# A BERT-Based Approach for Gender Inference from Place Reviews: Applications to Urban Representativeness

**Jemal Abate**[1], **Felipe Peixoto**[1], **João Bald**[1],
**Myriam Delgado**[1], **Thiago H. Silva**[1]

[1]Universidade Tecnológica Federal do Paraná (UTFPR)
Departamento Acadêmico de Informática
Curitiba, Brasil

jemalabatejilo,felipefiaspeixoto,bald@alunos.utfpr.edu.br
myriamdelg,thiagoh@utfpr.edu.br

***Abstract.*** *User-generated content from location-based platforms provides valuable insights into urban behavior, but the lack of explicit demographic information limits analyses of social representativeness. In particular, understanding gender differences in the use of urban space remains challenging due to the absence of structured user attributes. In this work, we investigate the use of natural language processing techniques to infer binary gender labels from textual reviews in Google Places. We evaluate two transformer-based approaches: a fine-tuned BERT classifier and a BERT-based model augmented with linguistic features for gender classification from review text. Experiments conducted on a large-scale dataset of place reviews show that the augmented BERT model achieves high performance, reaching an average F1-score of 0.95. Beyond predictive performance, we explore how inferred gender proxy labels can support urban representativeness analysis. Using New York City as a case study, we analyze the spatial distribution of gender imbalance across ZIP codes and assess the extent to which these patterns align with an external benchmark (Foursquare). These findings highlight both the potential and the limitations of using inferred demographic proxy attributes to study urban representativeness. Our results demonstrate that contextual language models can support demographic inference in location-based social data, enabling new perspectives on urban behavior while raising important considerations regarding bias, uncertainty, and representativeness.*

***Resumo.*** *O conteúdo gerado pelo usuário em plataformas baseadas em localização fornece informações valiosas sobre o comportamento urbano, mas a falta de informações demográficas explícitas limita as análises de representatividade social. Em particular, compreender as diferenças de gênero no uso do espaço urbano permanece um desafio devido à ausência de atributos estruturados do usuário. Neste trabalho, investigamos o uso de técnicas de processamento de linguagem natural para inferir rótulos binários de gênero a partir de avaliações textuais no Google Places. Avaliamos duas abordagens baseadas em Transformers: um classificador BERT ajustado e um modelo baseado em BERT aumentado com características linguísticas para classificação de gênero a partir do texto da avaliação. Experimentos conduzidos em um conjunto de*

*dados em larga escala de avaliações de locais mostram que o modelo BERT aumentado alcança alto desempenho, atingindo uma pontuação F1 média de 0,95. Além do desempenho preditivo, exploramos como os rótulos proxy de gênero inferidos podem apoiar a análise de representatividade urbana. Usando a cidade de Nova York como estudo de caso, analisamos a distribuição espacial do desequilíbrio de gênero entre os CEPs e avaliamos em que medida esses padrões se alinham com um benchmark externo (Foursquare). Essas descobertas destacam tanto o potencial quanto as limitações do uso de atributos demográficos inferidos para estudar a representatividade urbana. Nossos resultados demonstram que modelos de linguagem contextual podem auxiliar na inferência demográfica em dados sociais baseados em localização, possibilitando novas perspectivas sobre o comportamento urbano, ao mesmo tempo que levantam considerações importantes sobre viés, incerteza e representatividade.*

## 1. Introduction

The widespread adoption of Information and Communication Technologies (ICTs), particularly the Internet and Web 2.0, has fundamentally transformed how individuals interact with space, time, and each other [Yuan et al. 2018, de Souza e Silva 2007]. In this context, Location-Based Social Networks (LBSNs), such as Google Places and Foursquare, have emerged as valuable sources of data for understanding urban behavior, capturing user interactions with places at an unprecedented scale [Santos et al. 2024]. These platforms provide rich, user-generated content that reflects preferences, experiences, and mobility patterns across cities.

At the same time, the rapid growth of online platforms has led to a massive increase in textual data, including reviews, comments, and discussions. Such content often contains subtle linguistic cues that can reveal user attributes, including demographic characteristics, preferences, and behavioral tendencies. Traditional methods for collecting demographic information, such as surveys and registration forms, are often costly, intrusive, and difficult to scale. Consequently, natural language processing (NLP) and machine learning (ML) techniques have become a promising alternative, enabling the extraction of user profiles directly from textual data in a non-intrusive manner [Eke et al. 2019, Rogers et al. 2022].

Recent advances in NLP, particularly the development of transformer-based models such as BERT, have significantly improved the ability to capture contextual and semantic information from text [Devlin et al. 2019]. These models have demonstrated high performance in user profiling tasks, including the inference of demographic attributes from language use. However, despite these advances, important challenges remain—especially regarding bias, representativeness, and the reliability of inferred attributes in real-world datasets.

