Digital Pulse of Development: Constructing Poverty Metrics from Social Media Discourse[⋆]

Woojin Jung[a,*], Andrew H. Kim[a], Charles Chear[a], Vatsal Shah[a], Ying Hung[b] and Tawfiq Ammari[c]

[a] Rutgers University, School of Social Work, 120 Albany St, New Brunswick, New Jersey, 08901, United States
[b] Rutgers University, Department of Statistics, 501 Hill Center, 110 Frelinghuysen Road, Piscataway, New Jersey, 08854, United States
[c] Rutgers University, School of Communication and Information, 4 Huntington St, New Brunswick, New Jersey, 08901, United States

ARTICLE INFO

Keywords:
ICTD
Poverty
Development
Social media
Spatio-temporal analysis
Kriging
Topic model

ABSTRACT

Assessing poverty and the specific needs of local populations in a contextualized manner presents significant challenges. This is especially true in emerging economies, where data collection infrastructure is lacking. Using the case of Zambia, this paper builds a language model to explore Twitter discourse as a proxy for poverty metrics. To our knowledge, this is the first study to demonstrate how mixed methods topic modeling, with feature selection guided by domain knowledge, can produce parsimonious and explainable poverty prediction models. Our analysis reveals that Twitter discourse captures salient development issues across regions and time. In particular, poorer villages tend to discuss immediate, local development concerns, whereas wealthier villages are more likely to engage with broader, more abstract development concepts. In addition, the development-related content of tweets has predictive power in estimating the wealth of the villages to which they are geotagged. Parsimonious, interpretable topic-based features account for more than 60% of the variation in village-level wealth. Furthermore, we explore imputation methods aimed at improving data utility in data sparse contexts. The spatial interpolation method, kriging, outperforms commonly used imputation methods while also providing uncertainty quantification that guides adaptive sampling of social media data. Based on these findings, we propose a novel pipeline, building information infrastructure for citizens as sensors. By leveraging social media discourse, development becomes more participatory and inclusive.

1. Introduction

Measuring spatial poverty is essential for distributing aid and resources to areas of need (Bird, 2019; Jung, 2023). To effectively support communities that are in extreme poverty or affected by economic shocks, local governments and international development agencies need granular contextualized information detailing the needs of these communities in a timely manner (e.g., Gebru et al. (2017)). However, poverty-related data are not always available or representative at the point of interest (POI). The scarcity of surveyed wealth, income, and consumption data is a major challenge. Therefore, it is necessary to develop systematic and granular poverty measures that are consistently and frequently available in any locale across the world. A novel data source is social media (Niu et al., 2020), which is the focus of this study. We will examine the potential of using social media data, particularly Twitter (also known as X), as an indicator of poverty metrics using the case of Zambia.

One of the most well-known social media platforms is Twitter. Its microblog data provide public opinions on economic, social, and political issues (Jamali et al., 2019; Leetaru et al., 2013), serving as a bridge between citizens and government officials (Blasi et al., 2022). Specifically, the spatio-temporal distribution of textual and citizen-generated data provides high-resolution information that complement traditional development indicators in development studies. For example, studies examine the volume of tweets and nightlight luminosity (NTL) as predictors of gross domestic product (GDP) globally (Indaco, 2020) and wealth levels in sub-Saharan Africa (Kondmann et al., 2020). Beyond the

[⋆] This paper has received the Honorable Mention Paper Award at EAAMO'25.

[*] Corresponding author.

✉ wj153@ssw.rutgers.edu (W. Jung); ahjkim92@ssw.rutgers.edu (A.H. Kim); charles.chear@rutgers.edu (C. Chear); vatsal.b.shah99@rutgers.edu (V. Shah); yhung@stat.rutgers.edu (Y. Hung); tawfiq.ammari@rutgers.edu (T. Ammari)

http://jungwoojin.org (W. Jung)

ORCID(s): 0000-0001-9179-2573 (W. Jung); 0009-0006-1408-4964 (A.H. Kim); 0000-0003-0795-5064 (C. Chear); 0009-0007-4593-3540 (V. Shah); 0000-0001-7298-0966 (Y. Hung); 0000-0002-1920-1625 (T. Ammari)

volume of social media features, efforts have also been made to build classifiers to predict development-related topics in Africa using Twitter and YouTube (Harriet et al., 2024). Despite its potential applications, the literature on the use of Twitter and social media data as a proxy for development, wealth, and poverty remains limited.

