec and latent
786 topic analysis in a large set of tweets, *Cities* 92 (2019) 187–196.
- 787 [32] K. Joseph, C. H. Tan, K. M. Carley, Beyond local, categories and friends: clustering
788 foursquare users with latent topics, in: *Proceedings of the 2012 ACM conference on ubiq-*
789 *uitous computing*, ACM, 2012, pp. 919–926.
- 790 [33] T. H. Silva, P. O. V. de Melo, J. M. Almeida, M. Musolesi, A. A. Loureiro, You are what
791 you eat (and drink): Identifying cultural boundaries by analyzing food and drink habits
792 in foursquare, in: *Eighth International AAAI Conference on Weblogs and Social Media*,
793 2014.
- 794 [34] C. Wagner, P. Singer, M. Strohmaier, The nature and evolution of online food preferences,
795 *EPJ Data Science* 3 (1) (2014) 38.
- 796 [35] D. Arribas-Bel, K. Kourtis, P. Nijkamp, The sociocultural sources of urban buzz, *Environ-*
797 *ment and Planning C: Government and Policy* 34 (1) (2016) 188–204.
- 798 [36] A. Noulas, S. Scellato, C. Mascolo, M. Pontil, Exploiting semantic annotations for clus-
799 tering geographic areas and users in location-based social networks, in: *Fifth International*
800 *AAAI Conference on Weblogs and Social Media*, 2011.
- 801 [37] D. Karamshuk, A. Noulas, S. Scellato, V. Nicosia, C. Mascolo, Geo-spotting: mining on-
802 line location-based services for optimal retail store placement, in: *Proceedings of the 19th*
803 *ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM,
804 2013, pp. 793–801.
- 805 [38] X. Long, L. Jin, J. Joshi, Exploring trajectory-driven local geographic topics in foursquare,
806 in: *Proceedings of the 2012 ACM conference on ubiquitous computing*, ACM, 2012, pp.
807 927–934.
- 808 [39] F. Zhang, L. Wu, D. Zhu, Y. Liu, Social sensing from street-level imagery: A case study in
809 learning spatio-temporal urban mobility patterns, *ISPRS Journal of Photogrammetry and*
810 *Remote Sensing* 153 (2019) 48–58.
- 811 [40] N. Naik, J. Philipoom, R. Raskar, C. Hidalgo, Streetscore-predicting the perceived safety
812 of one million streetscapes, in: *Proceedings of the IEEE Conference on Computer Vision*

- and Pattern Recognition Workshops, 2014, pp. 779–785.
- [41] D. Jurafsky, V. Chahuneau, B. R. Routledge, N. A. Smith, Narrative framing of consumer sentiment in online restaurant reviews, *First Monday* 19 (4).
- [42] K. Reschke, A. Vogel, D. Jurafsky, Generating recommendation dialogs by extracting information from user reviews, in: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2013, pp. 499–504.
- [43] B. Kovács, G. R. Carroll, D. W. Lehman, Authenticity and consumer value ratings: Empirical tests from the restaurant domain, *Organization Science* 25 (2) (2014) 458–478.
- [44] A. Goldberg, M. T. Hannan, B. Kovács, What does it mean to span cultural boundaries? variety and atypicality in cultural consumption, *American Sociological Review* 81 (2) (2016) 215–241.
- [45] E. L. Glaeser, H. Kim, M. Luca, Nowcasting gentrification: using yelp data to quantify neighborhood change, in: *AEA Papers and Proceedings*, Vol. 108, 2018, pp. 77–82.
- [46] S. Zukin, S. Lindeman, L. Hurson, The omnivores neighborhood? online restaurant reviews, race, and gentrification, *Journal of Consumer Culture* 17 (3) (2017) 459–479.
- [47] C. D. Manning, P. Raghavan, H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press, New York, NY, USA, 2008.
- [48] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, *Journal of machine Learning research* 3 (Jan) (2003) 993–1022.
- [49] P. DiMaggio, M. Nag, D. Blei, Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of us government arts funding, *Poetics* 41 (6) (2013) 570–606.
- [50] A. Edelmann, T. Wolff, D. Montagne, C. A. Bail, *Computational Social Science and Sociology*, *Annual Review of Sociology* 46 (2020) 1–24. doi:10.1146/annurev-soc-121919-054621.
- [51] F. Kling, A. Pozdnoukhov, When a city tells a story: urban topic analysis, in: *Proceedings of the 20th international conference on advances in geographic information systems*, ACM, 2012, pp. 482–485.
- [52] J. Pennington, R. Socher, C. Manning, Glove: Global vectors for word representation, in: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.
- [53] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, *arXiv preprint arXiv:1301.3781*.
- [54] Q. Le, T. Mikolov, Distributed representations of sentences and documents, in: *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32, ICML14, JMLR.org*, 2014, p. II1188II1196.
- [55] D. Bollegala, C. Bao, Learning word meta-embeddings by autoencoding, in: *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 1650–1661.
- [56] S. C. AP, S. Lauly, H. Larochelle, M. Khapra, B. Ravindran, V. C. Raykar, A. Saha, An autoencoder approach to learning bilingual word representations, in: *Advances in Neural Information Processing Systems*, 2014, pp. 1853–1861.
- [57] J. Li, M.-T. Luong, D. Jurafsky, A hierarchical neural autoencoder for paragraphs and documents, *arXiv preprint arXiv:1506.01057*.
- [58] B. L. Ranard, R. M. Werner, T. Antanavicius, H. A. Schwartz, R. J. Smith, Z. F. Meisel, D. A. Asch, L. H. Ungar, R. M. Merchant, Yelp reviews of hospital care can supplement and inform traditional surveys of the patient experience of care, *Health Affairs* 35 (4) (2016)

