What Does a Neighborhood Look Like Online? Linking the Built Environment to Digital Information Diets

Hemant Magwana
IIT Kharagpur
hemant.mag99@kgpian.iitkgp.ac.in

Kiran Garimella
Rutgers University
kiran.garimella@rutgers.edu

Abstract

The relationship between our physical environment and our increasingly digital social lives remains poorly understood, particularly within private communication platforms like WhatsApp, that dominate the Global South. To address this gap, we construct and analyze a novel dataset linking granular WhatsApp usage data from 3,127 individuals across 100 towns in Uttar Pradesh, India, with high-resolution, quantitative metrics of their immediate physical surroundings. These offline metrics are systematically extracted by applying a state-of-the-art vision-language model to thousands of Google Street View images, quantifying the built environment, local economic conditions, and public cleanliness.

Our analysis reveals that the physical world is systematically correlated with digital behavior in specific and often non-intuitive ways: the quality of public infrastructure, for example, predicts the style of online interaction (e.g., conversational equity), while visible environmental cues related to health and sanitation are associated with increased privacy-seeking behaviors (e.g., the use of ephemeral messages). We also uncover an asymmetric relationship, finding that while the offline world shapes online activity, digital traces alone are poor predictors of the physical environment when divorced from demographics. Finally, our findings show that information diffuses via a dual-mode system, comprising a slow, statewide "broadcast" of broad topics and a rapid, "hyper-local bulletin" of urgent keywords. Together, this work provides the first large-scale, empirical evidence of the deep connection between our physical world and our private digital conversations, offering a new methodological paradigm for studying society in the digital age.

ACM Reference Format:
Hemant Magwana and Kiran Garimella. 2025. What Does a Neighborhood Look Like Online? Linking the Built Environment to Digital Information Diets. In . ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

We live in an ever growing digital world, where our digital and physical lives are inextricably linked. Yet, the precise nature of this connection remains unclear. While we understand that our online behaviors are not created in a vacuum, we lack a systematic understanding of how the tangible, observable characteristics of our physical world such as the built environment, condition of our roads, or the quality of our public spaces shape the substance and style of our private digital conversations. This paper addresses this fundamental question by investigating the relationship between the physical environment and digital communication within the world's largest democracy, India. The central question of this study is whether the physical character of a locality systematically shapes the substance and style of its inhabitants' private digital conversations.

The urgency and importance of this question are magnified in the context of India, where over 500 million people have come online in the last decade alone [12], creating a digital society whose behaviors are potent but poorly understood. For this massive new digital citizenry, and for billions across the Global South, private messaging platforms like WhatsApp are not just a tool for communication but are the primary interface to the digital world [3]. Despite this reality, the vast majority of computational social science has focused on public platforms like Twitter and on "WEIRD" (Western, Educated, Industrialized, Rich, and Democratic) societies, creating a significant blind spot in our understanding [19]. Probing the online-offline link in a setting like rural and semi-urban India is therefore not a niche inquiry; it is a critical step towards developing a more globally representative science of human society in the digital age. Understanding this link is essential for everything from designing effective public health campaigns to monitoring local economic distress and fostering civic engagement.

This problem has remained intractable due to profound methodological barriers on both sides of the online-offline divide. The digital side, particularly private communication, is a "black box" by design; end-to-end encryption and the absence of public APIs have made platforms like WhatsApp largely inaccessible to researchers, raising immense ethical and logistical hurdles. The offline side presents a parallel challenge: quantifying the physical environment at scale has traditionally required either prohibitively expensive and slow methods, like in-person neighborhood audits, or a reliance on coarse and often outdated census data. These dual challenges have made it nearly impossible to link a granular view of the physical world to the private digital behaviors of the people who inhabit it.

Previous research has thus been fragmented. Studies of online life, even when geolocated, typically lack a rich, quantitative understanding of the user's immediate environment [21]. Conversely, a groundbreaking new field uses computer vision on Google Street View (GSV) to quantify neighborhood characteristics at an unprecedented scale [10], yet these studies can only see the environment; they cannot hear the conversations of the people living there. This paper bridges this chasm by constructing and analyzing a novel, large-scale dataset that directly links these two worlds.

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

(a) Presence of shop/temple (b) Hospital (c) Bad road (d) Garbage

Figure 1: Examples of offline signals captured by our pipeline. Panels (a)–(b) show detected points of interest (POIs) that mark local functions: (a) a retail shop and a temple, (b): a diagnostic center. Panels (c)–(d) show complementary context from Google Street View where POIs alone are not sufficient, including poor road condition and low cleanliness with visible issues (garbage, open drains, stagnant water).

Our study utilizes a novel dataset constructed by linking consented WhatsApp data with computationally derived metrics of the physical environment. The online component consists of approximately 5 million messages from 6,500 groups, collected between April and July 2024 via data donation from participants across 100 towns in Uttar Pradesh, India's most populous state. Using the precise home location of each donor, we gathered corresponding Google Street View imagery and Point-of-Interest (POI) data. We then employed vision-language models to analyze this offline data, systematically scoring environmental features such as road condition, public amenities, cleanliness, visible sanitation, and green cover.

Figure 1 illustrates examples of POI and GSV data from various locations in our sample. The Figure shows how POIs and Street View complement each other. Panels (a)–(b) show clear POI detections near donor homes, including a diagnostic center (Sai Diagnostic Centre Pathology and Xray), a retail shop (Cakes N Bakes, Sitapur), and a temple (Siddheswar Nath Mandir). Panels (c)–(d) add context from imagery in places where POIs alone are not informative, capturing poor road condition and low cleanliness with visible issues such as garbage, open drains, and stagnant water.

Using this dataset of GSV and POIs we build five standardized indices that summarize the offline environment. Our analysis has four parts. First, we validate the imagery and POI measures against independent demographics to show they capture real differences across towns. Second, we estimate partial correlations between the offline environment and online outcomes while holding demographics fixed, which isolates the role of place. Third, we test the inverse prediction problem to ask whether online traces alone can recover the physical world. Fourth, we study how topics and keywords spread across space and time by locating their first appearance and tracking distance and lag.

Our study makes three primary contributions. First, we provide robust empirical evidence that the physical environment is systematically correlated with digital behavior in specific and often non-intuitive ways: the local economy shapes conversational topics, the quality of public infrastructure influences interaction styles, and environmental cues of health and sanitation appear to trigger privacy-seeking behaviors online. Second, we demonstrate that this relationship is asymmetric; while the physical world leaves a clear imprint on digital life, online behaviors are a poor standalone predictor of the physical environment, a role far better served by demographics. Finally, we uncover a dual-regime system of information diffusion, where broad topics function as a persistent, statewide "broadcast layer" while specific keywords operate as urgent, "hyperlocal bulletins." Together, these findings offer a new methodological paradigm for studying the online-offline nexus and provide the first comprehensive evidence of the deep, systematic connection between the streets we see and the digital words we share.

