Computers, Environment and Urban Systems 111 (2024) 102131

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

A hybrid deep learning method for identifying topics in large-scale urban text data: Benefits and trade-offs

Madison Lore [a,*], Julia Gabriele Harten [a], Geoff Boeing [b]

[a] School of Community and Regional Planning, University of British Columbia, Vancouver, BC V6T 1Z2, Canada
[b] Department of Urban Planning and Spatial Analysis, University of Southern California, Los Angeles, CA 90089, United States

ARTICLE INFO

Keywords:
Urban informatics
Natural language processing
Deep learning
Text analysis
Qualitative methods

ABSTRACT

Large-scale text data from public sources, including social media or online platforms, can expand urban planners' ability to monitor and analyze urban conditions in near real-time. To overcome scalability challenges of manual techniques for qualitative data analysis, researchers and practitioners have turned to computer-automated methods, such as natural language processing (NLP) and deep learning. However, the benefits, challenges, and trade-offs of these methods remain poorly understood. How much meaning can different NLP techniques capture and how do their results compare to traditional manual techniques? Drawing on 90,000 online rental listings in Los Angeles County, this study proposes and compares manual, semi-automated, and fully automated methods for identifying context-informed topics in unstructured, user-generated text data. We find that fully automated methods perform best with more-structured text, but struggle to separate topics in free-flow text and when handling nuanced language. Introducing a manual technique first on a small data set to train a semi-automated method, however, improves accuracy even as the structure of the text degrades. We argue that while fully automated NLP methods are attractive replacements for scaling manual techniques, leveraging the contextual understanding of human expertise alongside efficient computer-based methods like BERT models generates better accuracy without sacrificing scalability.

1. Introduction

Los Angeles, California has become a poster child for the modern urban crises of housing shortages and unaffordability. In recent decades, the county's demographics have shifted rapidly while its housing supply has fallen far behind demand, pricing out residents and stymying inter-metropolitan in-migration (Zhu, Burinskiy, De la Roca, Green, & Boarnet, 2021). Desperate households have responded with a variety of coping strategies, creating a complex mosaic of novel housing arrangements (Angst, Rosen, De Gregorio, & Painter, 2023). Meanwhile, planners and policymakers have struggled to stay ahead of the crisis. On one hand, political maelstroms around densification, preservation, gentrification, and homelessness have complicated traditional planning processes (Dillon & Mejia, 2022). On the other hand, it is difficult to stay ahead of a crisis you cannot monitor.

Although tasked with managing housing affordability, planners often lack the data necessary to effectively assess and monitor local housing markets (Boeing, Wegmann, & Jiao, 2023). This poses a challenge to both cities' economies and its residents' lives and livelihoods. While recent research has made some inroads using quantitative data from online platforms like Craigslist to measure asking rents and unit characteristics in near real-time (Boeing & Waddell, 2017), these studies have typically failed to leverage such data sources' full information potential by forgoing difficult qualitative analysis of the rich user-generated text descriptions accompanying listings.

This is a familiar challenge to many urban scholars and practitioners: we can harvest and automatically analyze massive volumes of quantitative data, but qualitative data require overwhelming manual labor to extract insights at scale. Yet it is such qualitative data that can reveal new facets of the rental market, such as the often-subtle language of discrimination and the nature of housing conditions and affordability at neighborhood scales—the kinds of information planners need to intervene in a time of crisis to make housing more accessible (Adu & Delmelle, 2022; Boeing et al., 2023; Harten, Kim, & Brazier, 2021).

This study takes up this challenge. We propose and compare manual, semi-automated, and fully automated techniques to generate actionable insights from unstructured, user-generated text data. Using Los Angeles's housing crisis and rental market as a case study, we demonstrate

* Corresponding author at: Frederick Lasserre Building, 433 - 6333 Memorial Road, Vancouver, BC V6T 1Z2, Canada.
E-mail address: madison.lore@ubc.ca (M. Lore).

https://doi.org/10.1016/j.compenvurbsys.2024.102131
Received 31 January 2024; Received in revised form 15 May 2024; Accepted 18 May 2024

how and when computational techniques can generate insights on par with traditional manual techniques—but at a far larger scale and requiring far less labor. Analyzing 90,000 Craigslist rental listings in Los Angeles we find that fully automated techniques perform best with more-structured text, such as that posted by rental management companies, but struggle to capture subtle meanings in less-structured texts. However, integrating small amounts of manual hand-labeling early in the process yields a semi-automated technique better able to identify such nuance. We argue that important trade-offs exist when replacing manual techniques with computer-based techniques and propose guidelines to help practitioners and scholars harness this information landscape through mixed methods.

The remainder of this article is organized as follows. First, we review the recent use of text analysis in urban planning and highlight the need for hybrid computational-qualitative techniques. Next, we explain our research design, including data sources, processing, and analysis. Then we describe our findings and discuss the trade-offs of different techniques, the introduction of human bias, and recommendations for practice, before concluding.

2. Background

Urban text data is increasingly ubiquitous and is now being used in planning applications. For example, planning practitioners and scholars have leveraged online text data from formal sources such as community plans and policy documents (Brinkley & Stahmer, 2021; Fu, Li, & Zhai, 2023) and informal sources such as social media (Bronsvoort & Uitermark, 2022; Schweitzer, 2014).

The urban housing market offers a ready case for the use of online information. As housing searches move online, the internet offers a plethora of new data sources that planners can harvest to better understand housing phenomena (Boeing & Waddell, 2017; Gurran, Maalsen, & Shrestha, 2022; Harten et al., 2021). The rise of digital platforms in particular enables new housing research approaches and insights. Listings on online housing platforms usually have required text fields (structured data) as well as freeform descriptive text bodies (unstructured data). Recent research has used these structured data to reveal rapid changes to the housing market: for example, to assess rents and affordability (Boeing & Waddell, 2017), measure spatial variations in rental information availability (Adu & Delmelle, 2022), and identify new submarkets (Gurran et al., 2022; Harten et al., 2021).