- 697–705.
- [59] Y. Yuan, Y. Lu, T. E. Chow, C. Ye, A. Alyaqout, Y. Liu, The missing parts from social media-enabled smart cities: Who, where, when, and what?, *Annals of the American Association of Geographers* 110 (2) (2020) 462–475.
- [60] J. Fogel, S. Zachariah, Intentions to use the yelp review website and purchase behavior after reading reviews, *Journal of theoretical and applied electronic commerce research* 12 (1) (2017) 53–67.
- [61] B. D. Carpenter, S. Buday, Computer use among older adults in a naturally occurring retirement community, *Computers in Human Behavior* 23 (6) (2007) 3012–3024.
- [62] Quantcast, Audience insights (Apr 2020).
URL <https://www.quantcast.com/measure/profile/site/p-M4yfUTCpES3vn/yelp.co.uk?countryCode=GB&tab=demographics>
- [63] J. Johnson, Gb: online audience of yelp.co.uk by age group gender 2020 (Feb 2020).
URL <https://www.statista.com/statistics/491925/distribution-of-the-online-audience-of-yelpco.uk-in-gb-by-age-group-and-gender/>
- [64] P. Ruch, R. Baud, A. Geissbühler, Evaluating and reducing the effect of data corruption when applying bag of words approaches to medical records, *International Journal of Medical Informatics* 67 (1-3) (2002) 75–83.
- [65] B. C. How, K. Narayanan, An empirical study of feature selection for text categorization based on term weightage, in: *IEEE/WIC/ACM International Conference on Web Intelligence (WI'04)*, IEEE, 2004, pp. 599–602.
- [66] C. A. Martins, M. C. Monard, E. T. Matsubara, Reducing the dimensionality of bag-of-words text representation used by learning algorithms, in: *Proc of 3rd IASTED International Conference on Artificial Intelligence and Applications*, 2003, pp. 228–233.
- [67] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, R. Harshman, Indexing by latent semantic analysis, *Journal of the American society for information science* 41 (6) (1990) 391–407.
- [68] M. Honnibal, I. Montani, spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing, to appear (2017).
- [69] S. Yu, C. Xu, H. Liu, Zipf’s law in 50 languages: its structural pattern, linguistic interpretation, and cognitive motivation, *CoRR abs/1807.01855*. [arXiv:1807.01855](https://arxiv.org/abs/1807.01855).
URL <http://arxiv.org/abs/1807.01855>
- [70] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research* 12 (2011) 2825–2830.
- [71] M. Patterson, D. Silver, The place of art: Local area characteristics and arts growth in canada, 2001–2011, *Poetics* 51 (2015) 69–87.
- [72] M. Bowman, Artists in the canadian urban fabric (Jan 2013).
URL <http://spacing.ca/toronto/2010/03/17/artists-in-the-canadian-urban-fabric/>
- [73] D. Churchill, American expatriates and the building of alternative social space in toronto, 1965–1977, *Urban History Review/Revue d’histoire urbaine* 39 (1) (2010) 31–44.
- [74] D. Silver, The city is constantly making new versions of itself (Jul 2018).
URL <https://urbangenome.utoronto.ca/node/36>
- [75] J. MacQueen, et al., Some methods for classification and analysis of multivariate observations, in: *Proceedings of the fifth Berkeley symposium on mathematical statistics and*