Three major gaps are identified: (1) using social media data to analyze wealth distribution in low and middle income countries (3) integrating geo-tagged social media data as features in wealth prediction; and (4) the management of of missing social media data.

In the realm of development-related topics, there is a relatively large body of research on disaster response (Nia et al., 2022) and public health crises (Boon-Itt et al., 2020). However, fewer studies explain how text-based features contribute to wealth prediction. Among the limited studies that focus on wealth, most either lack qualitative review by humans or concentrate on developed countries. In detail, research on developing countries predominantly relies on the quantity or raw count of tweets to predict poverty metrics (Kondmann et al., 2020; Walk et al., 2023). In prior work, such as (Harriet et al., 2024), some identified topics related to development and poverty fall outside the commonly accepted conceptual understanding of these topics. A potential explanation is that topic classifiers were built without qualitative human input.

In terms of geographic coverage, many analyses are conducted in the US (Giorgi et al., 2023a; Livermore et al., 2022; Jiang et al., 2018) or European countries (Bartelmeß et al., 2024; Abitbol and Morales, 2021; Levy Abitbol et al., 2018; Blasi et al., 2022; Preoţiuc-Pietro et al., 2015a) where ground-truth surveys or other data sources are relatively abundant. Studies highlight differences between high- and low-income users in developed countries. Higher-income users tend to engage in discussions about politics, technology, and corporations whereas lower-income users tend to focus on entertainment and sports (Abitbol and Morales, 2021; Preoţiuc-Pietro et al., 2015a), and express emotions more openly, using informal language (Preoţiuc-Pietro et al., 2015a). However, the topics and posting frequencies among lower-income users in developing countries may differ, necessitating further investigation. Other studies show that the aggregation of digital traces can predict well-being when compared to Gallup survey results in US counties (Jaidka and Ahmed, 2015; Giorgi et al., 2023a). Another study shows that topic modeling, combined with census and satellite imagery, can infer socioeconomic development (Levy Abitbol et al., 2018). However, the utility of social media to address data gaps is relatively lower in advanced countries, where reliable alternatives such as census data, income tax records, and other surveys are readily available and more accessible.

While some studies have focused on emerging economies, their analyses are typically conducted at a higher geographic level, such as cross-country studies (Kondmann et al., 2020; Bartelmeß et al., 2024) or city-level (Indaco, 2020) than at a finer, small area level. For example, Wikipedia embedding with NTL is used to predict wealth between countries (Sheehan et al., 2019). In fact, only a few studies make explicit use of the native geographic coordinates attached to tweets. Despite the potential of large geotagged data sets, fine-grained regional analysis is challenging due to non-representative samples (Giorgi et al., 2022) and data sparsity (Hoover and Dehghani, 2020) in small/remote areas. More granular analysis is needed to understand the use of social media data as a proxy for development.

Lastly, missing data is a challenge when using social media data. Twitter data often lack observations at a spatial region of interest where ground-truth labels exist. This absence is especially pronounced in traditionally data-deficient areas, where novel data is expected to complement traditional poverty indicators. While techniques such as using a user's profile location can help infer their whereabouts, not all users provide profile information, and even when they do, the data is often limited to coarse, city-level locations rather than precise coordinates.

To our knowledge, few methods test the accuracy of geospatial interpolation by leveraging spatial structure. One study used Gaussian processes, specifically kriging, to interpolate life satisfaction in small US counties by combining Twitter latent Dirichlet Allocation (LDA) topic data with census demographics, addressing the lack of representative Gallup survey data (Giorgi et al., 2023a). In data-rich settings, the abundance of tweets across thousands of counties enables Twitter data to boost predictions of life expectancy when combined with census data. However, the question of whether Twitter data can be effectively interpolated in data-sparse environments, both in terms of Twitter activity and ground-truth surveys, has largely been unexplored.

Based on these motivations, our paper aims to explore how development-related discourse in Twitter informs regional development and poverty in Zambia. Specifically, we ask three research questions.

> **RQ 1.** How can topic model features trained on Twitter data articulate differences across provinces and village-level wealth?

> **RQ 2.** To what extent can Twitter topic features provide parsimonious and explainable models for predicting village-level wealth?

RQ 3. How well do different imputation methods reconstruct missing social media data and replicate prediction accuracy?