2 Related Work

This research is situated at the intersection of four key domains: (i) the sociological link between physical environments and social behavior, (ii) the computational methods used to measure these environments, (iii) the study of communication on private digital platforms, and (iv) the dynamics of information diffusion.

Social Disorganization in the Digital Age. Social disorganization theory, a cornerstone of urban sociology, posits that structural characteristics of a neighborhood such as poverty and residential instability weaken social cohesion, leading to adverse outcomes [20]. A key component of this theory is the physical environment; visible signs of disorder are seen as both a symptom and a cause of diminished collective efficacy [17]. While early internet scholarship hypothesized a "death of distance" that would render geography irrelevant [18], subsequent work has shown that digital life is deeply embedded in physical place [22]. The physical world continues to shape social outcomes, with features like urban highways acting as tangible barriers to the formation of social ties [1].

This paper extends the social disorganization framework into the digital realm. We treat our AI-annotated Google Street View variables as powerful proxies for the theory's core concepts: variables extracted from street view images such as economic situation as a measure of poverty, and cleanliness and infrastructure condition as proxies for physical disorder. Our central hypothesis is that these offline indicators of social disorganization will manifest

linguistically within the private, localized conversations on WhatsApp. For instance, we expect that communities with poorer built environments may exhibit more conversational topics related to grievance, infrastructure failure, or economic hardship, effectively transferring offline frustrations into the digital commons.

Quantifying the Physical World with AI. Traditionally, measuring neighborhood characteristics required laborious methods like the in-person audits of Systematic Social Observation [16] or reliance on coarse, slow-moving census data. The advent of large-scale geospatial data began a computational turn, with researchers using satellite imagery to predict poverty [13] and commercial data to measure urban deprivation. More recently, the availability of street-level imagery has created a new frontier for urban analytics.

Pioneering work in this area used computer vision on GSV images to estimate the demographic makeup of neighborhoods from the prevalence of different cars [10] and to uncover visual predictors of urban change [15]. Our work builds directly on this tradition but advances it in two ways. First, we move beyond earlier computer vision techniques by leveraging modern Vision-Language Models (VLMs), following emerging best practices that allow for more nuanced, semantically rich feature extraction from GSV images [8]. Second, we apply this advanced methodology to interpret the sociability and character of urban spaces in a novel, large-scale, and cross-national context [7], specifically the understudied environment of rural and semi-urban India.

Analyzing Communication in the Global South. Studying communication on private, encrypted platforms like WhatsApp presents significant methodological challenges that distinguish it from research on public platforms like Twitter [21]. We overcome these hurdles by using a data donation methodology that prioritizes user consent and privacy [9]. Existing research on WhatsApp in India has been vital but has largely focused on its role in political campaigns and the spread of misinformation [2]. Our work diverges from this focus by analyzing the texture of everyday conversation. We contribute to this literature by moving beyond high-stakes content to understand the dynamics of routine social interaction and, crucially, by providing the first systematic link between these digital conversational patterns and the quantified physical world.

Information Diffusion in a Mobile-First Context. Classic models of information diffusion, from the two-step flow of mass media [14] to network science theories of "weak ties" [11], laid the groundwork for understanding how ideas spread. The digital era, with its massive datasets from mobile phones and social media, has allowed for these theories to be tested and refined at an unprecedented scale. Seminal work using mobile phone records, for instance by Blumenstock et al. [4], has demonstrated that communication networks in the Global South are not random; instead, they are highly structured by offline social realities like wealth and ethnicity, and these structures fundamentally shape economic outcomes and behavior [5]. This research established that the structure of the network itself is a critical factor in social processes.

More recent work has focused on the nature of the information being spread, proposing that not all diffusion is simple contagion. The theory of complex contagion, for example, argues that adopting new behaviors or complex ideas, unlike a simple virus or rumor, often requires reinforcement from multiple sources within a dense social cluster [6]. This helps explain why certain types of information remain highly localized and struggle to cross community boundaries. This stands in contrast to the spread of simple, non-controversial information which can travel vast distances through weak ties.

Our findings contribute a crucial, content-based perspective to this literature by empirically demonstrating the parallel operation of these distinct diffusion modes on a single platform. This bimodal pattern highlights how mobile-first, group-centric platforms simultaneously support both the high-fidelity, trust-based communication needed for local coordination and the low-cost, broad dissemination of a shared cultural and political discourse.

3 Data Collection

Our analysis is based on a novel dataset that links digital communication traces from WhatsApp with computationally-derived observations of users' physical environments. The data was collected in two stages, covering online user activity and offline environmental context.

3.1 WhatsApp data

The data was collected across 100 locations in Uttar Pradesh, India's most populous state with over 220 million people. These locations, which span urban, semi-urban, and rural areas, were strategically selected to ensure a representative sample of the state's demographic diversity in terms of age, religion, and caste, benchmarked against census data. For simplicity, we refer to all locations as 'towns'. Figure 2 shows the locations of the 100 towns on a map.

Within each town, our survey team conducted in-person visits to recruit participants. Following an informed consent process approved by our institution's ethical review board (IRB), 3,127 participants agreed to donate their WhatsApp data using a specialized tool developed by [9]. In addition to their WhatsApp data, each participant provided their precise home geolocation and completed a demographic survey covering age, income, education, caste, religion, and household asset ownership. For analytical tractability, we aggregated detailed demographic data into a standardized set of broader categories. This process involved collapsing fine-grained survey responses into variables suitable for modeling, such as organizing age into four brackets, income into three tiers (Low, Medium, High), and residential areas into Village, Town, or City. The complete mapping from raw survey responses to these final categories is provided in Table 1.

The data collection spanned four months (April–July 2024). The final corpus contains approximately 5 million messages from over 6,500 WhatsApp groups (an average of 2.1 groups per user). For each message, our dataset includes the text content, timestamp, and anonymized user and group identifiers.

Ethical Considerations and Privacy. All data was collected following a protocol approved by our university's IRB, and all participants provided informed consent prior to donation. To protect participant privacy and minimize re-identification risk, all analyses in this paper were conducted at the town level. While our raw data contains precise geolocations, these were used only to link the online and offline datasets; all subsequent correlations and models use aggregated features, obscuring individual data points.