From an urban computing perspective, demographic attributes such as age and gender are essential for understanding inequalities, behavioral differences, and the representativeness of digital traces [O'Connor et al. 2024]. These attributes enable the analysis of how different groups interact with urban spaces and help uncover disparities in access, participation, and experience. Nevertheless, most LBSN datasets do not provide explicit demographic information, limiting their applicability for such analyses. Further-

more, prior work has shown that location-based social media data are subject to multiple sources of bias, including uneven participation across demographic groups and spatial sampling distortions [Yuan et al. 2018].

Inferring demographic attributes from user-generated text offers a potential solution to the lack of explicit demographic information, but it also introduces new methodological and interpretability challenges. User profiling remains a difficult task, as it requires identifying latent user characteristics from noisy and heterogeneous textual data [Gómez et al. 2023]. In particular, it is still unclear to what extent such inferred attributes can be reliably used to study urban phenomena and represent population groups.

In this work, we address this gap by investigating the use of transformer-based models to infer gender from Google Places reviews and by examining how these inferred attributes can support urban analysis. Specifically, we (i) evaluate the performance of BERT-based approaches for gender classification from review text, and (ii) analyze the spatial distribution of predicted gender across urban areas, using New York City as a case study. By connecting demographic inference with spatial analysis, this study opens up opportunities for using LBSN data to study urban behavior.

The remainder of this study is organized as follows. Section 2 presents the related work. Section 3 describes the data and methods used in the study. Section 4 presents the experimental results. Section 5 discusses ethical considerations and limitations. Finally, Section 6 concludes the paper and provides final remarks.

## 2. Related Work

This work relates to three main research directions: (i) the use of location-based social networks (LBSNs) to study urban behavior, (ii) demographic bias and representativeness in social media data, and (iii) user profiling through textual analysis.

**LBSNs and urban behavior.** LBSNs have been widely used to analyze human behavior in urban environments, as they provide large-scale, user-generated data reflecting interactions with places [Gubert et al. 2024, Silva and Silver 2025, Santos et al. 2024]. Prior studies have examined how demographic factors such as gender influence mobility patterns and venue preferences [Muhammad et al. 2019], as well as how different platforms capture economic and social characteristics of cities [Bernabeu-Bautista et al. 2021]. These data sources enable the construction of interest networks and support analyses of how individuals engage with urban spaces. However, most LBSN datasets lack explicit demographic information, limiting their use for studying population differences.

**Bias and representativeness in social media.** A substantial body of work shows that social media data are not demographically representative and may introduce systematic biases [Yuan et al. 2018, Mueller et al. 2017, Hargittai 2020]. Socioeconomically advantaged users are typically overrepresented, while certain demographic groups and geographic areas remain underrepresented [Hargittai 2020, Blank and Lutz 2017, Sanderson et al. 2024]. In the context of LBSNs, [Yuan et al. 2018] demonstrated that gender participation varies across regions, and [Mueller et al. 2017] identified gender-based differences in venue preferences. Despite these advances, representativeness issues remain underexplored when demographic attributes are not directly available.

**User profiling from text.** User profiling aims to infer demographic and behavioral attributes from user-generated content [Ikae and Savoy 2022, Alekseev and Nikolenko 2017, Thome et al. 2025, Aletras and Chamberlain 2018]. Early approaches relied on stylometric features and traditional machine learning models, achieving moderate performance. More recent work has leveraged deep learning and transformer-based architectures, significantly improving accuracy. For instance, [Sarwar et al. 2024] employed multilingual transformer models for predicting gender from text, achieving an accuracy of 92%, while [Himdi and Shaalan 2024] proposed an enhanced BERT-based approach for author gender identification in Arabic text. However, these studies are often conducted on relatively small or domain-specific datasets, limiting their applicability to heterogeneous, real-world data such as LBSNs.

**Research gap.** Despite progress in gender inference from text, its application to place reviews remains limited. Moreover, few studies connect user profiling with representativeness analysis in an urban computing context. As a result, it is still unclear how reliably inferred demographic attributes can be used to assess representativeness and understand interactions with urban spaces—particularly in LBSNs, where such information is typically missing. In this work, we address this gap by combining transformer-based user profiling with spatial analysis to investigate gender representation in urban environments.

## 3. Data and Methods

### 3.1. Google Places Dataset

We use a publicly available Google Places dataset consisting of user reviews associated with establishments listed in Google Maps [He et al. 2017, Pasricha and McAuley 2018]. The data include textual reviews, user identifiers, timestamps, and metadata about places (e.g., category, location, and geographic coordinates).

The dataset initially contains 33,459,761 reviews. Each record includes textual, user-related, and spatial attributes. The *UserID* uniquely identifies each reviewer, and the *name* field is used for gender inference ground truth. The *text* field contains the review content, while *rating* and *time* represent the evaluation and timestamp of each review.