Unlike remote sensing data (e.g., satellite imagery data (Oshri et al. (2018); Elmustafa et al. (2022))) or crowd-sourced maps (e.g., Open Street Maps identifying points of interest like hospitals or schools (Muñetón-Santa and Manrique-Ruiz, 2023)) which are traditionally used in wealth prediction models, social media data provide signals for development agendas contextualized in the words of different stakeholders. Using social media data as signal for citizen interests (Goodchild, 2007), our first study (§4) focuses on analyzing how Twitter topics can provide descriptive wealth signals that show (1) differences across spatial dimensions (i.e., how topics are weighted in different provinces and other administrative locales); and (2) wealth distribution (i.e., wealthy, medium, and poor villages). We find that poorer villages tend to focus on practical local concerns, whereas wealthier villages are more inclined to discuss broader, macro-level policy issues.

The second study (§5) focuses on predicting wealth using a set of development-related topics derived from the Twitter topic model. We find that using a subset of only seven topics qualitatively determined to be development-related provides an explainable model with a minimal loss of performance. This shows that refined quantitative information on topic characteristics can serve as a proxy measure of poverty. We proceed to show how explainable Twitter data can boost predictive performance in models using other traditionally used development proxies (e.g., NTL and the Normalized Difference Vegetation Index) while also providing explainability.

Finally, to mitigate data sparsity in social media data, we simulate the effectiveness of different spatial imputation methods in §6. Across a wide range of missing proportions, kriging interpolation demonstrates strong performance. When Twitter data are missing completely at random, borrowing information from spatial structures helps reduce information loss. Additionally, we propose how missing values can be imputed from the kriging model to use data efficiently and create better models in the future.

Our paper advances the use of Twitter topic model features in wealth prediction. We explore the distribution of topics and the mechanisms linking them to development and poverty. Another key contribution is our innovative approach to analyzing social media corpora, uniquely integrating qualitative and quantitative analyses with interpretable machine learning techniques. To our knowledge, this is the first study to leverage topic features from topic modeling to contextualize wealth indices within a mixed-methods framework. Additionally, we introduce spatial interpolation methods for unstructured social media data to address challenges in highly data-deficient environments, where both ground-truth data and alternative sources like social media archives may be scarce. Imputation methods for social media data remain underexplored in existing literature, making our study a valuable contribution to this area.

In this research, we start by reviewing the literature in §2. We then present the datasets employed in our work and the language model we built based on the Twitter text in §3. The next section (§4) visually explores how our Twitter topics are represented in the geographic space of Zambia, linking topic features with their regional context (§4.2). We add to the regional context by analyzing wealth-related differences in §4.3. In §5, we present our wealth prediction models, followed by the imputation simulation in §6. Finally, we describe our contributions, design and policy recommendations in §7.

2. Related Work

In this section, we review early research in the field to establish social media—particularly Twitter—as an emerging source for measuring development, poverty, and wealth.

2.1. Situating Social Media in the Context of Development, Poverty, and Wealth

Development is often framed as economic growth aimed at significantly reducing poverty and social inequality (Sachs, 2006). However, Amartya Sen's seminal work on development as freedom (Sen, 2014) broadens this perspective. According to Sen, development extends beyond economic progress or income growth. It is a holistic process encompassing human capabilities, social and environmental transformation, and political empowerment. In this view, poverty is understood as a fundamental deprivation of well-being, standing in direct contrast to development. Traditionally, poverty is measured through asset or wealth deficiencies, serving as a proxy for long-term economic welfare. This approach is particularly relevant in contexts where reliable data on consumption, expenditures, and price levels are lacking (Sahn and Stifel, 2003). In our study, we approach development, poverty, and wealth not as isolated concepts but as interconnected dimensions, using them interchangeably to capture their complex relationships.

Recent literature in development studies highlights the role of state institutions and societal capacity in shaping the trajectory of a country's economic development (Dincecco, 2017; North and Weingast, 1989; Acemoglu et al., 2015; Besley and Persson, 2010; Cingolani, 2018; Savoia and Sen, 2015). Daron Acemoglu and James Robinson argue that nations thrive when there are inclusive institutions encouraging participation in economic and political activities by a population at large (Robinson and Acemoglu, 2012). Concurrently, a significant paradigm shift in international development emphasizes capacity-building and evidence-based policymaking. A key example of this sea change is the World Bank's toolsets for rapid data collection, designed to enhance the effectiveness of poverty measurement and reduction initiatives (Yoshida et al., 2022).