However, the unstructured freeform text bodies could offer deeper insights into rental housing markets, but remain underutilized. Urban housing markets are changing at unprecedented rates, so understanding qualitative restrictions—beyond simple location, size, and affordability—could help planners intervene in housing crises. For example, exclusionary language and information inequalities can shape housing outcomes: the language used and features included in rental listings differ based on neighborhood racial composition (Boeing, Besbris, Schachter, & Kuk, 2021; Kennedy, Hess, Paullada, & Chasins, 2021). Freeform text is also key to unlocking home sharing, an increasingly important housing strategy currently shifting from sharing with family and friends to sharing among strangers (Maalsen, 2020; Parkinson, James, & Liu, 2021). While initial qualitative studies have shown that listers reveal their desired roommate traits and exclusionary criteria in selecting such roommates (Flage, 2018; Gaddis & Ghoshal, 2015), the extent of such exclusion has not been measured due to the labor- and time-intensive nature of this qualitative research.

Advances in computer-based text processing mean that large-scale data that traditionally would have been too tedious for manual labeling—and hence unavailable for analysis—can now be processed in record time. Whereas housing researchers have yet to fully harness freeform text, other urban researchers are already utilizing computer-assisted techniques, such as natural language processing (NLP), to generate insights through automated topic identification or sentiment analysis (Cai, 2021; Fu, 2024). Examples include the analysis of community plans to identify areas of emphasis (Brinkley & Stahmer, 2021; Fu et al., 2023) or building permit applications for changes in topics over time (Lai & Kontokosta, 2019). Other researchers have used user-generated or informal text such as travel experience reviews to investigate urban mobility issues (Serna, Gerrikagoitia, Bernabé, & Ruiz, 2017), online reviews for user sentiments towards neighborhoods (Hu, Deng, & Zhou, 2019), social media data to understand sentiments towards different areas of a city (Shelton, Poorthuis, & Zook, 2015), or constituent feedback and sentiments to local planning changes (Fu, Sanchez, Li, & Reu Junqueira, 2024).

However, in this space, the use of large-scale data sources and NLP tools is still emerging and standards for its use are not yet in place (Cai, 2021; Fu, 2024). NLP is often used as a full replacement for established manual techniques, such as qualitative thematic content coding. Yet, language is complex and its meanings and nuances can be lost in translation if not carefully processed (Mohr, Wagner-Pacifici, Breiger, & Bogdanov, 2013). Additionally, bias from the underlying data or the processing pipeline design can carry and amplify through successive modeling (Hovy & Prabhumoye, 2021). As NLP gains popularity in urban research, researchers must recognize the trade-offs between increased automation and the reliability of manual techniques. To date, however, it remains unclear how these trade-offs vary for different data sources or research objectives. Although planners have explored each end of this manual-to-automated spectrum, in the center lie underexplored hybrid techniques in which NLP complements manual techniques rather than replaces them.

3. Methods

This study proposes and compares techniques for identifying topics in unstructured, human-generated, text-based urban data. We assess if—and when—computer-based techniques can capture detail equivalent to traditional manual techniques that rely on human understanding of language but are too labor intensive for the scale of online data. As data increase in size and processing increases in complexity, we investigate the trade-offs between methods and propose metrics to quantify them. How much detail can different NLP techniques capture and how does this compare to traditional, qualitative thematic content coding? In particular, how can planners better utilize the hybrid techniques at the center of the manual-to-automated spectrum? We answer these questions using Los Angeles online rental listing data and assess analysis tools with regard to their ability to generate planning-relevant insights.

3.1. Data sources and preprocessing

Using a web scraper, we collected Craigslist rental listings in Los Angeles County between June 2020 and March 2021 from both the “apts / housing” and “rooms / shared” categories, which represent full units for rent and rooms in shared units for rent, respectively. The scraper collected 500 listings from each category at a random time each day to reduce time-of-day bias. Each listing includes fields manually entered by the lister, including square footage, price, number of bedrooms/bathrooms, and a freeform text box for detailed descriptions of the unit or living arrangement.

The initial collection yielded 255,436 listings. We then cleaned this data set to remove listings 1) outside of Los Angeles County (removed 4095 listings, roughly equal between listing types), 2) with duplicate listing IDs or body texts indicating the listing was reposted (removed 123,081 listings, majority from shared room listings), or 3) with implausible values in variables such as full unit rents above $30,000 or shared unit rents above $10,000 (removed 3928 listings, roughly equal between listing types). This yielded a data set of 124,332 listings comprising 76,406 full units for rent and 47,926 rooms in shared units for rent. For this study, we assess modeling techniques (detailed below) on these listing types separately, so we select equal-size random samples of each type for a final data set of 45,000 full units and 45,000 rooms in shared units.

We preprocess the rental listings' body text to remove extraneous symbols, convert all lettering to lowercase, remove stopwords, and lemmatize the text for consistency (Vijayarani, Ilamathi, & Nithya, 2015). Next, we separate the listings by sentence, defining sentence breaks by a tab, line or paragraph break, or common punctuation symbols (e.g., ".", ",", "!", "?", "|"). There are many ways to define breaks and we choose sentences here as most listings adhere to at least loose punctuation.

3.2. Topic identification

To generate insights from text data, topic identification (the process of identifying what the text is about) is a common task. We test four topic identification techniques ranging from fully manual to fully automated. This includes one common manual technique—qualitative thematic content coding (Williams & Moser, 2019)—and two common computer-based techniques—*k*-means and latent Dirichlet allocation (LDA). These latter two are fully automated unsupervised techniques that use NLP for topic identification, leveraging NLP's strength in modeling semantic relationships (Brinkley & Stahmer, 2021; Fu et al., 2023; Hu et al., 2019). In each of these three techniques, topical labels are assigned to a piece of text. In the manual technique, humans assign these labels based on their understanding of the text. The fully automated unsupervised techniques instead use algorithms to discover topics without human training and then assign labels. This is useful when "true" labels assigned by humans are not available, as is often the case with unstructured text data.

Finally, we also propose and test a fourth, novel computer-based technique—a hybrid (supervised) technique—that uses the manual technique labels as an input to finetune a pre-trained, deep learning, Bidirectional Encoder Representations from Transformers (BERT) model. Table 1 summarizes these four techniques and the following subsections describe them in detail.