Spatial information includes the place identifier (*gmapid*), administrative divisions (e.g., *reviewcity*, *reviewzipcode*, *reviewcounty*), hierarchical spatial indices, and precise geographic coordinates (latitude and longitude). Additionally, the dataset includes a *gender* field corresponding to the inferred label obtained through the gender labeling process.

This combination of textual and spatial data enables the analysis of user behavior and the investigation of gender-related patterns in urban environments.

For this study, we focus on reviews from New York City, providing a diverse set of user-generated content across multiple venue types. The review text is used as input for the user profiling model, while spatial attributes (e.g., latitude and longitude) support the subsequent urban analysis. Reviews may be written in multiple languages, reflecting the multilingual nature of the platform.

### 3.2. User Gender Identification Model

We model gender inference as a text classification task based on user reviews. To this end, we adopt a transformer-based approach using the *bert-base-multilingual-cased* model,

which is well-suited for handling multilingual textual data.

We consider two model configurations. The first uses BERT as a standalone classifier, while the second augments BERT embeddings with handcrafted linguistic features.

Compared to traditional approaches based on stylometric features or static embeddings, this strategy leverages contextual information and long-range dependencies in text, which are critical for capturing subtle linguistic signals associated with user attributes.

Figure 1 illustrates the architecture of this study, particularly the modeling section, starting from data collection to model training.

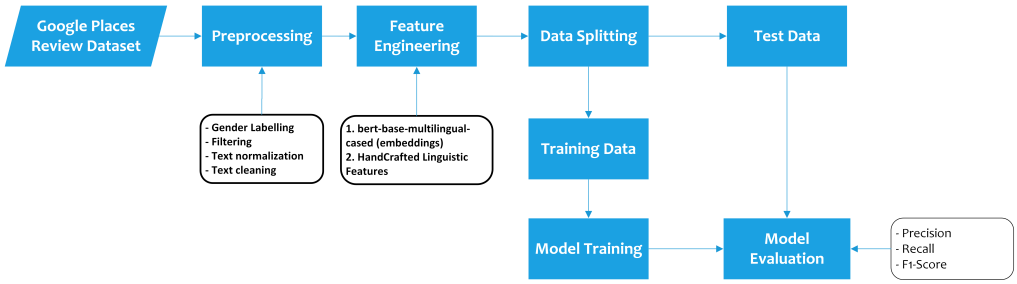



**Figure 1. Key steps of the gender classification approach**

The illustration shown above in Figure 1 has a step-by-step process that we have performed throughout this study, particularly for model training. Below, we have discussed the activities that we have performed on each section:

**Google Places Review Dataset**: The process begins with loading the Google Places review dataset, for New York City, which contains textual reviews along with associated UserID and spatial attributes.

**Preprocessing**: In the preprocessing stage, the raw text is cleaned and normalized to make it suitable for analysis. This involves converting all text to lowercase and removing unwanted elements, such as URLs, HTML tags, and hashtags. Additionally, custom patterns such as timestamp-like strings are removed, and extra whitespace is normalized. These steps help reduce noise and standardize the textual data. Furthermore, to ensure data quality and consistency, the dataset is filtered by retaining only reviews whose length falls within a specified range (500 to 2000 characters) and also selects the users whose reviews are more than five (5) posts. This step removes excessively short or overly long texts that may introduce noise or bias into the model.

**Gender Labeling**: The gender labeling component automatically infers user gender from names to support supervised learning. First, each user's first name is extracted from the full name using basic string processing. These names are then passed to the R-based "gender" package[1], accessed in Python via rpy2, which predicts gender using historical datasets such as the U.S. Social Security Administration (SSA). The model returns both gender labels and confidence measures (e.g., proportion of male/female associations), helping address ambiguity in certain names. The predicted results are merged back into the original dataset, producing a labeled dataset that is subsequently used for

[1]https://github.com/lmullen/gender

model training, as illustrated in the architecture. Gender labels are inferred from user names and used solely for training (supervision); the model itself relies exclusively on textual review content. To reduce noise and cultural bias, only predictions with a confidence of at least 95% are considered during the labeling process.

**Feature Engineering**: Feature engineering involves extracting both linguistic and contextual representations of the text. Linguistic features include statistical and stylistic measures such as word count, sentence count, average word length, lexical diversity, punctuation usage, and noun ratio obtained through part-of-speech tagging. In parallel, the text is processed using the *bert-base-multilingual-cased* transformer tokenizer to generate contextual embeddings. These two types of features are combined to provide a richer representation of the data. The *bert-base-multilingual-cased* model was selected due to its strong multilingual capability and its ability to capture rich contextual information from user-generated text, making it suitable for handling diverse and noisy inputs [Gardazi et al. 2025, Vaswani et al. 2017, Devlin et al. 2019, Liu et al. 2019].