Since the early 2010s, these transformations, alongside technological advancements, have driven the emergence of novel approaches in computational socioeconomics (Gao et al., 2019). A key aspect of this evolving development paradigm is the active participation of individuals in collecting, interpreting, and contextualizing data about their daily experiences (Rowlands, 1995, 1997). Academic interest in leveraging social media data for measuring development continues to grow. This research spans various data sources, including metadata (Fatehkia et al., 2020) and user-generated content (Indaco, 2020). Methodologies developed at both global and national scales position social media as a cost-effective alternative or a complementary tool to traditional poverty measurement approaches (Fatehkia et al., 2020; Leidig and Teeuw, 2015).

A key advantage of social media data is its high-frequency and high-resolution nature, which enables the inference of development levels and priorities across space and time. Poverty and needs assessments conducted at smaller community levels can produce more precise evaluations due to the relative homogeneity of these units compared to larger administrative areas (Jung, 2023). Such granular analyses are invaluable for policymakers in designing, monitoring, and evaluating public resource allocation based on localized needs. In contrast, traditional poverty assessments often rely on aggregated data, constrained by the lower spatial resolution and frequency of standard survey methodologies (Jung, 2023). Common surveys in developing countries, such as the Demographic and Health Surveys (DHS) and Living Standards Measurement Studies (LSMS), typically provide geographic data only at broad administrative levels, such as states or districts, due to limited sample sizes. Additionally, data quality issues, such as missing data in census-based measurements, pose significant challenges in many developing contexts (Carr-Hill, 2013; Kuffer et al., 2022). The increasing use of social media in developing countries (Poushter et al., 2018) further enhances its potential as a valuable data source for detailed spatiotemporal analyses. For instance, as of early 2023, Zambia had 2.7 million social media users, representing 13.3% of its population, with Twitter alone accounting for 0.15 million users (0.8% of the total population). [1]

2.2. Imputation Methods for Spatial Data in Social Media

Despite its promising potential as a proxy for poverty measurement, using social media features in wealth prediction models is not without its challenges. Social media platforms tend to have a higher representation of users located in large urban areas, males, and specific language groups (Lasri et al., 2023; Mislove et al., 2011; Jiang et al., 2018), leading to an oversampling of a socioeconomically advantaged segment of the population (Hargittai, 2020). This selection bias, coupled with data quality, necessitates a multifaceted research methodology that incorporates quantitative, qualitative, and data science approaches. The application of natural language processing techniques, particularly topic modeling, remains limited in this context.

Additionally, social media data are often incomplete, posing a significant challenge. The reliability of such proxies heavily depends on data quality, making it essential to address missingness through rigorous imputation techniques. In statistics, various imputation methods have been developed to estimate and replace missing data (Murray, 2018; Little and Rubin, 2002). To effectively impute missing social media data, it is crucial to leverage spatial autocorrelation within existing data. For example, there is a strong correlation between Twitter topic counts and their spatial locations. To account for this spatial dependency, kriging—also known as the Gaussian process model—is employed as an imputation technique (Stein, 1999). This method effectively utilizes spatial patterns to enhance data completeness, improving the robustness of social media-based poverty estimation.

2.3. Twitter as a Virtual Public Square

Recognized as one of the most publicly accessible platforms for social media data, Twitter is a site of public discourse and political engagement. Through the use of hashtags, Twitter allows for the creation of ad hoc publics (Lynn et al., 2020; Bruns and Burgess, 2011), which enable people to cohere the discourse over a topic of interest

[1] https://datareportal.com

and an imagined audience (Litt and Hargittai, 2016). This, in turn, creates what Boyd (Boyd, 2010) terms *networked publics*, a group of people whose discourse and norms are shaped by networked technologies like Twitter. Twitter has been used to study public discourse in Information and Communications Technology for Development (ICTD) scholarship - c.f., (Pal, 2017; Panda et al., 2020; Meena et al., 2020; Das et al., 2023b).

Research has explored the use of Twitter data in studying poverty and development, highlighting its intersection with various dimensions, including economic, social, environmental, and political aspects. In terms of economic development, Twitter data has been employed for cross-country economic analyses, with tweet volume serving as a proxy for economic outputs such as GDP and gross national product (GNP). This approach is particularly useful in regions where reliable economic statistics are scarce (Indaco, 2020; Nia et al., 2022; Kondmann et al., 2020). The political dimension, especially the relationship between Twitter and public engagement, has also received significant attention. Studies across three African countries suggest that more democratic nations exhibit higher levels of public policy discourse on Twitter (Best and Meng, 2015). Additionally, Twitter has played a role in mobilizing political movements, such as the #FeesMustFall protests against poverty and corruption in South Africa (Ngidi et al., 2016). In India, the platform serves as a crucial space for political engagement (Jaidka and Ahmed, 2015) and as a tool for analyzing politicians' focus on development issues (Bozarth and Pal, 2019). Beyond politics and economics, Twitter has been leveraged to study social, human, and environmental development, including research on COVID-19, violence, and natural disasters (Osuagwu et al., 2021; Jongman et al., 2015; Simon et al., 2014).