3.2.1. Manual technique

The traditional manual technique of qualitative thematic content coding entails multiple researchers independently reading and hand-labeling text data to understand what it is about (topic identification). Here, two researchers read and hand-labeled a sample of listing text bodies following standard qualitative coding practices to understand what information listers provide in their rental listings. In particular, this involves iterative co-generation of labels from the data, testing for consistency in label interpretation and assignment, and multiple rounds of independent labeling until labels are established and labeling convergence was achieved (Williams & Moser, 2019).

Table 1
Summary of topic identification techniques.

Technique	Description and Coverage	Automation	Point of Human Intervention
Manual	Traditional qualitative thematic content coding with multiple independent coders to ensure rigor on small subset of listings	None	Generation of initial topics from reading, labeling of all text
k-means	Basic unsupervised technique to find *k* topics in a set of text data run on full set of listings	High	Thematic interpretation of automated topics
LDA	Sophisticated unsupervised technique to find common topics across a set of text data run on full set of listings	High	Thematic interpretation of automated topics
Hybrid	Supervised NLP technique using a BERT-based model trained on manually created labels from a small set, then run on the full set of listings	Medium	Generation of initial topics through manual technique

We perform the manual technique on two sets: a first set to identify the general topics that are commonly covered in all rental listings and a second set to identify relational qualities, such as desirable/undesirable tenant traits and behaviors, which generally appear in listings for shared units only. These two distinct sets allow us to test and contrast methods for topic identification both for more well-defined, general topics as well as topics based on nuanced language and subtle verbal cues. The analysis of the second set in particular is expected to be more challenging for computer-assisted methods, but humans (and in turn the manual technique) are good at extracting meaning from opaque or implicit language: this is the value of qualitative text analysis.

Given that the listings for full units versus rooms in shared units exhibit content differences, we randomly sample 100 listings from each to fill out our first set, totaling 2551 sentences. Because these listings will also be used as the training set in our hybrid technique, this sample size balances performance and pragmatic limits of manual labeling.[1] As language fluctuates much more when listers mention relational qualities, for our second set, we sample 3000 sentences isolated from nearly 3000 different room listing text bodies labeled as pertaining to relational qualities.

3.2.2. k-means topic identification

To prepare our text data for computer-based analysis, we transform the sentence-based text data into machine readable vectors through 1) count vectorization and 2) term frequency-inverse document frequency (TF-IDF). Count vectorization takes the full "dictionary" of words across the listings and counts the frequency of each word in each sentence. TF-IDF further represents the importance of each word in a sentence through assigning weights that take into account both its frequency within the sentence and its frequency across all sentences (Uther et al., 2011). If a word appears in a sentence but is otherwise uncommon across all sentences, TF-IDF ascribes it a higher importance in understanding that sentence's topic. For example, "oceanfront" may not be common across all sentences, so it would signify an important word in a sentence that could describe location. Conversely, "room" may be common across many rental listing sentences and would hold low importance in each sentence in which it appeared.

Once the data is transformed in this way, it can be used as the input into a *k*-means algorithm that generates *k* topics of data by first randomly choosing *k* points to act as topic centers, then assigning every other point to the closest topic, in which the distance is that between the center and each TF-IDF vector representing a sentence. The center of each topic is updated after all of the data points, representing sentences, have been assigned to a topic and the process iterates until the topic assignments no longer change. The model outputs a topic label for each sentence in our full data set of 90,000 listings. This is akin to grouping sentences by similarity of words, but human interpretation is still needed to identify what real-world topic each label represents by looking across the label's sentences.

3.2.3. Latent Dirichlet Allocation (LDA)

As with *k*-means, LDA uses the initial count vectors, which represent the frequency of each word in each sentence. Unlike *k*-means, however, LDA inherently determines weights and importances of the words in each sentence so only counts (and not TF-IDF) are needed (Blei, Ng, & Jordan, 2003). With the data transformed into the count vectors, an LDA model works iteratively to determine both a set of words that comprise each topic and which topics are present in each sentence. The LDA

[1] Studies have found that as few as 100 samples are necessary for training to reach acceptable standards of error, with more samples needed if the set of topics are rarer (Hopkins & King, 2010).

model outputs, for each sentence, a k-length set of probabilities from 0 to 1 representing the probability of that sentence belonging to each topic. One sentence can have a high probability of belonging to more than one topic. For example, an LDA model might determine that Topic 1 is composed of words like "patio", "deck", and "balcony" while Topic 2 is composed of words like "hardwood", "stone", and "concrete". The model would also estimate that the sentence, "There is a large balcony off the bedroom" contains Topic 1 with a high probability, while the sentence "Includes hardwood floors and a stone patio" would have a high probability of containing both Topic 1 and Topic 2. Because this is the final output of the model, human interpretation is still needed to explain what the topic might be representing.

3.2.4. Hybrid (supervised) technique

Finally, we introduce a novel hybrid technique combining elements of both the manual technique and machine-based text data analysis. This method uses the output of the manual technique as labeled training data for a supervised NLP model. In particular, we split the hand-labeled sentences into training and testing sets. The training set is used to update (i.e., finetune) a pre-trained deep learning BERT model,[2] while the testing set is held aside to evaluate the model after training has completed. Pre-trained language models learn word representations and relationships from a large corpus of text without a specific task in mind and can be finetuned with a smaller amount of labeled data for a particular task, such as identifying topics. Here, we utilize a BERT model designed for sequence classification tasks (Devlin, Chang, Lee, & Toutanova, 2018). BERT models excel at understanding nuanced phrases by considering the context of the surrounding sentence when learning relationships between the words. Because it is pre-trained on a large corpus of sentences and contexts, variations in language use have been shown to be adequately captured in downstream analytics (Devlin et al., 2018; Peinelt, Nguyen, & Liakata, 2020; Qasim, Bangyal, Alqarni, & Ali Almazroi, 2022).