**Data Splitting**: The processed dataset is divided into training and testing subsets, using an 80/20 split. In the study, we consider two experiments. The first experiment employed a random train–test split, whereas the second experiment utilized a UserID-based splitting strategy to ensure that data from the same user did not appear in both training and testing sets. Under the UserID-based split, all reviews from the same user are assigned exclusively to either training or test data. For evaluation, each review is classified independently; user-level predictions are then obtained by majority voting over a user's review-level predictions (ties broken randomly), and performance under this split is reported based on these user-level labels. Class imbalance in the training set is addressed using random over-sampling to ensure balanced supervision during model fitting. The test set is kept separate for evaluation under the chosen split protocol.

**Training Data**: The training dataset consists of preprocessed text, corresponding linguistic feature vectors, contextual data, and gender labels. These inputs are structured into a format compatible with the model, ensuring that both textual and numerical features are aligned correctly for training.

**Model Training**: During model training, a hybrid architecture is employed that combines transformer-based embeddings with handcrafted linguistic features. The BERT model generates contextual embeddings from the input text, specifically using the [CLS] token representation. These embeddings are concatenated with the extracted linguistic features and passed through a classification layer. The model is trained using a weighted cross-entropy loss function to account for class distribution, and optimization is performed over multiple epochs with a learning rate scheduler. Importantly, names are used only to generate gender labels for supervision and are not used as input features to the model. The classifier relies exclusively on review text.

**Test Data**: The test dataset undergoes the same preprocessing and feature engineering steps as the training data to maintain consistency. It is kept separate from the training process and is used solely for evaluating the model's performance.

**Model Evaluation**: The trained model is evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score. Both weighted and macro-averaged metrics are computed to provide a comprehensive assessment of performance. A

detailed classification report is also generated to analyze the model's effectiveness across different classes.

**Trained Model**: The final output of the pipeline is a trained model capable of predicting gender from textual reviews. By leveraging both deep contextual embeddings from the transformer model and handcrafted linguistic features, the model achieves a robust understanding of the text, enabling accurate classification.

### 3.3. Experimental Setup

This section describes the experimental setup used to evaluate two model variants: (i) a Pure BERT baseline and (ii) a BERT model augmented with linguistic features. For each model, we consider two train–test partitioning strategies introduced previously—a random split and a UserID-based split—allowing us to assess both predictive performance and cross-user generalization.

**Pure BERT configuration.** In the first setup, we adopt a Pure BERT architecture in which `bert-base-multilingual-cased` serves as the sole text encoder. Raw review text is tokenized and passed through BERT, and the [CLS] token representation (768 dimensions) is used as a document-level embedding. This representation is fed into a dropout layer (rate = 0.1), followed by a linear classification layer for gender prediction. The model is fine-tuned end-to-end using a batch size of 64, a learning rate of $2 \times 10^{-5}$, and 30 training epochs. Optimization is performed using AdamW with a cosine annealing scheduler, while class-weighted cross-entropy loss is used to mitigate class imbalance.

**BERT with Linguistic Features.** In the second setup, we extend the Pure BERT model by incorporating 11 handcrafted linguistic features alongside BERT embeddings for classification. All training hyperparameters and optimization settings remain identical to the Pure BERT configuration, allowing us to isolate the contribution of the additional linguistic features.

Performance is evaluated using precision, recall, and F1-score. This design allows us to assess both the added value of linguistic features beyond contextual embeddings and the robustness of each approach under user-disjoint evaluation settings.

**External spatial benchmark.**

As an additional validation analysis, we compare the ZIP code-level gender patterns inferred from Google Places reviews with those observed in a Foursquare dataset for which gender labels are available [Mueller et al. 2017]. We treat Foursquare not as population ground truth, but as a reference platform for evaluating whether inferred labels from Google Places reproduce similar spatial patterns of gender imbalance.

For each ZIP code, gender imbalance is computed as $p_{male} - p_{female}$, which is the difference in proportions when the two groups exhaust the sample. We restrict the comparison to ZIP codes with at least 25 observations per gender in both datasets and evaluate concordance using Pearson and Spearman correlations, directional agreement, bias, and mean absolute error between Google Places and Foursquare estimates.

## 4. Results

In this section, we present the main findings of our study. We begin by understanding linguistic patterns across users. We then evaluate the proposed model's performance and

examine the spatial distribution of inferred gender.

### 4.1. Textual Characteristics Between Male and Female

We begin by analyzing the textual content of the dataset. To explore potential differences across user groups, Figure 2 presents word clouds for reviews associated with female and male users.


**Figure 2. Word Cloud by Gender**

Both groups share a common vocabulary centered on place-related experiences (e.g., food, service, and quality), which reflects the nature of the platform. However, differences in term prominence suggest the presence of subtle linguistic variations between groups.

While these differences are not easily separable through simple frequency-based analysis, they indicate that demographic signals may be embedded in language use. This observation supports the use of contextual models, such as BERT, which are capable of capturing more complex semantic and stylistic patterns beyond surface-level word frequencies.