3. Dataset and Methods for Topic Modeling

In this section, we describe our main datasets, the topic modeling process, and how we engineer features from the topic model. We start by introducing the ground-truth data (§3.1.1), which serves as the outcome we aim to predict using Twitter data. Next, we detail our Twitter dataset, including data collection and the process of building a language model (§3.1.2). Finally, we present topic features: topic weights and top topic count, and highlight seven development-related topics with illustrative example tweets (§3.2.3).

3.1. Data

3.1.1. Ground-Truth Wealth

In this study, we used the 2018 Zambian Demographic and Health Surveys (DHS) datasets. The DHS datasets serve as a standard for assessing asset-based poverty and various household indicators related to health, education, and living standards. They provide a composite measure of household living standards through two key variables: the wealth factor score and the wealth index. The DHS household-level dataset includes 12,831 nationally representative samples. To facilitate analysis at a broader spatial scale, household wealth data were aggregated and averaged across the 535 georeferenced village clusters.

Subsequently, we spatially joined the Twitter data with our ground-truth data, using the georeferenced DHS village cluster. To achieve this, we established a 10 km radius around each village cluster centroid, assigning Twitter features within this buffer to the corresponding cluster. The buffer size was chosen to account for the geographic displacement of village cluster coordinates (Perez-Heydrich et al., 2013) and the varying pixel size of geospatial covariates. Through this joining process, our main dataset comprises 313 village clusters (See Figure 1).

3.1.2. Twitter Data

We compiled a corpus of Twitter data in Zambia spanning from 2019 to 2021. For data collection, we utilized the Twitter API V2 Academic Research product track, which enabled full-archive searches and provided a higher tweet retrieval cap. This yielded a total of 2,120,809 tweets. Each tweet was formatted as a JSON object containing named attributes and associated values, such as user metadata, tweet content, timestamps, and geolocation information where available

This API allows for country-level tweet selection, even for tweets that are not explicitly geolocated. Our analysis focused solely on unique tweets (excluding retweets) originating from Zambia, without filtering by specific keywords or hashtags. This gave us a total of 53,790 geo-located tweets. Having done so, we tagged each Tweet by geo-coordinates of village clusters (presented in §3.1.1), ward, district, and province.

Recognizing the challenges identified in previous ICTD research—particularly failures resulting from insufficient collaboration with local entities (cf. (Saha et al., 2022; Toyama, 2015, 2014))—we engaged with domain experts from Innovations for Poverty Action (IPA) and the Ministry of Community Development and Social Services (MCDSS) of