The smallest administrative unit in our sample has a population of over 30,000, ensuring a high degree of anonymity. In accordance with our ethical commitments, no individual-level data, particularly user geolocations, will be publicly released.

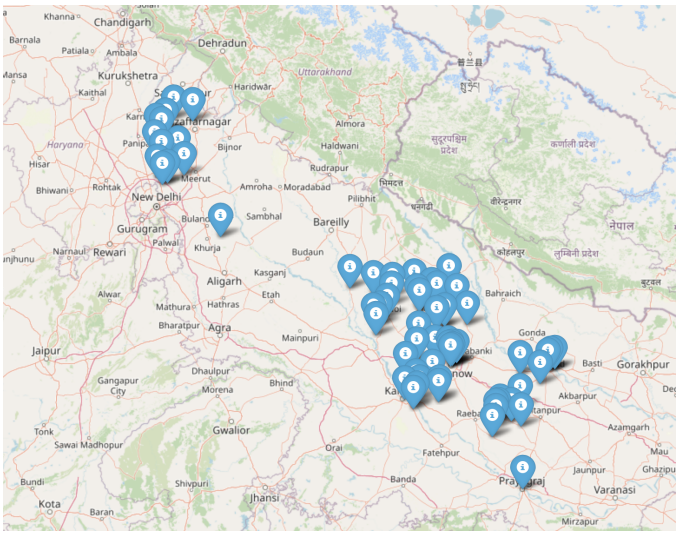

Figure 2: Geographic distribution of the 100 sampled towns across Uttar Pradesh. In each town, we collected data from around 30 users from their homes.

3.2 Offline data

The offline component of our dataset was computationally derived from the home geolocations provided by each participant. We used these coordinates to systematically query two key sources of environmental data: Google's Point-of-Interest data (POI) and Google Street View imagery (GSV).

3.2.1 Google Point of Interest data. To capture the local service and commercial environment, we queried the Google Places API for all establishments within a 200-meter radius of each of the 3,127 household locations. The raw API responses, which contained 81 distinct establishment "types" and 158 unique combinations of type labels, were too granular for direct analysis. To create a tractable feature set, we manually mapped the 81 unique POI types into 15 high-level, mutually exclusive categories relevant to daily life and economic activity (e.g., Healthcare, Retail, Religious; see Table 2 for the complete mapping). For each of the 3,127 locations, we then computed the total count of POIs falling into each of these 15 categories. The final dataset includes substantial variation across categories, from highly prevalent types like Retail (8,838 instances), Health care (5,731 instance), Religious (4,336 instances) to rarer ones like Beauty/Fitness (864), Government (493), and Transportation (52).

3.2.2 Google Street View Data. To visually characterize the built environment, we collected GSV imagery for each location. We implemented an automated "virtual walk" protocol to create a comprehensive visual record of the immediate neighborhood. Rather than relying on a single static image, this protocol captured multiple perspectives: (i) 360° View: Four images were taken from the origin coordinate at cardinal headings (0°, 90°, 180°, and 270°); (ii) Street Transect: Additional images were captured at successive points along the street's natural transect. An adaptive step-distance algorithm ensured that each image corresponded to a new, unique panorama, maximizing spatial coverage and preventing redundancy.

GSV imagery was available for 1,570 of the 3,127 locations (50.2%), covering 91 of the 100 sampled towns. This coverage rate reflects the relatively recent introduction and expansion of GSV in India and its limited availability in less accessible rural areas. To assess the potential for sampling bias introduced by this data attrition, we compared the POI characteristics of locations with and without GSV coverage.

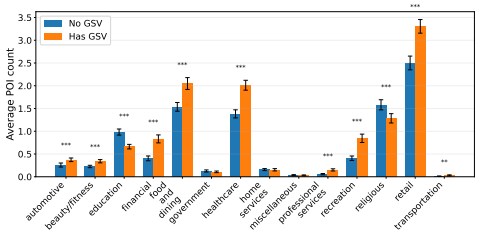

Figure 3: Mean POI counts for locations with and without Google Street View (GSV) coverage. Locations with GSV are in more commercially developed areas, while those without have slightly more education and religious POIs. Error bars represent 95% confidence intervals. $***$ on top of the bars indicate statistically significant $p < 0.001$ difference in means.

The comparison reveals a systematic difference: locations with GSV coverage are situated in more commercially developed areas, exhibiting significantly higher densities of Retail (Δ=+0.80,p <0.001), Healthcare (Δ=+0.63,p <0.001), and Food & Dining (Δ=+0.51,p <0.001) POIs. Conversely, locations lacking GSV had slightly more Education and Religious POIs. This indicates that our visual analysis is necessarily focused on the more developed and accessible segments of our sample, a factor we acknowledge in our interpretation of the results.

4 Methods

Our methodology is designed to systematically quantify and link the online and offline worlds. This section details the two primary analytical steps: first, the extraction of topics and metadata from multimodal WhatsApp content, and second, the computation of locality metrics from Google street view images.

4.1 Quantifying Online Communication: Topics & Engagement

To create a comprehensive picture of online activity, we first processed the content and metadata of all 5 million messages. A significant portion of the dataset (>60%) was non-textual, necessitating a multimodal approach. We unified all content into a standardized text format by generating detailed English descriptions for visual

media; we used the Gemma-27B model for images and a frame-by-frame analysis with Qwen-2.5-32B-VL for videos. This created a uniform corpus for analysis. We then employed an LLM-driven classification system, Qwen2.5-7B-Instruct, to systematically categorize each piece of content description. This system assigned both a single primary topic from a 14-category taxonomy adapted for the Indian social context (i.e., Religion, Politics, Entertainment, Commerce, Poetry, Employment, Violence, Health, Education, Environment, Greetings, Relationships, Sexual Content, and Miscellaneous) and a set of 5-10 secondary keywords that capture finer-grained semantic details like specific objects or named entities.

Alongside content analysis, we computed a suite of metrics to create a behavioral fingerprint for each of the 6,500 WhatsApp groups. These metrics characterized the group's interactional dynamics from four perspectives. We quantified the overall volume and temporal rhythm of conversation through raw message counts and burstiness indices. We measured content richness by calculating the distribution and entropy of different media types (text, image, video). To understand interaction and reciprocity, we analyzed the proportion of forwarded messages and the frequency of "tie-switching" between authors. Finally, we assessed conversational equality using the Shannon entropy and Gini coefficient of the author distribution to distinguish between equitable discussions and those dominated by a few individuals. We then aggregated all groups for a user in a specific location to obtain engagement metadata metrics per location.