We add a classification head to the model for our specific data purpose and labels. This head uses binary cross-entropy loss to evaluate how well the model learns by minimizing the *dissimilarity* between predicted labels and true labels over multiple iterations (Liu & Qi, 2017). Additionally, the model uses a learning rate of 1×10^{-5} and the Adam optimizer with weight decay to update the model weights during training. Over five epochs of four batches each, the model is updated to learn relationships between the manually created labels and the listing text bodies while utilizing the strong understanding of relationships between words themselves from the pre-trained model. This model is a multilabel classifier, so each sentence can have one or more topics present. This means that for each sentence, the model outputs the probability from 0 to 1 of belonging to each label generated through the manual technique.

3.3. Time and complexity analysis

To measure the trade-offs across this spectrum of topic identification methods, we measure time and complexity costs for each. First, we calculate each method's total runtime. For the manual technique, we calculate this based on the total number of words to read given an average reading speed of 238 words per minute (Brysbaert, 2019). This does not account for the additional time spent on generating and establishing the labels, assigning labels to sentences, and comparing results between coders, as these tasks can vary greatly with the researcher(s) and the complexity of the text data and labels. Instead, relying on reading speed alone yields a conservative yet robust estimate. For each computer-based technique, we calculate the real elapsed runtime from start to end. As with the manual technique, here, we do not include the time to write or test the code necessary to execute these techniques.

3.4. Individual performance metrics

Next, we compute individual measures of accuracy for each technique. For the manual technique, we calculate two interrater reliability metrics at the label level: general percent agreement and Cohen's kappa. The former measures the number of times both researchers agreed on the binary classification (yes or no) of a topic's presence in each sentence divided by the total number of sentences. The latter accounts for agreement due to random chance and includes terms for the actual agreement and the estimated agreement if both researchers were to randomly guess each label (McHugh, 2012). Together these metrics measure the consistency of labeling across independent coders and establish reliability of the labels.

Each computer-assisted method is constructed slightly differently, so we rely on distinct, but analogous measures of individual performance to understand how well the topics generated by each method describe the data points that lie within them. For k-means, we compute the *silhouette score* (ranging from -1 to 1) to measure how well each point fits within its labeled topic. Scores closer to 0 indicate poorer performance and suggest that the topics are heavily overlapping. For LDA, we compute the *coherence score*, which measures how closely related the words in each topic are to the words in other topics. While coherence scores do not have a set minimum, a score that is negative or closer to zero indicates higher overlap in the words underlying each topic.

For the hybrid technique, we calculate the *precision, recall, and F1 score* of the assigned sentence labels against the manually labeled testing set as this technique uses the output from the manual technique as the "true label". Precision (the positive predictive value) measures what share of all true + false positives are true positives: that is, out of the labels generated for each sentence by the hybrid method, how many matched the "true label" for each sentence? Recall (the true positive rate) measures what share of all true positives + false negatives are true positives: that is, out of the "true labels" for each sentence, how many of the hybrid method labels matched? The F1 score is the harmonic mean of precision and recall.

3.5. Comparing across techniques

To evaluate and compare how well each technique captures the nuance of complex texts at scale, for each topic, we calculate the level of agreement (LOA) between (i) the hand-labeled data generated via the manual technique, taken here as the "gold standard" and (ii) those produced via the other techniques. Similar to interrater reliability scoring, one can think of this level of agreement measure as asking: if one coder were human and the other a computer-based technique, how well would they agree on the classification of each sentence against a defined codebook?

To facilitate comparisons, for the fully automated unsupervised techniques (i.e., k-means and LDA) we generate the same number of topics as were generated through the manual technique. Then we compute the percent overlap of sentences belonging to each generated topic and each manually labeled topic. The output is a matrix of how many sentences are shared between each of the generated topics and the manually labeled topics. For example, for the first set of six general topic labels, we have a 6×6 matrix in which each value is the percent of sentences shared between i-th LDA (or k-means) topic and the j-th manually labeled topic. To calculate LOA we need to find a one-to-one match between the two sets of labels to prevent one all-encompassing topic having artificially high agreement with all of the manually labeled topics. For this, we use the Hungarian algorithm to find the maximum overlap which adds the constraint that each generated topic can only match one to one with each manually labeled topic (Kuhn,

[2] Many pre-trained models are available at https://huggingface.co. For this model, we used the "bert-based-uncased" model available at: https://huggingface.co/google-bert/bert-base-uncased

1955). Now, each generated topic is matched to a "best fit" hand-labeled topic and we compute the recall. For the hybrid technique, because each sentence is inherently labeled according to the topics from the manually labeled set, the labels are already matched one-to-one and we simply compute the recall.

Finally, we examine the most common or important words for each topic generated from each of the computer-based techniques. For each technique, these words are drawn out and interpreted to label the topic. This step is necessary for fully automated unsupervised techniques where the determined topics are not labeled a priori but left for human interpretation after the fact. For the hybrid technique, these words reveal how closely our understanding of the words aligns with the expected topic label. Together, an evaluation of the top words from each topic gives qualitative meaning to the level of agreement values computed for each technique. If a technique has a low LOA, this might be supported by a set of words that differs greatly from those most important to the manually labeled topics.

4. Findings

4.1. Topic identification

Our manual technique for topic identification reveals six general topics present across rental listings (i.e., set 1, manual technique): information about the 1) unit, 2) location, 3) logistics of renting, 4) personal qualities of the prospective/current renter/roommate, 5) restrictions on renter behavior, and 6) privacy specifications. The first three topics describe "environmental conditions" of the listing whereas the latter three describe "relational conditions," and the final topic is particularly relevant for shared rentals.

While over 87% of both full unit and shared unit listings contain descriptions of environmental conditions, just 45% of full units describe relational conditions whereas 92% of rooms in shared units do. Focusing on relational conditions only, we identify a set of finer-grained topics (i.e., set 2, manual technique) detailing the specific information listers provide when seeking someone to share a unit. We determine this second set of sentences by utilizing the labels from the hybrid method applied to the first set of general topics. Some listings describe current or desired tenants, including their age (11%), gender (17.5%), or sexuality (2%), some of which are protected classes under the US Fair Housing Act. Many listings specify lifestyle restrictions like no smoking, drugs, or pets (29.5%), or require full-time professional employment (26.3%) or proof of income covering rent and expenses (16.9%).