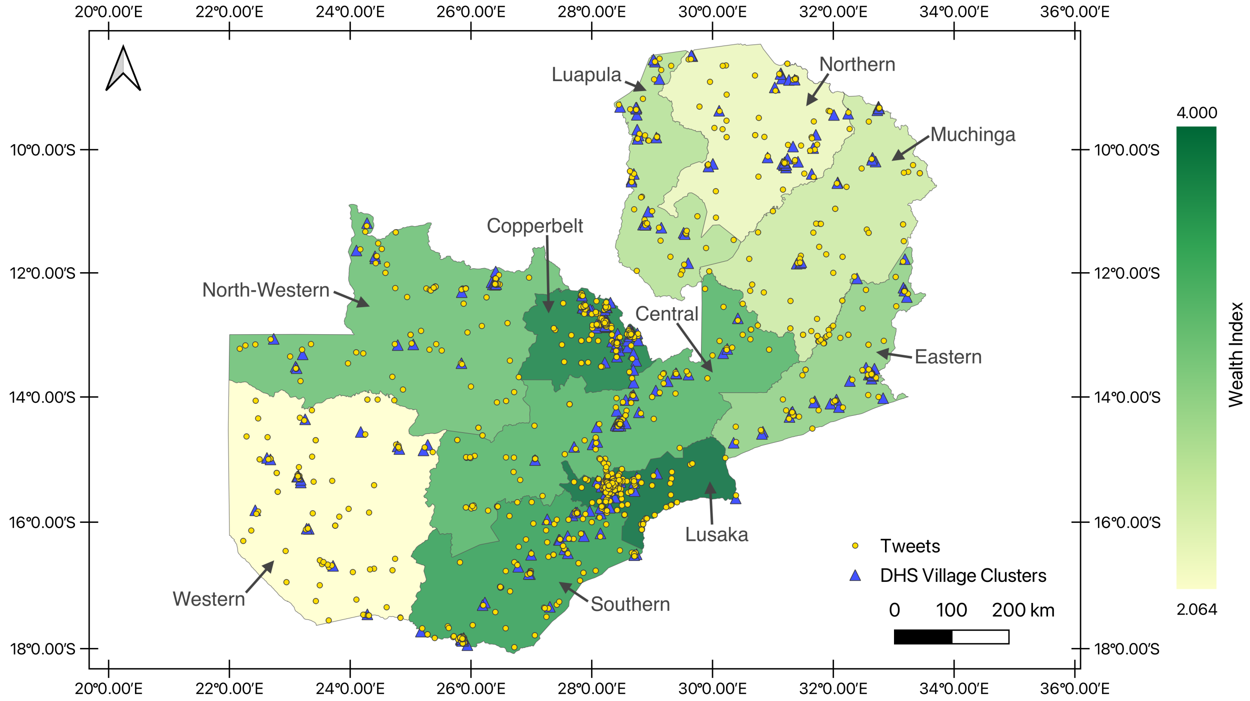


Figure 1: The distribution of village clusters. Blue triangles are village clusters included in the analytic sample (n = 313). Yellow circles are Twitter tweets (count = 20,235). Province colors denote province-level mean wealth, e.g., lighter is less wealthy, darker is more wealthy. Axes are geocoordinates (latitudes and longitudes).

Zambia. Their expertise facilitated access to the most up-to-date geospatial information and provided essential local context for our analysis.[2]

3.2. Topic Modeling with BERTopic

We employed Bidirectional Encoder Representations from Transformers (BERT) for contextualized topic modeling. BERT is a natural language processing (NLP) model that generates contextualized embeddings that enhance topic coherence (Reimers and Gurevych, 2019). To extract topics from these embeddings, we utilized the BERTopic package (Grootendorst, 2022). Traditional topic modeling techniques like LDA (Blei and Lafferty, 2006; Hays, 1998), often yield less interpretable results (Egger and Yu, 2022) due to the limited length of tweets (maximum of 280 characters) as well as the frequent use of characters such as hashtages (#) and mentions (@). To address this, we opted for a topic modeling approach based on clustering word embeddings. Word embeddings capture the semantic context of words by mapping corpus terms into a vector space where proximity represents semantic association (Mikolov et al., 2013). This aligns with the perspective that semantic space reflects meaningful word relationships (Griffiths et al., 2007), as cited in (Angelov, 2020).

3.2.1. Topic Models

Since word embeddings are represented in high-dimensional spaces, applying a dimensionality reduction method is necessary. This method must "preserve the pairwise distance structure amongst all data samples" while maintaining "local distances over global distances" within a corpus (McInnes et al., 2018, P.13). One such technique is UMAP (Uniform Manifold Approximation and Projection), which we employed with default hyperparameters (number of

[2] They played a crucial role in clarifying recent changes in the names and boundaries of districts and wards, essential details for spatial analysis. In addition to other contributions, domain experts helped standardize naming conventions and explained district changes, some of which resulted from redistricting. For instance, Mansa district was divided into two separate districts: Mansa and Chembe. Similarly, Chienge district is also spelled Chiengi, and Kapiri Mposhi district is sometimes referred to as Kapiri district. These clarifications were particularly important given that such changes have occurred rapidly, are not always intuitive, and are sparsely documented.

neighbors = 14 and distance metric = 'cosine').[3] After dimensionality reduction, contextually similar words and documents can be identified using a clustering technique, specifically Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN). HDBSCAN is particularly well-suited for this task as it is a density-based clustering algorithm (Campello et al., 2013). Similar documents in the corpus will be closer to each other. Each document will also be closer to words semantically closer to it. To generate the topic models, we utilized the BERTopic library.[4]