4.2 Deriving Environmental Metrics from GSV Data

To convert the raw visual (GSV) and geographic (POI) data into structured profiles, we developed an automated annotation pipeline. Given the infeasibility of manually annotating all 1,570 locations, we leveraged a state-of-the-art multimodal model (Gemini 2.0 Flash) to generate a consistent, quantitative analysis for each location. The methodology involved developing a rigorous analytical framework, guiding the model with high-quality examples, and validating its output against human annotations.

First, based on iterative manual coding, we established a comprehensive data dictionary and scoring rubric to precisely defined each variable to be extracted, grouping them into three types: categorical keys requiring a selection from a predefined list (e.g., `locality_type`), indexable keys scored on a 1-5 Likert scale to capture quality (e.g., `infrastructure_condition`, `cleanliness`), and count-based keys for tallying visible features (e.g., `visible_sanitation_issue`). The rubric provided explicit decision rules and scoring criteria for each variable, specifying whether the model should use GSV images, POI data, or both. This entire framework was embedded within the model's prompt to ensure consistent interpretation. The complete list of keys is provided in Table 3.

To guide the model's performance, we employed a few-shot prompting strategy. We manually curated a "golden set" of 15 diverse locations, representing the full spectrum of environments in our dataset, from dense urban centers to undeveloped rural areas. For each of these locations, we created a meticulous, human-authored JSON annotation that served as an exemplar of the desired output. This curated set was used to instruct the model on the specific annotation task, effectively fine-tuning its analysis without altering the model's underlying weights.

With this framework in place, a unique, context-rich multimodal prompt was programmatically constructed for each of the 1,570 target locations. Each prompt began with the rubric instructions, followed by the 15 few-shot exemplars (including their filtered POI data, GSV images, and golden JSON annotations), and concluded with the target location's own POI and GSV data. The model was instructed to synthesize this information and return its analysis exclusively in a structured JSON format. This pipeline was executed iteratively across all locations, generating a rich, machine-readable dataset quantifying the offline environment. The complete prompt we used, containing the manual example annotations for a sample location can be found here.

Finally, to validate the reliability of this automated approach, we compared the AI-generated annotations against a human-annotated ground truth. A random sample of 50 locations was manually annotated by experts following the same rubric. We then measured the agreement between the AI and human reports, allowing for a tolerance of ±1 for indexable and count-based keys while requiring an exact match for categorical keys. The comparison revealed a high degree of concordance, confirming that our automated pipeline produced valid and reliable metrics for the subsequent analysis. Table 4 provides the performance of our Gemini model in annotating various categories in our GSV images.

5 Results

Our analysis proceeds in two stages. First, we establish the validity of our offline environmental metrics by correlating them with demographic data (Section 5.1). Next, we present various analyses to investigate the primary research question of connecting the offline lived environment and online WhatsApp behaviors (Sections 5.2-5.4).

5.1 Validation: Offline Metrics Reflect On-the-Ground Realities

A prerequisite for linking digital behavior to the physical world is establishing the veridicality of our offline metrics. To do so, we tested whether our novel offline measures, derived from GSV and POI data, robustly align with the demographic and socioeconomic structure of the 100 towns in our study. We aggregated all datasets to the town level and computed Spearman correlations between our offline features and a suite of demographic variables (detailed in Table 1).

First, our analysis demonstrates that GSV-derived features systematically map the infrastructural and developmental cleavage between urban and rural settings. 'City' type households anchor one end of this spectrum, exhibiting strong positive correlations with composite scores for development level (Spearman ρ =0.43), urbanization (ρ =0.39), and the condition of public infrastructure (ρ =0.37). 'Village' households define the opposite pole, showing significant negative correlations with these same metrics (ρ =-0.35). This core finding is further substantiated by specific visual features that trace the contours of daily life; for example, the prevalence of green spaces is tightly linked to agricultural proxies like tractor

ownership (ρ =0.35), while visible sanitation issues are, as expected, negatively associated with higher-quality housing (ρ =-0.27).

Second, POI densities provide a granular fingerprint of the local economic and civic landscape. The architecture of local commerce is clearly visible: a higher concentration of retail and banking POIs is a powerful predictor of household wealth, correlating strongly with asset ownership such as televisions and indoor toilets ($0.60 \leq \rho \leq 0.66$). Beyond commerce, these metrics reveal disparities in public infrastructure. More rural, 'Village'-classified areas exhibit lower access to healthcare facilities (ρ =-0.28) but a higher density of religious sites (ρ =0.26), illustrating how the civic commons is constituted differently across the urban-rural divide.

These correlations are presented not as an exhaustive analysis but as a quantitative foundation, which we supplemented with a qualitative audit. We manually inspected the top and bottom 20 correlations and confirmed that the overwhelming majority represented intuitive, real-world relationships. This verification process ensures our metrics are grounded in observable reality and are not statistical artifacts.

With the fidelity of our offline lens established, our findings confirm that these metrics distill the complex physical environment into a valid set of quantitative signals. We therefore proceed to the core of our inquiry: investigating the systematic relationship between this quantified physical world and the digital interactions that take place within it.

5.2 The Interplay Between Offline Environments and Online Behavior

Having validated our offline metrics, we next investigated their direct relationship with online activity. Simple bivariate correlations between environmental features and WhatsApp behaviors were generally weak (typically $|\rho| \approx 0.1$), a likely consequence of confounding demographic variations between towns. To isolate the specific influence of the physical environment, we therefore computed partial correlations, controlling for a comprehensive set of demographic factors. This approach allows us to compare towns with similar population structures, revealing how differences in their physical surroundings systematically align with their digital lives. This controlled analysis revealed several significant relationships.

A town's economic fabric appears to shape the substance of its online conversations. A higher density of general POIs—a proxy for commercial vibrancy—was a powerful predictor of discussions related to education, a relationship that remained highly significant after controlling for demographics ($\rho_{\text{partial}} \approx 0.50, p \approx 1.4 \times 10^{-6}$). Separately, the concentration of financial institutions was significantly associated with an increase in religion-focused messaging ($\rho_{\text{partial}} \approx 0.42$). These findings suggest that the ambient presence of a formal economy, independent of residents' own socioeconomic status, fosters specific thematic discourses.

Beyond conversational topics, the physical quality of public infrastructure correlated strongly with the style and modality of digital interaction. Better road conditions, as measured from GSV images, were not linked to transportation talk but rather to greater conversational equity—that is, more evenly distributed participation among authors in WhatsApp groups ($\rho_{\text{partial}} \approx 0.40$). Similarly, higher-quality public amenities predicted a greater entropy in the types of media shared ($\rho_{\text{partial}} \approx 0.33$), indicating a richer and more diverse media-sharing environment. This suggests that better public infrastructure may facilitate more inclusive and expressive forms of online engagement.