4.2. Time and complexity

A key drawback of the manual technique is the amount of time needed to process large-scale text data. In our case, one person simply reading through the entire set of 90,000 listings would take 1144 h, or the equivalent of 143 8-h workdays. As described earlier in section 3.2.1., identifying relational conditions (set 2) involves more nuance and more subtle verbal cues in comparison to identifying general topics (set 1). We do not explicitly benchmark the time to create and identify the thematic coding labels, but the time scales with label complexity.

The time required to run the computer-assisted techniques on the other hand is significantly lower (Table 2). Instead of days, the processing time for all 90,000 listings is on the order of minutes for *k*-means and LDA and on the order of minutes to single digit hours for the hybrid technique after initial training. Each technique's processing time grows linearly with listing count, but the computer-based techniques have much lower overall processing times. For the hybrid technique there is an upfront time cost for labeling the training data (here, approximately two hours to read the training set at 238 words per minute and 5 min to train the finetuned model using a Tesla T4 GPU through Google Cloud Engine). After that processing time increases with listings count at 39 milliseconds/listing. Processing times for both the general topics and the relational qualities are similar as the computer-based techniques are less dependent on the number of topics.

Table 2
Processing time for each technique on the first labeled set by number of listings and words.

Listing count	Avg Word Count	Manual (hours)	*k*-means (sec)	LDA (sec)	Hybrid (training + runtime)
100	18,972	1.3	0.3	0.9	–
1000	175,224	12	2.5	6.2	2.5 h + 0.6 min
10,000	1,793,908	126	8.4	51.9	2.5 h + 6.5 min
90,000	16,337,796	1144	232.0	480.0	2.5 h + 60.0 min

4.3. Individual performance accuracy

Interrater reliability analysis reveals an average percent agreement of 95% and an average Cohen's Kappa of 84.3% across the first set (general topics). The average percent agreement is 94.5% across full unit listings and 95% across rooms in shared unit listings, and average Cohen's Kappa is 79% and 86%, respectively. When testing the interrater reliability metrics of the second set of labeled listings (relational qualities), the average percent agreement is 99% and the average Cohen's Kappa is 93%.

Across all listings for the first set of general topics, the *k*-means silhouette score was 0.019. Similar values result from separating the data by listing type, with the full unit listings only slightly higher than the rooms in shared unit listings (0.023 and 0.015, respectively). Applying *k*-means to identify the second set (relational qualities) yields a comparable silhouette score of 0.027.

For LDA, we compute coherence scores. For full unit and shared unit listings together for the general topics, the coherence score is 0.481. For the same set, the full unit listings and shared unit listings separately have scores of 0.603 and 0.530, respectively. For relational qualities, the coherence score is slightly lower, 0.348.

The hybrid technique's precision for the general topics is 81%, its recall is 84%, and its F1 score is 82%. The values for the full unit listings and rooms in shared unit listings are similar, with the full unit listings slightly higher compared to the shared unit listings. For the relational qualities, the hybrid technique has a precision of 80%, a recall of 77%, and an F1 score of 78%.

4.4. Comparing across techniques

Next, we compute the level of agreement (LOA) for the general topics and all computer-assisted methods vs the manual approach. For the full unit listings (which generally have a higher degree of structure) *k*-means topics have only a 25% agreement, while LDA topics have 81%. The hybrid technique exhibits 86% agreement with the labels obtained through the manual technique. In other words, *k*-means labels just 25% of the sentences "correctly", whereas LDA and the hybrid technique achieve "correct" labels 81% and 86% of the time, respectively. We also compute these measures for the rooms in shared unit listings. Here, the language was generally less structured and more heterogeneous. For the shared unit listings, the values drop to 18% agreement for *k*-means, 74% for LDA, and 83% for the hybrid technique.

Finally, we compute the LOA for relational qualities in shared unit listings, which involves the most nuanced and varied language. The values are 25% agreement for *k*-means, 32% for LDA, and 77% for the hybrid technique. To buttress these findings qualitatively, Table 3 presents the 15 most common words in each general topic generated via the different techniques, from both full and shared unit listings. The manual technique generates expected results under each topic, such as words like "close", "located", and "beach" describing the location of the listing and "respectful", "clean", and "roommate" describing personal qualities

Table 3
Most common 15 words (as exemplars) in each topic for each of the compared techniques, plus our interpretation of the topic accordingly.

Manual	*k*-means	LDA	Hybrid
Manual label: Unit	*Generated Topic 1*		
room, apartment, floor, parking, bedroom, kitchen, amenity, closet, laundry, unit, pool, home, community, feature, new	room, month, private, rent, house, utilities, shared, bathroom, included, available, furnished, bedroom, quiet, deposit, clean	room, rent, utilities, bathroom, bedroom, parking, month, available, private, looking, please, kitchen, deposit, furnished, clean	apartment, room, bedroom, kitchen, floor, parking, unit, home, new, private, closet, living, bathroom, area, large
Manual label: Location	*Generated Topic 2*		
located, apartment, los, location, angeles, close, city, freeway, minute, restaurant, neighborhood, ca, park, hollywood, downtown	apartments, community, apartment, contact, center, amenities, los, angeles, access, show, info, fitness, home, call, pet	los, angeles, center, us, amenities, community, downtown, contact, home, la, hour, offer, show, district, info	los, angeles, located, park, close, city, hollywood, minute, freeway, ca, home, restaurant, center, access, beach
Manual label: Logistics	*Generated Topic 3*		
contact, month, info, show, rent, please, deposit, call, lease, utility, available, move, tour, text, room	contact, room, info, show, rent, bedroom, de, call, private, house, available, bath, parking, bathroom, please	housing, info, application, show, contact, fee, property, income, deposit, id, lease, call, check, monthly, bathrooms	contact, month, apartment, rent, show, info, pet, available, room, call, deposit, please, bedroom, unit, lease
Manual label: Personal Qualities	*Generated Topic 4*		
roommate, clean, respectful, looking, work, responsible, home, female, quiet, someone, time, male, professional, one, working	room, looking, house, month, bedroom, rent, roommate, bathroom, apartment, clean, available, please, private, someone, also	contact, show, info, air, apartment, parking, access, call, laundry, amenities, community, gated, dishwasher, refrigerator, ca	looking, clean, roommate, home, work, room, time, professional, someone, quiet, friendly, house, respectful, like, female
Manual label: Restrictions	*Generated Topic 5*		
pet, credit, deposit, allowed, smoking, income, rent, dog, cat, must, friendly, check, female, please, policy	unit, new, building, parking, contact, info, stove, kitchen, show, floors, bedroom, apartment, laundry, large, deposit	appliances, stainless, pool, steel, community, amenities, views, living, features, washer, home, bedroom, outdoor, style, center	pet, credit, smoking, month, income, check, please, room, rent, drug, must, application, year, deposit, person
Manual label: Privacy	*Generated Topic 6*		
private, room, bathroom, shared, bedroom, kitchen, access, house, master, space, roommate, area, full, share, available	access, amenities, select, apartment, community, air, availability, call, disposal, gated, controlled, conditioning, units, laundry, refrigerator	show, info, contact, de, el, la, se, en, para, renta, con, una, cuarto, college, persona	private, room, bathroom, shared, bedroom, rent, one, bath, available, house, bed, month, share, kitchen, furnished