3.2.2. Topic Coherence and Topic Score

To determine the optimal number of topics for the BERTopic model, we trained 18 different models, systematically varying a single HDBSCAN hyperparameter: the minimum cluster size (Asyaky and Mandala, 2021). This hyperparameter directly influences the number of topics in each model, with larger minimum cluster sizes leading to fewer clusters and identified topics. The first model was initialized with a minimum cluster size of 15, while the last model had a cluster size of 195. To select the best-performing model, we computed the coherence score for each using the Gensim CoherenceModel feature (Řehůřek and Sojka, 2010). Coherence values have been shown to better approximate human ratings of a topic model's "understandability" (Röder et al., 2015) compared to other metrics, such as perplexity (Chang et al., 2009). The model with the highest coherence score (0.72) was selected, corresponding to a minimum cluster size of 115, which resulted in 103 topics. Given that more than 92% of tweets in our dataset are in English, we use an English-specific sentence embedding model (all-MiniLM-L6-v2) and we auto-translated the remaining Tweets. Established previous work shows that monolingual models perform better in single language corpora than their multilingual counterparts (Li et al., 2020; Chen et al., 2020).

3.2.3. Qualitative Analysis

To interpret each of the topics identified above, we qualitatively analyzed a sample of 50 tweets per topic (using topic scores as a measure of the Tweet valence). Following established practices in human-computer interaction (HCI), we relied on coder consensus rather than reporting interrater reliability scores (McDonald et al., 2019). Six coders, including the co-authors of this paper and graduate students from social science and computer science disciplines, conducted a thematic analysis of tweets from Zambia between 2019 and 2021. The analysis had two primary objectives: (1) to qualitatively assess the topics generated by topic modeling and (2) to identify the most relevant topics. Each coder was assigned a subset of topics and independently analyzed tweets associated with them. To support the coding process, we provided a list of keywords most representative of each topic along with a sample of 50 tweets where the topic weight was highest. Coders reviewed the same tweets through three rounds of analysis to ensure consistency and depth in thematic identification.

Sen's capability approach views poverty not just as a lack of income, but as a deprivation of the basic capabilities needed to live a life one values. Building on this perspective, the multidimensional approaches offer a broader understanding of poverty that goes beyond monetary measures (Alkire and Santos, 2010). In alignment with these principles, the Sustainable Development Goals (SDGs) provide a global framework for addressing interconnected social, economic, and environmental challenges. By interpreting the final set of topics through the lens of multidimensional poverty and the SDGs, we can better understand how Twitter discourse in Zambia captures the overlapping issues related to poverty and regional development.

The first round involved coders conducting individual analyses before meeting to compare their codes, leading to the consolidation or contestation of certain classifications. Some topics, such as 'wildlife' and 'sports', were deemed irrelevant and excluded. Meanwhile, multiple coders identified recurring themes related to government, food, and health, coding them accordingly. In the second round, coders revisited the tweets to refine topic labels and identify more precise distinctions. For example, topics initially categorized under 'government' were further differentiated into 'election corruption' and 'frustration with government' to capture more specific themes. During the third and final round, the coders reviewed the classifications of each other to further refine the topic labels. In one instance, a topic coded as 'food insecurity' by one coder and 'agricultural production' by another was reassigned as 'food systems' after multiple coders determined it to be a more accurate representation of the sampled tweets.

A final list of seven topics was generated: *election corruption, food systems, social progress, mining, frustration with government policy, public health-related challenges, and social inequality*. Table 1 shows sample tweets for each

[3] `https://maartengr.github.io/BERTopic/getting_started/parameter%20tuning/parametertuning.html#umap`

[4] `https://maartengr.github.io/BERTopic/index.html`

Table 1
The Seven Relevant Development Topics Along with Their Descriptions and Example Quotes.

Topic Name	Topic Description	Sample Tweet
Election corruption	Behaviors and choices of politicians that suggest corrupt motive.	Never!!! After that stupid move by the party, I can never even vouch for a politician. After asking people for donations for his campaign, he betrayed us and our constitution.
Food systems	Characteristics of a food system such as production, supply and demand, agriculture, and policies.	Food for thought #Zambia Is there a #food crisis or shortage of #maize ? We must figure the source of problems.
Social progress	Programs, policies, and social movements for development and democracy.	I am currently attending the inclusive innovation training as a beneficiary of the Healing Hands program hosted by university in Lusaka Zambia.
Mining	The mining industry in Zambia and other African countries.	The mining system forms a complex part of the #copper and #cobalt supply from lone workers, community groups to cooperatives. #Chinese companies in the #CopperBelt mining hubs take in artisanal units, which are crushed, bagged and loaded onto flatbeds.
Frustration with government policy	Frustration with government issues such as bureaucracy, the economy, and directionless.	My poor Zambia, how did you fall in the hands of the aimless one. Open your heart for the future of our children's children who, though not yet born, owe nations mammoth sums of monies. We can't run away, we must stay, we can change things starting next years elections.
Public health-related challenges	Challenges affecting the health of individuals and population.	Doctors at the University Teaching Hospital have successfully removed the 10 needles that had remained in John Mwa a five year boy old boy in the North-Dr John Makupe. #Zambia #Malawi #BREAKING #AA19
Social inequality	Inequality in society such as economic, social class, and gender.	we can have a wealth tax DRM strategy for Africa: If the world's rich were taxed 1% of their wealth in a year, billions could solve problems for the people. Does Africa have the capacity to tax her rich?