Strikingly, environmental cues related to health and sanitation were linked to privacy-seeking behaviors online. Both a higher density of healthcare POIs and more visible sanitation issues in GSV images were significantly correlated with an increased use of revoked or ephemeral messages (ρ_{partial} ranging from 0.36 to 0.38). This suggests that physical environments that prime residents to be mindful of health—whether through the presence of services or the visibility of risks—may foster more cautious and private communication patterns.

Finally, the online-offline relationship is not always one of direct reflection; in some cases, the physical world appears to suppress related digital discourse. For instance, poor infrastructure was associated with less discussion about health ($\rho_{\text{partial}} \approx -0.42$), contrary to what a problem-oriented model of communication might predict. Likewise, a higher density of retail POIs correlated with less talk about outdoor activities ($\rho_{\text{partial}} \approx -0.36$), and more transportation POIs with slightly less commerce-related talk ($\rho_{partial} \approx -0.29$). These counterintuitive findings are crucial, as they underscore the complex, non-linear nature of the interplay between physical context and digital expression, confirming our results are not mere artifacts of demographic composition.

In sum, our analysis reveals a systematic and often non-intuitive relationship between the physical environment and digital communication. We find that the local economic landscape shapes conversational topics, public infrastructure quality influences interaction styles, and specific environmental cues can even prompt privacy-seeking behaviors. Crucially, this relationship is complex, with certain physical deficits appearing to suppress, rather than stimulate, related online discourse. These findings establish a clear directional link from the offline world to online behavior. This raises a pivotal inverse question: how strong is this connection in the opposite direction? To test this, we next explore whether these digital traces can, in turn, be used to predict the tangible characteristics of a town, thereby probing the symmetry of the online-offline relationship.

5.3 Predicting the Offline Environment from Online Data

Having shown that the offline world may shape online behavior, we next tested the inverse: how well can digital traces predict the observable, physical environment? To do this, we reframed the analysis as a prediction task. First, we constructed five composite indices to quantify the offline environment by integrating the GSV annotations and POI data. These indices, detailed in Table 5, were standardized (z-scored) to represent: (1) Cleanliness & Health, (2) Economic Vitality, (3) Civic & Political Visibility, (4) Religious & Cultural Presence, and (5) Public Services Access. We then used LASSO regression models to predict each index from three distinct feature sets: (i) user demographics, (ii) WhatsApp metadata, and (iii) conversation topics. In our results, we report the standardized coefficients (β) from our regression models, which represent the

change in standard deviations of the outcome for a one standard deviation change in the predictor.

Our analysis reveals a stark asymmetry. A town's demographic composition is a modest but significant predictor of its physical environment, whereas its population's digital behaviors offer virtually no standalone predictive power.

The performance of our models makes this gap clear. Demographics alone could predict a town's Economic Vitality with a cross-validated R^2 of 0.19, and also offered modest predictive power for Cleanliness & Health and Religious & Cultural Presence ($R^2 \approx 0.08$ for both). In sharp contrast, models using only WhatsApp metadata or topic distributions consistently failed to outperform a simple baseline ($R^2 \leq 0$), with the LASSO model often collapsing to an intercept-only solution. This highlights that while digital behavior is an outcome correlated with the environment, it is not a reliable proxy for it.

An examination of the model coefficients reinforces the primacy of demographics and reveals which features were most salient:

- Economic Vitality was best predicted by direct demographic markers of wealth, such as household appliance ownership ($\beta = 0.14$) and income ($\beta = 0.10$). This was the only index where WhatsApp usage provided even a faint signal, with the total number of users ($\beta = 0.11$) showing a small positive association.
- Cleanliness & Health was primarily associated with a town's urban character ($\beta = 0.10$ for city residents). Interestingly, while talk of "formal support" was positively correlated ($\beta = 0.09$), explicit health-related discourse was negatively correlated ($\beta = -0.07$), echoing our earlier finding that such topics may arise more from need than from the presence of good infrastructure.
- Public Services Access and Religious & Cultural Presence were also almost exclusively predicted by demographic variables like housing type, education, and income (β values ranging from 0.05 to 0.12).
- Civic & Political Visibility was the sole exception, being unpredictable by demographics but showing a weak link to conversation topics, such as discussions of formal support ($\beta = 0.07$) versus themes of violence ($\beta = -0.06$).

Collectively, these results demonstrate that the predictive link from the online to the offline world is primarily mediated by the underlying demographic structure of the population. The dynamics of digital conversations, on their own, are a faint echo, not a clear reflection, of the physical world.

5.4 The Geographic Dynamics of Information Spread

Finally, we broadened our scope from the static relationship between an environment and its internal conversations to the dynamic process of how information travels between towns. To trace these pathways, we first established a spatiotemporal origin for each unique topic and keyword, defined as the location of its earliest appearance in the dataset. This "ground zero" served as a fixed point from which we measured the geographic distance and time lag of every subsequent mention. This analysis reveals two distinct and parallel regimes of information diffusion.

The first regime, evident in broad primary topics, functions as a persistent, statewide broadcast layer. These general themes (e.g., commerce, health, politics) are not contained by local geography. On average, only 6.3% of a topic's message volume occurs within 5 km of its origin. The median distance for a message from its topic's point of origin is approximately 95 km, with a median time lag of roughly 5,000 hours (≈210 days). This pattern indicates that broad topics do not spread contagiously from village to village; rather, they are disseminated widely and resurface continually over many months, creating a shared discursive space across the entire state.

In sharp contrast, the analysis of specific secondary keywords reveals a bimodal, local-global distribution. This second regime consists of two distinct information flows: (i) Hyper-Local Bulletins: A significant subset of keywords is intensely localized. We find that 8.8% of high-volume keywords retain over 50% of their message traffic within a 5 km radius of their origin. These terms such as "working hours," "shutdown," and "supply" are characterized by short adoption lags (median 18–60 hours) and function as urgent, community-specific alerts about practical, local matters. (ii) Hyper-Global Memes: Conversely, another subset of keywords spreads just as broadly as primary topics. We find that 5.6% of keywords have less than 1% of their message volume near their origin. These terms, often related to aspirational or consumer content like "trading," "accessories," and "fashion," disseminate almost immediately to dozens of towns, behaving like non-local viral content.