- 905 probability, Vol. 1, Oakland, CA, USA, 1967, pp. 281–297.
- 906 [76] T. Caliński, J. Harabasz, A dendrite method for cluster analysis, *Communications in*
 907 *Statistics-theory and Methods* 3 (1) (1974) 1–27.
- 908 [77] S. Nanda, B. Mahanty, M. Tiwari, Clustering indian stock market data for portfolio man-
 909 agement, *Expert Systems with Applications* 37 (12) (2010) 8793–8798.
- 910 [78] N. Czink, P. Cera, J. Salo, E. Bonek, J.-P. Nuutinen, J. Ylitalo, A framework for automatic
 911 clustering of parametric mimo channel data including path powers, in: *IEEE Vehicular*
 912 *Technology Conference*, IEEE, 2006, pp. 1–5.
- 913 [79] S. Łukasik, P. A. Kowalski, M. Charytanowicz, P. Kulczycki, Clustering using flower pol-
 914 lination algorithm and calinski-harabasz index, in: *2016 IEEE Congress on Evolutionary*
 915 *Computation (CEC)*, IEEE, 2016, pp. 2724–2728.
- 916 [80] V. Satopaa, J. Albrecht, D. Irwin, B. Raghavan, Finding a” kneedle” in a haystack: Detect-
 917 ing knee points in system behavior, in: *2011 31st international conference on distributed*
 918 *computing systems workshops*, IEEE, 2011, pp. 166–171.
- 919 [81] S. Kay, Opinion: Yelp reviewers’ authenticity fetish is white supremacy in action (Jan
 920 2019).
 921 URL [https://ny.eater.com/2019/1/18/18183973/](https://ny.eater.com/2019/1/18/18183973/authenticity-yelp-reviews-white-supremacy-trap)
 922 [authenticity-yelp-reviews-white-supremacy-trap](https://ny.eater.com/2019/1/18/18183973/authenticity-yelp-reviews-white-supremacy-trap)
- 923 [82] L. Rainford, Toronto’s weston area revitalized by housing and transit developments (Oct
 924 2018).
 925 URL <https://www.toronto.com/news-story/8946043-toronto-s-weston-area-revitalized-by-hous>
- 926 [83] E. C. Delmelle, Differentiating pathways of neighborhood change in 50 us metropolitan
 927 areas, *Environment and planning A* 49 (10) (2017) 2402–2424.
- 928 [84] E. C. Delmelle, Mapping the DNA of urban neighborhoods: Clustering longitudinal se-
 929 quences of neighborhood socioeconomic change, *Annals of the American Association of*
 930 *Geographers* 106 (1) (2016) 36–56. doi:10.1080/00045608.2015.1096188.

931 **Appendix A. Yelp Methodology**

932 While Yelp does make available its own ‘academic dataset’ for research, this
 933 dataset contains only a subset of the businesses available in any given city. For
 934 this reason, we decided that it was necessary to obtain our own, more complete
 935 version of the Yelp dataset for one city. We chose Toronto in part due to our
 936 residence there, but also due to its substantial public dataset availability. The city
 937 of Toronto makes available an official set of neighbourhood shapefiles, defining
 938 the 140 neighbourhoods considered to exist by the city government (Figure 1).
 939 The scraping method for obtaining this data comprises a two-stage process. The
 940 first stage employs the official Yelp API to discover all businesses in Toronto,
 941 while the second stage uses a scraping script to obtain detailed data about these
 942 businesses.

943 Yelp exposes an official API for developers, which only provides basic infor-
 944 mation about businesses on the site. Critically, this API only provides the first
 945 twenty reviews for any business, selected using a proprietary Yelp relevancy al-

946 gorithm. The information provided by this API is sufficient to obtain a list of
947 all businesses in Toronto. However, it is not possible to simply pass an empty
948 search to Yelp to obtain a full list of businesses in a defined region. A more so-
949 phisticated approach is required for two reasons: (a) Yelp’s API requires some
950 text in the search field, and (b) Yelp only returns the first one hundred businesses
951 that match any query. In order to work within these constraints, the following
952 recursive method was created. For each category officially defined by Yelp (e.g.
953 ‘Chinese Restaurant’, ‘Rock Climbing’), search only the name of the category
954 using the API. Then, if the number of businesses returned is greater than one
955 hundred, sub-divide the boundaries into one hundred grid-based regions. Call
956 the function on each of these regions and append the output to the total list of
957 businesses. Otherwise, return the list of businesses returned by the API for the
958 query. Using this method, we obtained the Yelp listing for every business in
959 Toronto. The number of businesses obtained, approximately 24000, was vali-
960 dated against results from [7], in which they found that across the US, Yelp had
961 18.7% of the listings that the official business patterns dataset did. Our Toronto
962 dataset comprises a similar proportion of all businesses officially registered in
963 Toronto (which includes many non-consumer facing businesses that would not
964 be expected to have Yelp pages).