### 4.2. Gender Inference Performance

#### 4.2.1. Pure BERT Performance

The performance of the Pure BERT model is presented in Table 1, which reports precision, recall, and F1-score under both the random-based split and the UserID-based split evaluation settings. The Pure BERT model achieved high performance, reaching an average F1-score of 0.92 under the random train–test split and 0.86 under the more challenging UserID-based split. The performance decrease in the UserID-based setting likely reflects the increased difficulty of generalizing to unseen users and provides a more realistic assessment of the model's ability to generalize across unseen users. Nevertheless, performance under the UserID-based split remains high, indicating that the model retains substantial predictive power even when evaluated in a stricter cross-user generalization setting.

#### 4.2.2. BERT with Linguistic Features

We further evaluate the BERT model augmented with linguistic features for gender classification from place reviews. The results are presented in Table 2.

**Table 1. Pure BERT Performance considering Random and UserID-based split**

| | Random Split | | | UserID-based Split | | |
|---|---|---|---|---|---|---|
| **Class** | **Precision** | **Recall** | **F1-score** | **Precision** | **Recall** | **F1-score** |
| Female | 0.94 | 0.93 | 0.93 | 0.84 | 0.88 | 0.86 |
| Male | 0.91 | 0.91 | 0.91 | 0.87 | 0.83 | 0.85 |
| Average | 0.92 | 0.92 | 0.92 | 0.86 | 0.86 | 0.86 |

**Table 2. BERT with Linguistic Features performance considering Random and UserID-based split**

| | Random Split | | | UserID-based Split | | |
|---|---|---|---|---|---|---|
| **Class** | **Precision** | **Recall** | **F1-score** | **Precision** | **Recall** | **F1-score** |
| Female | 0.97 | 0.96 | 0.96 | 0.93 | 0.96 | 0.94 |
| Male | 0.93 | 0.94 | 0.93 | 0.96 | 0.92 | 0.94 |
| Average | 0.95 | 0.95 | 0.95 | 0.94 | 0.94 | 0.94 |

The BERT model augmented with linguistic features achieves high performance, with an average F1-score of 0.95 under the random split and 0.94 under the UserID-based split, demonstrating robust performance even under the stricter cross-user evaluation setting.

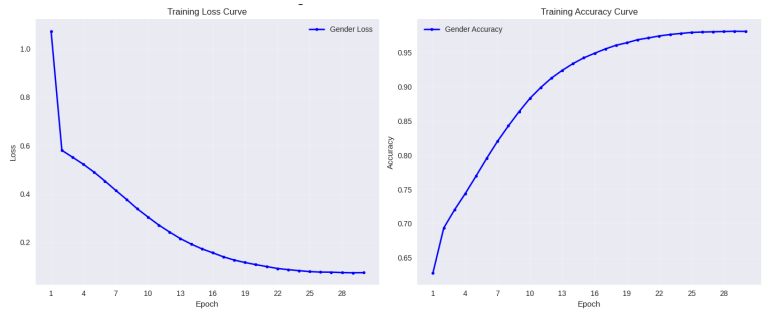



**Figure 3. Training loss and accuracy curves for the BERT model augmented with linguistic features on the Google Places dataset (random split, 30 epochs)**

Figure 3 shows the training dynamics of the model. The results indicate stable convergence, with decreasing loss and increasing accuracy over training epochs, suggesting that the model effectively learns meaningful representations from the data.

### 4.3. Gender Inference: Comparison with Previous Approaches

We compare our results with prior studies on gender identification from textual data, as summarized in Table 3.

**Table 3. Related works on gender prediction based on user textual content**

| **Author(s)** | **Techniques** | **Best Result** |
|---|---|---|
| [Alekseev and Nikolenko 2017] | Word2Vec, clustering, logistic regression | 0.81 (F1-Score) |
| [Ikae and Savoy 2022] | LR, DT, KNN, SVM, NB, NN, RF, LightGBM | 0.84 (F1-Score) |
| [Sarwar et al. 2024] | DistilBERT, mBERT, XLM-RoBERTa, mDE-BERTa | 0.92 (Accuracy) |
| [Himdi and Shaalan 2024] | CNNs, LSTM, BiLSTM, BERT | 0.91 (F1-Score) |

Among these, [Sarwar et al. 2024] reports one of the strongest results, achieving an accuracy of 92.03% using multilingual transformer models. However, differences in datasets, preprocessing strategies, and evaluation protocols limit direct comparability. To complement this comparison, we further evaluate our model using the same train–test split protocol on the dataset introduced by [Sarwar et al. 2024].

### 4.4. Evaluation on an External Benchmark

To further assess the robustness of our approach beyond Google Places reviews, we evaluate our model on the external benchmark introduced by [Sarwar et al. 2024], which consists of multilingual news articles annotated for author gender. This dataset differs substantially from our primary place-review corpus, providing a distinct domain for evaluating transferability.