Table 6 shows the raw statistics of the locality and spread of the primary topics. Tables 7, 8 show the raw statistics for most and least local 10 secondary keywords.

This bimodal nature is captured succinctly in the overall distribution: while the 25th percentile of keyword spread is 0 km due to the high volume of local messages, the median distance leaps to 69 km, reflecting the large number of keywords that travel widely.

This distinction has important implications. The structure of the digital network facilitates a "broadcast" pattern of long-distance hops for general topics, rather than a contagious village-to-village spread. At the same time, it supports a highly efficient mechanism for localized communication. The interplay between these two diffusion regimes offers a dynamic view of community information flow, where spikes in local-alert keywords could indicate emerging community stress or events, while the persistent statewide themes reflect more stable, structural concerns.

6 Discussion

This study set out to investigate the relationship between the observable physical world and private digital behavior. Our analysis, grounded in a novel dataset linking WhatsApp usage to AI-quantified Google Street View imagery in Uttar Pradesh, yielded four primary findings. First, we validated that our computational metrics of the offline environment are robust proxies for on-the-ground socioeconomic realities. Second, we discovered that specific environmental features are systematically correlated with online behavior; the local economy shapes conversational topics, public infrastructure influences interaction styles, and cues related to health and sanitation prompt privacy-seeking behaviors. Third, we found this relationship to be asymmetric: while the physical world leaves an imprint on the digital, online behaviors are a poor standalone

predictor of the physical world, a role far better served by demographics. Finally, we identified a dual-mode system of information diffusion, where broad topics operate as a statewide "broadcast layer" while specific keywords function as "hyper-local bulletins."

It is crucial to underscore that our findings are correlational and do not imply direct causality. Establishing a causal link—for instance, that poor road quality *causes* a change in conversational style—is exceedingly difficult with observational data, given the complex interplay of social, economic, and environmental factors. However, the absence of a simple causal claim does not diminish the importance of these relationships. Our findings are significant because they systematically uncover the latent environmental context of digital life, revealing the otherwise invisible structural factors that consistently co-vary with how communities communicate. These strong, systematic correlations demonstrate that the physical world is a powerful explanatory layer for understanding digital society, highlighting features that have predictive value even if the precise causal mechanisms remain a subject for future investigation.

The Environment as a Digital Primer. Perhaps our most insightful finding is that the physical environment does not merely mirror online discussions but appears to actively shape the nature of digital interaction in non-obvious ways. The link between better road quality and more equitable conversational participation—rather than more talk about transport—suggests that the quality of public infrastructure has a second-order effect on social dynamics. Well-maintained environments may foster a sense of public order and fairness that translates into more inclusive online dialogue.

Furthermore, the correlation between health-related cues (both positive, like clinics, and negative, like sanitation issues) and the use of ephemeral messages points towards a form of psychological priming. An environment that constantly reminds residents of health and hygiene may also induce a more cautious or private mindset, which then manifests in their digital communication practices. This suggests the online-offline link is not just a simple reflection of needs (e.g., bad roads leading to complaints) but a more subtle process where the ambient physical world sets the cognitive and social tone for digital life.

The Asymmetry of the Online-Offline Link. Our prediction task results offer a crucial dose of humility regarding the power of "digital footprints." The finding that demographics are far better predictors of the physical world than WhatsApp metadata or topics is significant. It suggests that while digital behavior is an outcome influenced by place, it is a faint and noisy echo, not a clear reflection. The underlying socioeconomic and demographic structure of a community remains the primary determinant of its physical character. This has important implications for a growing field of research that seeks to use social media data as a direct, real-time proxy for on-the-ground realities. Our work suggests that without robust demographic controls, such approaches may be unreliable, as the digital signals are heavily mediated by the population's composition.

A New Model for Digital Diffusion. The discovery of two parallel diffusion regimes offers a nuanced model for how information flows in a mobile-first, group-centric society. The "statewide broadcast" of broad, persistent topics on health, commerce, and politics functions as a shared discursive background, creating a sense of statewide community and sentiment. This slow-moving layer is a conduit for culture and identity. In contrast, the "hyper-local bulletins" about jobs, supplies, or shutdowns represent the network's logistical backbone, facilitating the urgent, practical coordination essential for community life. This dual structure implies that the same digital platform serves fundamentally different purposes simultaneously. For policymakers or observers, this means that tracking broad, viral topics may reveal long-term sentiment, but understanding acute, local events requires listening for spikes in these fast-moving, small-radius keywords.

Limitations and Future Directions. While this study offers several novel insights, its limitations should be acknowledged. First, our analysis is correlational and cannot establish causality. Future work could explore natural experiments, such as over time analysis of GSV data to obtain before-and-after studies of infrastructure projects, to isolate causal effects. Second, the incomplete GSV coverage means our visual analysis is concentrated in more developed and accessible areas; while we analyze this bias, it remains a constraint. Third, our study is situated in a single Indian state, and the generalizability of these specific findings to other cultural and technological contexts requires further research. Finally, data donation, while ethically robust, may introduce self-selection biases. Future research should aim to replicate these findings in diverse geographic settings, incorporate longitudinal data to track changes over time, and integrate other data sources, such as public health or economic records, to further enrich the online-offline link.

7 Conclusion

This research demonstrates that the streets we walk and the communities we inhabit have a tangible, measurable imprint on our digital lives. By developing a scalable methodology to see and quantify the physical world through the eyes of AI, we have shown that the connection between our offline and online worlds is not only real but also complex, systematic, and often surprising. The digital traces of society are not created in a vacuum; they are a faint but interpretable echo of the physical world in which they are produced.

References

[1] Luca Maria Aiello, Anastassia Vybornova, Sándor Juhász, Michael Szell, and Eszter Bokányi. 2025. Urban highways are barriers to social ties. *Proceedings of the National Academy of Sciences* 122, 10 (2025), e2408937122.

[2] Syeda Zainab Akbar, Anmol Panda, and Joyojeet Pal. 2024. Political hazard: Misinformation in the 2019 Indian general election campaign. In *Political Campaigning in Digital India*. Routledge, 133–151.

[3] Eric Bellman. 2017. The End of Typing: The Next Billion Mobile Users Will Rely on Video and Voice. https://www.wsj.com/articles/the-end-of-typing-the-internets-next-billion-users-will-use-video-and-voice-1502116070. [Accessed 08-10-2025].

[4] Joshua Blumenstock, Gabriel Cadamuro, and Robert On. 2015. Predicting poverty and wealth from mobile phone metadata. *Science* 350, 6264 (2015), 1073–1076.