of the current or prospective tenant. Since the hybrid technique builds on these results, its top words in each topic demonstrate only slight variations in vocabulary and order of importance.

The fully automated unsupervised techniques require deeper interpretation and do not necessarily align with the topics identified via the manual technique. For LDA, words in each topic are generally separable, but (near) synonyms appear across topic groupings such as "room", "house", and "home". Topics 1 and 5 generally describe aspects of the rental listing, with topic 5 offering more specific details on the apartment space. Topics 3, 4, and 6 include details on the logistics of renting. Topic 2 generally describes the listing's surrounding location, but also includes logistical elements. The *k*-means words in each topic are not as separable and feature many common words and (near) synonyms such as "bedroom", "bathroom", and "home." Each of the *k*-means topics describes rental listings in vague terms—distinct interpretations are not readily observable.

5. Discussion

5.1. Summary

Planners and policymakers need a stronger evidence base for interventions into modern urban crises, such as the housing crisis in Los Angeles. Although the housing market's quantitative data exhaust provides some empirical footing, a rich mine of qualitative data remains untapped due to the practical time and labor limits of processing huge text datasets manually, heretofore the gold standard for rigor and accuracy. Accessing the information in these large informal data sources could provide planners with important insights, for instance regarding discrimination or emerging housing vulnerabilities in near real time. Conversely, fully automated, unsupervised computer-based techniques are becoming more accessible and widely used, but they perform best on highly structured texts, which rental listings and many other forms of user-generated online data are not.

This study collected a large dataset of online rental listings to propose a new approach to scaling qualitative research in an era of rapidly advancing artificial intelligence tools. While this exact dataset itself cannot be re-hosted due to terms of use constraints, online archival tools, including the Internet Archive's Wayback Machine (https://web.archive.org/), collect snapshots of Craigslist that could reproduce a dataset matching this time period. Our study explores a spectrum of techniques from fully manual to fully automated to identify strengths and weaknesses of each. It then proposes a novel hybrid supervised technique offering the best of both worlds: manual rigor augmented by automated scalability. Using these rental listings, we demonstrate the trade-offs of these techniques across two key dimensions: time and accuracy.

For example, while systematically analyzing the content of the listings would provide a deeper understanding of the nuances of the rental housing market, manually labeling the full dataset of 90,000 listings would take thousands of hours. Meanwhile the fully automated techniques (i.e., LDA and *k*-means) can complete the task in just a few minutes. This affords researchers more time for finetuning models, interpreting results, and performing subsequent analyses—which they will need because these techniques yield far worse accuracy than the manual technique does. In particular, we find that the *k*-means models are generally unfit for identifying either general or more fine-grained topics. While the LDA models perform better, performance degrades when tasked with labeling less structured text and more complex meaning, such as listings for rooms in shared units and for identifying relational qualities.

Our proposed hybrid supervised technique exists in the middle of this spectrum. To re-summarize, we generate topic labels through a traditional, qualitative, manual thematic content coding process then use these as training data in a supervised NLP model. This employs a pre-trained BERT model (designed for sequence classification tasks and fit with an additional classification head for the specific tasks at hand) which we fine-tune with the labeled data. This requires upfront processing time (to hand-label its training data) that the fully automated techniques do not. However, compared to the manual technique, the full processing time is much lower even as the runtime scales with the number of listings. With just over two hours of manual labeling and five minutes of training, our hybrid technique achieved 78–82% accuracy levels across a variety of tasks, far exceeding the fully automated techniques. By design, the method's topic identification also aligns well with that of the manual technique.

LOA offers another way to measure performance of the computer-assisted methods relative to the gold standard, the manual technique hand labels. The hybrid technique has the highest level of agreement, followed by LDA, then *k*-means. Notably, structure determines the topic models' usefulness: LOA declines (particularly for LDA and *k*-means) when moving from listings for full units, to rooms in shared units, to only sentences describing relational conditions. Unsupervised techniques are fast and excel with structured and separable texts, but struggle with nuance and implicit meaning.

5.2. Human Bias

Manual labeling remains the gold standard of topic identification. However, it relies on human intervention and experience, which inherently introduce human bias. Superficially, the fully automated methods seem to mitigate this problem. However, we argue that each of these techniques merely shifts the point at which human bias can enter the analysis—either at the data gathering, topic determination, or sentence labeling stages. Understanding the scope and potential for bias is critical when deciding which technique to employ.

At the *data gathering* stage, human bias enters the analysis from the lister rather than the researcher. Any analysis using online, volunteered information inherits biases introduced through factors like technology access, language, and communication. Asking questions such as "Who uses this platform?" and "Who is the lister signaling to attract or exclude?" are essential to understand the data's quality. Over- and under-representation of certain groups affects downstream analysis. Importantly, revealing this bias is exactly the kind of value-add that processing text data for qualitative analysis can deliver. Such nuance can only be extracted from freeform text, rather than structured user input, and unlocks new possibilities for housing research on equity and discrimination.

The second stage, *topic determination* happens *before* the sentences are labeled in the manual and hybrid techniques, but *after* the sentences are labeled by LDA and *k*-means. Although the sequencing differs, a human intervenes at this stage in all techniques.