Among prior work, [Sarwar et al. 2024] reported high performance using multilingual transformer models (DistilBERT, mBERT, XLM-RoBERTa, and mDEBERTa) under a random train–test split. Using the same split protocol, we benchmark our BERT model augmented with linguistic features on their dataset (Figure 4), with results summarized in Table 4.

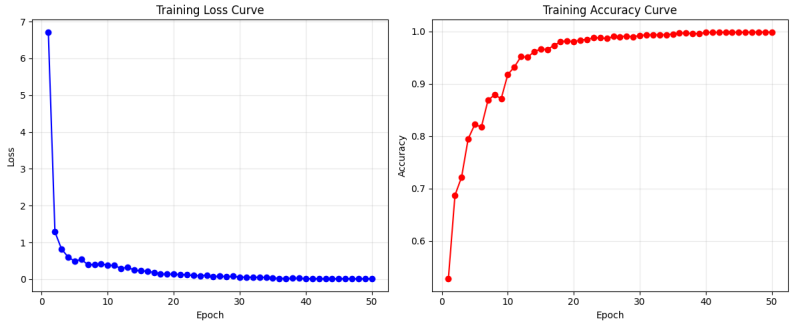



**Figure 4. Training loss and accuracy curves for the BERT model augmented with linguistic features on the [Sarwar et al. 2024] dataset (random split)**

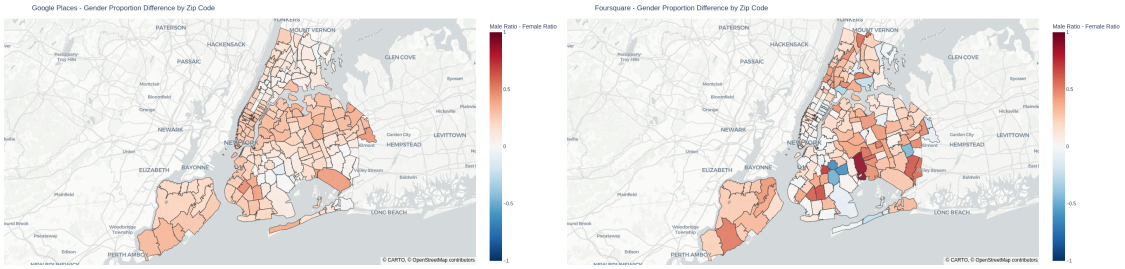
**Table 4. Comparison of our method on the dataset of [Sarwar et al. 2024]**

| Method | Techniques | Accuracy |
|---|---|---|
| [Sarwar et al. 2024] | DistilBERT, mBERT, XLM-RoBERTa, mDEBERTa | 0.92 |
| **Proposed method** | BERT + linguistic features | **0.98** |

Our implementation achieved 98% accuracy on this benchmark under our preprocessing and feature setup, compared with the 92% reported in [Sarwar et al. 2024], suggesting that the proposed approach performs competitively in a distinct textual domain. This external benchmark provides preliminary evidence that the approach may transfer beyond place-review data, although further evaluation on diverse datasets is needed.

### 4.5. Spatial Distribution of Gender

We analyze the spatial distribution of inferred gender across the city. Figure 5 presents the gender imbalance across ZIP codes for Google Places and Foursquare, computed as the difference between male and female review proportions.




(a) Google Places (b) Foursquare

**Figure 5. Spatial distribution of male versus female user imbalance by ZIP code**

Figure 5a highlights substantial spatial heterogeneity in gender representation, with some areas exhibiting stronger male- or female-skewed participation than others. These patterns may reflect differences in venue composition, economic activity, or urban functions across neighborhoods, as well as variation in how demographic groups engage with urban spaces.

At the same time, these distributions should be interpreted cautiously. Because they reflect patterns of platform activity rather than population composition, they may also capture platform-specific biases, including uneven participation across places or user groups. Thus, the observed spatial patterns should be understood as signals of gendered platform usage rather than direct estimates of demographic distributions.

To assess whether inferred gender labels in Google Places reproduce plausible urban patterns, we compare ZIP code-level gender imbalance with the same measure computed from Foursquare (Figure 5b), used here as a benchmark platform. Visual comparison suggests some similarities in high-activity areas, but the quantitative comparison reveals limited concordance overall. Across all matched ZIP codes, correlation is weak (Pearson $r = 0.05$, Spearman $\rho = -0.03$), indicating that Google Places does not fully reproduce the spatial pattern observed in Foursquare.