[5] Joshua Blumenstock and Lauren Fratamico. 2013. Social and spatial ethnic segregation: A framework for analyzing segregation with large-scale spatial network data. In *Proceedings of the 4th Annual Symposium on Computing for Development*. 1–10.

[6] Damon Centola. 2010. The spread of behavior in an online social network experiment. *science* 329, 5996 (2010), 1194–1197.

[7] Kieran Elrod, Katherine Flanigan, and Mario Bergés. 2025. Street View Sociability: Interpretable Analysis of Urban Social Behavior Across 15 Cities. *arXiv preprint arXiv:2508.06342* (2025).

[8] Jon E Froehlich, Alexander Fiannaca, Nimer Jaber, Victor Tsara, and Shaun Kane. 2025. Streetviewai: Making street view accessible using context-aware multimodal ai. *arXiv preprint arXiv:2508.08524* (2025).

[9] Kiran Garimella and Simon Chauchard. 2024. WhatsApp explorer: A data donation tool to facilitate research on WhatsApp. *Mobile Media & Communication* (2024), 20501579251326809.
[10] Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei. 2017. Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States. *Proceedings of the National Academy of Sciences* 114, 50 (2017), 13108–13113.
[11] Mark S Granovetter. 1973. The strength of weak ties. *American journal of sociology* 78, 6 (1973), 1360–1380.
[12] Rishi Iyengar. 2018. The future of the internet is Indian — edition.cnn.com. https://edition.cnn.com/interactive/2018/11/business/internet-usage-india-future/. [Accessed 08-10-2025].
[13] Neal Jean, Marshall Burke, Michael Xie, W Matthew Alampay Davis, David B Lobell, and Stefano Ermon. 2016. Combining satellite imagery and machine learning to predict poverty. *Science* 353, 6301 (2016), 790–794.
[14] Elihu Katz, Paul F Lazarsfeld, and Elmo Roper. 2017. *Personal influence: The part played by people in the flow of mass communications.* Routledge.
[15] Nikhil Naik, Scott Duke Kominers, Ramesh Raskar, Edward L Glaeser, and César A Hidalgo. 2017. Computer vision uncovers predictors of physical urban change. *Proceedings of the National Academy of Sciences* 114, 29 (2017), 7571–7576.
[16] Robert J Sampson and Stephen W Raudenbush. 1999. Systematic social observation of public spaces: A new look at disorder in urban neighborhoods. *American journal of sociology* 105, 3 (1999), 603–651.
[17] Robert J Sampson, Stephen W Raudenbush, and Felton Earls. 1997. Neighborhoods and violent crime: A multilevel study of collective efficacy. *science* 277, 5328 (1997), 918–924.
[18] Don E Schultz. 1998. The Death of Distance-How the Communications Revolution will Change Our Lives. *International Marketing Review* 15, 4 (1998), 309–311.
[19] Ali Akbar Septiandri, Marios Constantinides, and Daniele Quercia. 2024. WEIRD ICWSM: How Western, Educated, Industrialized, Rich, and Democratic is Social Computing Research? *arXiv preprint arXiv:2406.02090* (2024).
[20] Clifford Robe Shaw and Henry Donald McKay. 1942. Juvenile delinquency and urban areas. (1942).
[21] Yuri Takhteyev, Anatoliy Gruzd, and Barry Wellman. 2012. Geography of Twitter networks. *Social networks* 34, 1 (2012), 73–81.
[22] Barry Wellman. 2001. Physical place and cyberplace: The rise of personalized networking. *International journal of urban and regional research* 25, 2 (2001), 227–252.

Table 2: POI Keyword to Google Maps Category Mapping

POI Keyword	Google Maps Categories
government	local_govt_office, post_office, courthouse, city_hall, police
religious	church, hindu_temple, mosque, synagogue, place_of_worship
healthcare	hospital, pharmacy, doctor, dentist, physiotherapist, drugstore, vet, health
professional	lawyer, real_estate, travel_agency
education	school, university, library, primary_school, secondary_school
financial	bank, atm, accounting, insurance, finance
home_services	contractor, electrician, painter, moving, laundry
food_dining	restaurant, cafe, bakery, bar, meal_delivery, night_club, liquor_store, food
recreation	park, tourist_attraction, movie_theater, lodging
beauty/fitness	beauty_salon, hair_care, spa, gym
automotive	car_repair, gas_station, car_dealer, car_rental, car_wash
transportation	transit_station, train_station, parking
retail	shopping_mall, supermarket, dept_store, clothing, shoe, electronics, book, hardware, home_goods, furniture, jewelry, pet, bicycle, florist, convenience, store
miscellaneous	storage, cemetery
general	point_of_interest, establishment

A Demographic variables

Table 1 categorizes the demographic variables.

B Variable definitions

B.1 Point of Interest variables

Table 2 shows the POI categories mapping.

B.2 GSV Variables

Table 3 shows the GSV variables and their definitions.

C Validation of the GSV image annotation

Table 4 provides the performance of our Gemini model in annotating various categories in our GSV images. The Field with N/A in Tolerant column are categorical type varibles requiring exact match. We fixed a minimum of 60% match for each field and improved our results to cross this threshold over different attempts.

D Offline indices

Table 5 shows the variables we used to construct the five offline indices.

Table 1: Categorization of Demographic Variables

Variable	Category	Composition / Description
Age	18–24	Four distinct age brackets.
	25–34	
	35–44	
	45+	
Religion	Hinduism	The two predominant religions in the sample.
	Islam	
Caste	General	Based on self-reported social groups.
	Non-Hindu	
	Other Backward Caste	
	Scheduled Caste	Combines Scheduled Caste and Scheduled Tribe.
Education	Low / Basic	No Schooling, Primary (Up to 5th), Middle (5th–9th).
	Secondary	10th Pass, 12th Pass, Vocational College.
	Higher	University First/Higher Degree, Professional Education.
Monthly Income (INR)	Low Income	Below 10,000 INR.
	Medium Income	10,000 – 25,000 INR.
	High Income	Above 25,000 INR.
Residential Area	Village	Village, Small Town.
	Town	Medium Town.
	City	Metropolitan City, Large City.
House Type	Basic	Kutcha, Slum/Hut, Semi-Pucca, 1–2 Room House/Flat.
	Medium	3 Room House/Flat.
	Large	4+ Room House/Flat, Pucca House.
Resources	car/jeep/van, scooter/motorcycle/moped, air_conditioner computer/laptop/ipad, fan/cooler, washing_machine, fridge television, bank/post_office_account, atm/debit/credit_card lpg_gas, home_internet_connection, indoor_toilet pumping_set, tractor	Resources the user has access to. Binary variable for each variable.