By design, the manual and hybrid techniques share the same topic determination stage. When human intervention occurs early in the process, it offers a chance to add context, nuance, and meaning before each sentence is labeled either by hand or computer. Even when researchers follow established standards to minimize human bias, such as multiple independent coders, the output's quality is only as good as the quality of the topics. This quality measures the topics' ability to adequately capture the meaning of the text data. Meanwhile, LDA and *k*-means generate topics without human input up to this stage, allowing for a less-biased determination of common topics. However, to draw meaningful conclusions, a human must examine each set of words to determine its topic, again bringing in their own bias. Importantly, unlike in the manual technique, standards for reducing bias in interpreting topics have not been established.

Finally, the *sentence labeling* stage happens *after* the topic determination in the manual and hybrid techniques, but *before* it in LDA and *k*-means. In the manual technique, this is usually performed by multiple researchers to reduce the effect of a single person's bias but can also introduce inconsistencies in labeling between coders. However, given a robust codebook, this process is rigorous and typically able to capture the nuance and complexity of text data well (MacQueen, McLellan, Kay, & Milstein, 1998). For all computer-based techniques, this stage is automated.

5.3. Recommendations

Planners must navigate a complicated methodological landscape to leverage the power of large-scale text data. We argue that the choice of techniques is context-dependent and varies with the underlying data, analytical goals, and researchers themselves.

First, the underlying data's volume and velocity largely determine whether computer-based techniques are necessary. High-velocity text data sources (such as new rental listings added to Craigslist every day) benefit from computer-based techniques keeping pace with new listings using the same models with minimal additional processing time.

Second, the data's structure should influence the technique choice. More structured and formalized data work better with fully automated unsupervised techniques to deliver reliable, well-defined, meaningful topic identification. As the formality or structure of the text degrades, introducing human interpretation earlier becomes important. Human intervention at the topic determination stage, before sentence labeling (such as in the hybrid technique), leverages superior human text processing capabilities to more reliably identify subtlety and nuance.

Third, researcher time and skill may tip the scales towards or away from automation. Computer-based techniques have high barriers to entry, as they require programming knowledge. However, this barrier continues to fall as more online tools appear and basic aptitude in programming become commonplace. While limited technical capacity may favor more manual techniques, lower time availability favors automation.

5.4. Future agenda

This study opens the door to further research into when and how to best utilize emerging NLP techniques. First, here we only examine a handful of computer-based topic identification techniques and future work should develop best practices for additional techniques.

Second, replicating this work on non-English, shorthand, and other kinds of text will demonstrate these techniques' feasibility or lack thereof across a more diverse set of contents and structures. This seems especially important as this study highlights how the text's structure and nuance heavily influence the computer-based techniques' performance. Additionally, many models, especially pre-trained language models, are built on formal English text. Continuing to build models on such text widens the accuracy gap for text in other languages or less-formal capacities, leading to a spiral of unintended biases (Hovy & Prabhumoye, 2021). Developing and testing new models for nontraditional, complicated, or non-English language text helps attenuate such bias.

Finally, the hybrid technique is only as good as the hand-labeled training data on which it is built. Examining the positionality of the researchers who perform the manual labeling process can reduce differences between the social norms of the lister and of the coder, as well as potentially enhance insights from the data themselves.

6. Conclusion

Large-scale text data can shed new light on evolving social processes, such as who is implicitly welcomed or excluded in housing listings. However, traditional manual analysis techniques quickly become infeasible when facing the enormous volumes of unstructured text-based digital content now available. This study investigated emerging techniques for computer-based topic identification and introduced a novel hybrid technique in which manual technique labels inform a supervised model to identify topics in online rental listings' freeform body text. Without sacrificing computer-based techniques' processing power, the hybrid technique incorporates all-important human expertise early on to perform at large-scale with high accuracy, importantly in capturing the nuance of unstructured, freeform text. However, no technique across the spectrum from fully manually to fully automated can escape human bias. We argue that text-processing models should not strive to be fully automated, but instead utilize both computer processing power for handling large-scale datasets as well as human cognition to unpack the assumptions of the modeling process at each stage.

Funding

Social Science and Humanities Research Council of Canada's Canada Research Chairs Program

CRediT authorship contribution statement

Madison Lore: Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Julia Gabriele Harten:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Geoff Boeing:** Writing – review & editing, Writing – original draft, Supervision, Investigation, Data curation.

Declaration of competing interest

none.

Acknowledgements

Dr. Harten gratefully acknowledge funding from the Canada Research Chair Program.

References

Adu, P., & Delmelle, E. (2022). Spatial variations in exclusionary criteria from online rental advertisements. *The Professional Geographer, 74*, 1–11. https://doi.org/10.1080/00330124.2022.2061537

Angst, S., Rosen, J., De Gregorio, S., & Painter, G. (2023). How do renters survive unaffordability? Household-level impacts of rent burden in Los Angeles. *Journal of Urban Affairs*, 1–24.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*(Jan), 993–1022.

Boeing, G., Besbris, M., Schachter, A., & Kuk, J. (2021). Housing search in the age of big data: Smarter cities or the same old blind spots? *Housing Policy Debate, 31*(1), 112–126. https://doi.org/10.1080/10511482.2019.1684336

Boeing, G., & Waddell, P. (2017). New insights into rental housing markets across the United States: Web scraping and analyzing craigslist rental listings. *Journal of Planning Education and Research, 37*(4), 457–476. https://doi.org/10.1177/0739456X16664789

Boeing, G., Wegmann, J., & Jiao, J. (2023). Rental housing spot markets: How online information exchanges can supplement transacted-rents data. *Journal of Planning Education and Research, 43*(3), 525–537. https://doi.org/10.1177/0739456X20904435

Brinkley, C., & Stahmer, C. (2021). What is in a plan? Using natural language processing to read 461 California city general plans. *Journal of Planning Education and Research, 44*(2), 632–648, 0739456X21995890.

Bronsvoort, I., & Uitermark, J. L. (2022). Seeing the street through Instagram. Digital platforms and the amplification of gentrification. *Urban Studies, 59*(14), 2857–2874.