However, concordance improves when restricting the analysis to the highest-volume areas, although this agreement depends on the scale considered. For the top 10 ZIP codes with the largest number of observations (based on Foursquare), we observe a strong positive linear association (Pearson $r = 0.70$, $p = 0.025$), while the top 25 ZIP codes show moderate rank agreement (Spearman $\rho = 0.46$, $p = 0.02$) and a moderate positive linear association (Pearson $r = 0.39$, $p = 0.053$). In contrast, concordance weakens substantially when expanding to the top 50 ZIP codes. These results suggest that inferred labels capture meaningful spatial structure primarily in the highest-activity areas, while agreement diminishes as lower-volume areas are included.

At the same time, Google Places tends to produce systematically more male-skewed estimates than Foursquare, reflected in a positive mean bias of 0.072 (median = 0.091), a mean absolute error of 0.201 (median = 0.176), and higher imbalance values in 64.47% of matched ZIP codes. Together, these findings suggest that the inferred labels recover some broad urban structure, particularly in high-volume areas, but also exhibit a systematic tendency toward higher male-skew estimates that should be considered when interpreting representativeness patterns derived from the inferred labels.

## 5. Ethical Considerations and Limitations

This study involves inferring demographic attributes from user-generated content, which raises important ethical and methodological considerations. First, the gender labels used in this work are not directly observed but inferred from user names using external datasets. As such, these labels may contain inaccuracies, cultural biases, and ambiguities, particularly for names that are gender-neutral or underrepresented in the reference data. Consequently, the model learns to predict name-inferred gender rather than ground-truth demographic attributes.

Importantly, user names are used solely to generate supervision labels and are not included as input features in the model; all predictions are based exclusively on review text. However, reliance on weak supervision labels introduces uncertainty into both training and evaluation, and reported performance should therefore be interpreted as agreement with inferred labels rather than true demographic classification accuracy. Although additional experiments on external datasets suggest the proposed approach remains effective beyond the original data, these results should be interpreted with caution, given the limitations of the labeling process. Similarly, the comparison with Foursquare should be understood as a benchmark of cross-platform consistency rather than validation against population-level ground truth.

In addition, this study adopts binary gender categories, which do not reflect the full spectrum of gender identities. This simplification is driven by limitations in available data and labeling methods, but it inherently excludes non-binary and gender-diverse individuals. Future work should explore more inclusive approaches to demographic inference.

More broadly, the use of inferred demographic attributes in urban analysis may amplify biases already present in location-based social network data. Such platforms are known to be demographically unrepresentative, with uneven participation across social groups and geographic areas. As a result, the spatial patterns identified here should not be interpreted as direct representations of population-level distributions, but rather as reflections of platform-specific user activity and participation.

Finally, inferring demographic attributes from textual data raises concerns related to privacy and potential misuse. Although this study relies on publicly available data, the ability to infer sensitive attributes highlights the need for responsible use of such techniques. We emphasize that the goal of this work is to support aggregate-level urban analysis, not individual profiling, and we encourage future research to further examine ethical safeguards for demographic inference in urban data science.

## 6. Conclusions

In this work, we investigated the use of transformer-based models for user profiling in location-based social networks, focusing on gender inference from Google Places reviews. Our results show that contextual language models, such as *bert-base-multilingual-cased*, can capture semantic and stylistic signals associated with user attributes, achieving strong classification performance. The proposed approach demonstrates competitive performance and promising robustness across distinct datasets.

Beyond predictive performance, this study explored how inferred demographic

attributes can support urban analysis. By examining spatial patterns of gender imbalance and their alignment with an external benchmark, we found that user representation is not spatially uniform, with certain areas exhibiting stronger gender imbalances than others. These findings highlight the potential of combining NLP-based user profiling with spatial analysis to provide new perspectives on urban behavior and representativeness.

At the same time, our results emphasize important limitations. Inferred demographic attributes introduce uncertainty, and the observed spatial patterns may reflect not only behavioral differences but also platform-specific biases and uneven participation across user groups. In addition, the comparison with Foursquare suggests that while inferred Google Places labels recover some broad spatial structure, they do not fully align with the benchmark platform and tend to overestimate male-skewed patterns in many ZIP codes. These findings reinforce the need for caution when interpreting inferred demographic signals as proxies for broader population characteristics.

Future work can extend this study in several directions. Exploring alternative language models and incorporating additional data modalities (e.g., behavioral or temporal signals) may improve the robustness of user profiling. Evaluating the approach across different cities and cultural contexts would help assess its generalizability, while further work is needed to better understand and mitigate biases in inferred demographic attributes when applying these methods to urban studies.

## 7. Acknowledgments

National Council for Scientific and Technological Development - CNPq (processes 314603/2023-9, 441444/2023-7, 409669/2024-5, and 444724/2024-9). This research is also part of the INCT TILD-IAR funded by CNPq (proc. 408490/2024-1).

## References

Alekseev, A. and Nikolenko, S. (2017). Word embeddings for user profiling in online social networks. Computacion y Sistemas, 21(2):203–226.