Table 5: Index Composition of Offline Environmental Factors

Index	Contributors
Cleanliness & Health	cleanliness, health_facility, public_amenity, infrastructure_condition, green_spaces, poi_healthcare, poi_home_services, (–) sanitation_issue
Economic Vitality	development_level, urbanization, building_density, public_amenity, road_condition, poi_financial, poi_retail, poi_professional, poi_food_dining, poi_transportation
Civic & Political	politics_messaging, infrastructure_condition, public_amenity, road_condition, poi_government, poi_professional, poi_transportation, poi_miscellaneous
Religious & Cultural	religion_presence, green_spaces, poi_religious, poi_recreation, poi_education
Public Services	education_presence, public_amenity, health_facility, green_spaces, poi_education, poi_healthcare, poi_transportation, poi_home_services, poi_government

(-) indicates negative contributor

E Geographic Dynamics of Information

Tables 6 show all the primary topics. Tables 7, 8 show the top and bottom 10 secondary keywords respectively and their geographic and temporal spread. Notes: all percentages are shares of message volume per topic/keyword; median lags are hours between a topic/keyword's first appearance anywhere and its appearance in the reported town. Tables show the extremes of locality versus statewide broadcast while meeting a ≥200-message support threshold.

Table 3: GSV Variables and Scoring Rubric

GSV Variable	Possible Values	Description
locality_type	Residential, Commercial, Mixed-Use/Institutional, Industrial, Transportation, Agricultural, Open-Undeveloped	Categorizes the purpose or setting of the locality
building_density	1-5 (Integer)	Describes the relative space occupied by the buildings
building_height	1-5 (Integer)	Score for the average height of buildings
road_type	paved/unpaved/mixed	Dominant road material visible
road_condition	1-5 (Integer)	Physical quality of the road surface
infrastructure_condition	1-5 (Integer)	Score for the general state and quality of fixed public structures
development_level	1-5 (Integer)	Score for the overall physical and economic maturity of the location
urbanization	1-5 (Integer)	Score reflecting the general degree of urban influence
economy_status	deprivedclass, lowerclass, middleclass, upperclass, mixed	Visually assessed socioeconomic class of the primary residents or activity
primary_economy	present/absent	Indicates the presence/absence of Primary Sector activity (e.g., farming)
secondary_economy	present/absent	Indicates the presence/absence of Secondary Sector activity (e.g., manufacturing)
tertiary_economy	present/absent	Indicates the presence/absence of Tertiary Sector activity (e.g., services, retail)
dominant_sector	unclear/primary/secondary/tertiary	The single most visible economic sector driving the area
health_facility	1-5 (Integer)	Composite score for the availability and proximity of health services (visual and POI)
cleanliness	1-5 (Integer)	General tidiness and lack of visible litter
green_spaces	1-5 (Integer)	Abundance and quality of natural or cultivated greenery
religious_diversity	present/absent	Indicates the presence/absence of visible symbols/structures from multiple distinct religions
politics_messaging	1-5 (Integer)	Visibility and frequency of political posters/banners
education_presence_score	1-5 (Integer)	Composite score for the density and accessibility of educational institutions
religion_presence_score	1-5 (Integer)	Prominence of religious structures or symbols
visible_sanitation_issue	Count	Direct count of observable sanitation problems (e.g., open drains)
public_amenity	Count	Direct count of publicly accessible shared structures (e.g., street lights)

Table 6: Primary topics

Topic	Towns	Origin (%)	≤ 5 km (%)	≤ 10 km (%)	Median lag (h)
Commerce	92	2.9%	2.9%	2.9%	4759
Education	88	4.9%	4.9%	4.9%	4917
Employment	80	3.1%	3.1%	3.1%	5003
Entertainment	91	5.9%	5.9%	5.9%	5097
Environment	91	5.2%	5.2%	5.2%	5006
Greetings	93	7.6%	7.6%	7.6%	4884
Health	92	5.7%	5.7%	5.7%	5183
Miscellaneous	95	9.8%	10.2%	10.2%	5143
Poetry	83	11.8%	11.8%	11.8%	5094
Politics	84	6.0%	6.0%	6.0%	4894
Relationships	92	12.6%	12.8%	12.8%	5108
Religion	92	4.6%	4.6%	4.6%	4934
Sexual Content	52	0.7%	0.7%	0.7%	5722
Violence	87	3.8%	3.8%	3.8%	5287

Table 7: Secondary keywords – most local

Keyword	Messages	Towns	Origin (%)	≤ 5 km (%)	Median lag (h)
normal	200	7	94.0%	94.0%	1271
supply	351	13	91.2%	91.2%	2637
triangular flags	234	18	78.2%	91.0%	0
shutdown	349	27	77.1%	77.1%	2737
working hours	280	16	74.6%	74.6%	1420
salary range	316	11	72.5%	72.5%	1140
pigeon	220	6	69.1%	69.1%	1413
feeder	238	10	66.0%	66.0%	2841
linked devices	205	19	57.1%	64.9%	23
job posting	288	19	58.3%	58.3%	1239

Table 4: Comparison of Human vs. AI for GSV Annotations

Field	Exact Match Rate	Tolerant (±1) Rate
locality_type	60%	N/A
building_density	53%	97%
building_height	77%	100%
road_type	83%	N/A
road_condition	20%	87%
infrastructure_condition	80%	100%
development_level	67%	97%
urbanization	47%	100%
economy_status	67%	N/A
primary_economy	90%	N/A
secondary_economy	97%	N/A
tertiary_economy	77%	N/A
dominant_sector	67%	N/A
health_facility	33%	73%
cleanliness	37%	73%
green_spaces	37%	90%
religious_diversity	63%	N/A
politics_messaging	67%	87%
education_presence_score	47%	93%
religion_presence_score	47%	87%
visible_sanitation_issue	40%	77%
public_amenity	53%	100%

Table 8: Secondary keywords – least local

Keyword	Messages	Towns	Origin (%)	≤ 5 km (%)	Median lag (h)
trading	468	31	0.2%	0.2%	1841
cap	464	22	0.2%	0.2%	4481
individuals	452	69	0.2%	0.2%	4789
dashboard	391	41	0.3%	0.3%	2373
accessories	749	58	0.3%	0.3%	1697
SIM card	341	27	0.3%	0.3%	802
sandals	339	33	0.3%	0.3%	4567
pigeons	331	7	0.3%	0.3%	3270
embroidery	318	41	0.3%	0.3%	4085
creativity	615	73	0.3%	0.3%	4843