Brysbaert, M. (2019). How many words do we read per minute? A review and meta-analysis of reading rate. *Journal of Memory and Language, 109*, Article 104047. https://doi.org/10.1016/j.jml.2019.104047

Cai, M. (2021). Natural language processing for urban research: A systematic review. *Heliyon, 7*(3). https://doi.org/10.1016/j.heliyon.2021.e06322

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. In *, 1 (Long and Short Papers). In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics.

Dillon, L., & Mejia, B. (2022, October 19). Why it's so hard to fix housing overcrowding in Los Angeles. *Los Angeles Times*. https://www.latimes.com/homeless-housing/story/2022-10-19/overcrowding-los-angeles-housing-fix.

Flage, A. (2018). Ethnic and gender discrimination in the rental housing market: Evidence from a meta-analysis of correspondence tests, 2006–2017. *Journal of Housing Economics, 41*, 251–273.

Fu, X. (2024). Natural language processing in urban planning: A research agenda. *Journal of Planning Literature, 0*(0), 08854122241229571.

Fu, X., Li, C., & Zhai, W. (2023). Using natural language processing to read plans. *Journal of the American Planning Association, 89*(1), 107–119. https://doi.org/10.1080/01944363.2022.2038659

Fu, X., Sanchez, T. W., Li, C., & Reu Junqueira, J. (2024). Deciphering public voices in the digital era: Benchmarking ChatGPT for analyzing citizen feedback in Hamilton, New Zealand. *Journal of the American Planning Association*, 1–14.

Gaddis, S. M., & Ghoshal, R. (2015). Arab American housing discrimination, ethnic competition, and the contact hypothesis. *The Annals of the American Academy of Political and Social Science, 660*(1), 282–299. https://doi.org/10.1177/0002716215580095

Gurran, N., Maalsen, S., & Shrestha, P. (2022). Is 'informal'housing an affordability solution for expensive cities? Evidence from Sydney, Australia. *International Journal of Housing Policy, 22*(1), 10–33.

Harten, J. G., Kim, A. M., & Brazier, J. C. (2021). Real and fake data in Shanghai's informal rental housing market: Groundtruthing data scraped from the internet. *Urban Studies, 58*(9), 1831–1845.

Hopkins, D. J., & King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science, 54*(1), 229–247. https://doi.org/10.1111/j.1540-5907.2009.00428.x

Hovy, D., & Prabhumoye, S. (2021). Five sources of bias in natural language processing. *Language and Linguistics Compass, 15*(8), Article e12432.

Hu, Y., Deng, C., & Zhou, Z. (2019). A semantic and sentiment analysis on online neighborhood reviews for understanding the perceptions of people toward their living environments. *Annals of the American Association of Geographers, 109*(4), 1052–1073. https://doi.org/10.1080/24694452.2018.1535886

Kennedy, I., Hess, C., Paullada, A., & Chasins, S. (2021). Racialized discourse in Seattle rental ad texts. *Social Forces, 99*(4), 1432–1456. https://doi.org/10.1093/sf/soaa075

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly, 2*(1–2), 83–97.

Lai, Y., & Kontokosta, C. E. (2019). Topic modeling to discover the thematic structure and spatial-temporal patterns of building renovation and adaptive reuse in cities. *Computers, Environment and Urban Systems, 78*, Article 101383.

Liu, L., & Qi, H. (2017). Learning effective binary descriptors via cross entropy. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 1251–1258). https://doi.org/10.1109/WACV.2017.144

Maalsen, S. (2020). 'Generation share': Digitalized geographies of shared housing. *Social & Cultural Geography, 21*(1), 105–113.

MacQueen, K. M., McLellan, E., Kay, K., & Milstein, B. (1998). Codebook development for team-based qualitative analysis. *Cam Journal, 10*(2), 31–36.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*(3), 276–282.

Mohr, J. W., Wagner-Pacifici, R., Breiger, R. L., & Bogdanov, P. (2013). Graphing the grammar of motives in National Security Strategies: Cultural interpretation, automated text analysis and the drama of global politics. *Poetics, 41*(6), 670–700.

Parkinson, S., James, A., & Liu, E. (2021). Luck and leaps of faith: How the digital informal economy transforms the geographies of shared renting in Australia. *Social & Cultural Geography, 22*(9), 1274–1290.

Peinelt, N., Nguyen, D., & Liakata, M. (2020). tBERT: Topic models and BERT joining forces for semantic similarity detection. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics* (pp. 7047–7055).

Qasim, R., Bangyal, W. H., Alqarni, M. A., & Ali Almazroi, A. (2022). A fine-tuned BERT-based transfer learning approach for text classification. *Journal of Healthcare Engineering*, 2022.

Schweitzer, L. (2014). Planning and social media: A case study of public transit and stigma on twitter. *Journal of the American Planning Association, 80*(3), 218–238. https://doi.org/10.1080/01944363.2014.980439

Serna, A., Gerrikagoitia, J. K., Bernabé, U., & Ruiz, T. (2017). Sustainability analysis on urban mobility based on social media content. *Transportation Research Procedia, 24*, 1–8.

Shelton, T., Poorthuis, A., & Zook, M. (2015). Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information. *Landscape and Urban Planning, 142*, 198–211. https://doi.org/10.1016/j.landurbplan.2015.02.020

Uther, W., Mladenić, D., Ciaramita, M., Berendt, B., Kołcz, A., Grobelnik, M., , … Zheng, F., et al. (2011). TF–IDF. In C. Sammut, & G. I. Webb (Eds.), *Encyclopedia of machine learning* (pp. 986–987). Springer US. https://doi.org/10.1007/978-0-387-30164-8_832.

Vijayarani, S., Ilamathi, M. J., & Nithya, M. (2015). Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks, 5*(1), 7–16.

Williams, M., & Moser, T. (2019). The art of coding and thematic exploration in qualitative research. *International Management Review, 15*(1), 45–55.

Zhu, L., Burinskiy, E., De la Roca, J., Green, R. K., & Boarnet, M. G. (2021). Los Angeles' housing crisis and local planning responses. *Cityscape, 23*(1), 133–160.