Aletras, N. and Chamberlain, B. P. (2018). Predicting Twitter User Socioeconomic Attributes with Network and Language Information. In Proceedings of the 29th on Hypertext and Social Media, pages 20–24, New York, NY, USA. ACM.

Bernabeu-Bautista, A., Serrano-Estrada, L., Perez-Sanchez, V. R., and Marti, P. (2021). The geography of social media data in urban areas: Representativeness and complementarity. ISPRS International Journal of Geo-Information, 10(11).

Blank, G. and Lutz, C. (2017). Representativeness of social media in great britain: Investigating facebook, linkedin, twitter, pinterest, google+, and instagram. American Behavioral Scientist, 61(7):741–756.

de Souza e Silva, A. (2007). Cell phones and places: The use of mobile technologies in brazil. In Societies and cities in the age of instant access, pages 295–310. Springer.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proc. of NAACL-HLT, volume 1, pages 4171–4186.

Eke, C. I., Norman, A. A., Shuib, L., and Nweke, H. F. (2019). A survey of user profiling: State-of-the-art, challenges, and solutions. IEEE Access, 7:144907–144924.

Gardazi, N. M., Daud, A., Malik, M. K., Bukhari, A., Alsahfi, T., and Alshemaimri, B. (2025). Bert applications in natural language processing: a review. Artificial Intelligence Review, 58(6):166.

Gubert, F. R., Santos, G. H., Delgado, M., Silver, D., and Silva, T. H. (2024). Culture Fingerprint: Identification of Culturally Similar Urban Areas Using Google Places Data. In Proc of ASONAM, Rende, Calabria, Italy.

Gómez, J.-C., Moreno, J., Manzano, M. A. I., and Ojeda, D. L. A. (2023). Reconstructive classification for age and gender identification in social networks. IEEE Trans. on Comput. Social Sys., 11(2):2291–2301.

Hargittai, E. (2020). Potential biases in big data: Omitted voices on social media. Social Science Computer Review, 38(1):10–24.

He, R., Kang, W.-C., and McAuley, J. (2017). Translation-based recommendation. In Proc of RecSys, page 161–169, New York, NY, USA. Association for Computing Machinery.

Himdi, H. and Shaalan, K. (2024). Advancing author gender identification in modern standard arabic with innovative deep learning and textual feature techniques. Information, 15(12).

Ikae, C. and Savoy, J. (2022). Gender identification on twitter. Journal of the Association for Information Science and Technology, 73(1):58–69.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Mueller, W., Silva, T. H., Almeida, J. M., and Loureiro, A. A. (2017). Gender matters! analyzing global cultural gender preferences for venues using social sensing. EPJ Data Science, 6(1):5.

Muhammad, R., Zhao, Y., and Liu, F. (2019). Spatiotemporal analysis to observe gender based check-in behavior by using social media big data: A case study of guangzhou, china. Sustainability, 11(10).

O'Connor, K., Golder, S., Weissenbacher, D., Klein, A. Z., Magge, A., and Gonzalez-Hernandez, G. (2024). Methods and annotated data sets used to predict the gender and age of twitter users: Scoping review. Journal of Medical Internet Research, 26:e47923.

Pasricha, R. and McAuley, J. (2018). Translation-based factorization machines for sequential recommendation. In Proc of RecSys, page 63–71, New York, NY, USA. Association for Computing Machinery.

Rogers, D., Preece, A., Innes, M., and Spasić, I. (2022). Real-time text classification of user-generated content on social media: Systematic review. IEEE Transactions on Computational Social Systems, 9(4):1154–1166.

Sanderson, R., Franklin, R., MacKinnon, D., et al. (2024). Left out and invisible?: Exploring social media representation of 'left behind places'. GeoJournal, 89:37.

Santos, G., Gubert, F., Delgado, M., and Silva, T. (2024). Redes de interesse: comparando o google places e foursquare na captura da escolha de usuários por áreas urbanas. In Proc of CoUrb, pages 99–112, Niterói/RJ. SBC.

Sarwar, R., An Ha, L., Teh, P. S., Sabah, F., Nawaz, R., Hameed, I. A., and Hassan, M. U. (2024). Agi-p: A gender identification framework for authorship analysis using customized fine-tuning of multilingual language model. IEEE Access, 12:15399–15409.

Silva, T. H. and Silver, D. (2025). Using graph neural networks to predict local culture. Environment and Planning B: Urban Analytics and City Science, 52(2):355–376.

Thome, B., Hertweck, F., and Conrad, S. (2025). Predicting perceived text complexity: The role of person-related features in profile-based models. Journal of Educational Data Mining, 17(1):276–307.

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (NeurIPS), volume 30, pages 5998–6008.

Yuan, Y., Wei, G., and Lu, Y. (2018). Evaluating gender representativeness of location-based social media: a case study of weibo. Annals of GIS, 24(3):163–176.