Note: Quoted tweets are disguised, per Bruckman (2002), to ensure the privacy of social media users in Zambia.

respective topic.[5] *Election corruption*, tied to SDG 16: Peace, Justice, and Strong Institutions, reflects governance issues that erode institutional trust. *Food systems*, which are linked to SDG 2: Zero Hunger, raise concerns about security, sustainability, and nutrition. Calls for *social progress*, reflected in SDG 10: Reduced Inequalities and SDG 5: Gender Equality, signal demands for greater rights and inclusion. Discussions of *mining* relate to SDG 8: Decent Work and Economic Growth, revealing tensions between growth and labor conditions. *Frustration with government policy*

[5] In keeping with earlier work on protecting user privacy, we follow Bruckman's (Bruckman, 2002) recommendation to engage in multiple levels of user disguise when quoting social media users in a research study. Thus, quoted tweets in this paper are disguised to ensure the privacy of social media users in Zambia.

reflects SDG 16: Peace, Justice, and Strong Institutions, and points to institutional failures that sustain poverty. This sentiment also appears in conversations on *public health-related challenges*, connected to SDG 3: Good Health and Well-Being, which highlight barriers to care and disease management. Across these areas, *social inequality*, which is related to SDG 1: No Poverty and SDG 10: Reduced Inequalities, underscores the structural disparities shaping poverty outcomes. Framing the thematic findings this way highlights that, in the sampled tweets, poverty extends beyond income to encompass political, social, and environmental dimensions.

3.3. Topic Features

Throughout this paper, we focus on two Twitter features. First, to analyze the regional distribution of tweets, we utilize Twitter topic weights or the count of the topic weights to build our models. Within our methods, we will identify whether we used topic weights, or the top topic counts. Both are described below.

3.3.1. Twitter Topic Weights

Topic distributions are averages aggregated at specific regional areas (e.g., province, low-income villages), an analysis of which shows distinct regional trends in public discourse across Zambia (Figure 2). Given how weights can present better distinctions between real-world values across different regions, different wealth areas (Kraak and Ormeling, 2020), and for determining the mean of variations over time (Killick et al., 2012), we use topic weights for our analysis in Study I (§3.1.2).

3.3.2. Twitter Top Topic Count

To calculate topic counts, we assign each tweet to the topic with the highest weight (see §3.2 for details). The total topic count for each village cluster is the sum total of tweets for each topic in a 10 km buffer from the village cluster centroid. The top topic count for village cluster v can be expressed as follows:

$$T_{c,v} = \sum_{i \in B_v} \mathbb{1}(T_i = \arg\max(W_i)) \tag{1}$$

where:

- $T_{c,v}$ is the total topic count for village cluster v,
- B_v is the set of tweets within a 10km buffer from the centroid of village cluster v,
- T_i is the topic assigned to tweet i,
- W_i is the vector of topic weights for tweet i,
- $\arg\max(W_i)$ selects the topic with the highest weight for tweet i,
- $\mathbb{1}(\cdot)$ is an indicator function that equals 1 if the condition is true and 0 otherwise.

Top topic counts improve decision tree and ensemble regression models because they simplify tree splits and thus make the models more interpretable (Breiman et al., 2017)[6]. Using top topic counts also contributes to a more stable kriging model given the rounding errors introduced by the skewed topic weights (Goovaerts, 2005). This value was employed in the classifiers presented in §5 and imputations presented in §6.

4. RQ1: Distribution of Twitter Data Across the Country

Variations in regional development influence Twitter usage patterns and topics. This section examines how Twitter features vary across locations and economic contexts.

4.1. Methods

To address RQ1, we examine whether key development issues across regions and wealth levels are reflected in people's tweets within their respective contexts.

[6] This is also the case for interpreting the coefficients of logistic regression models (Hastie et al., 2009)

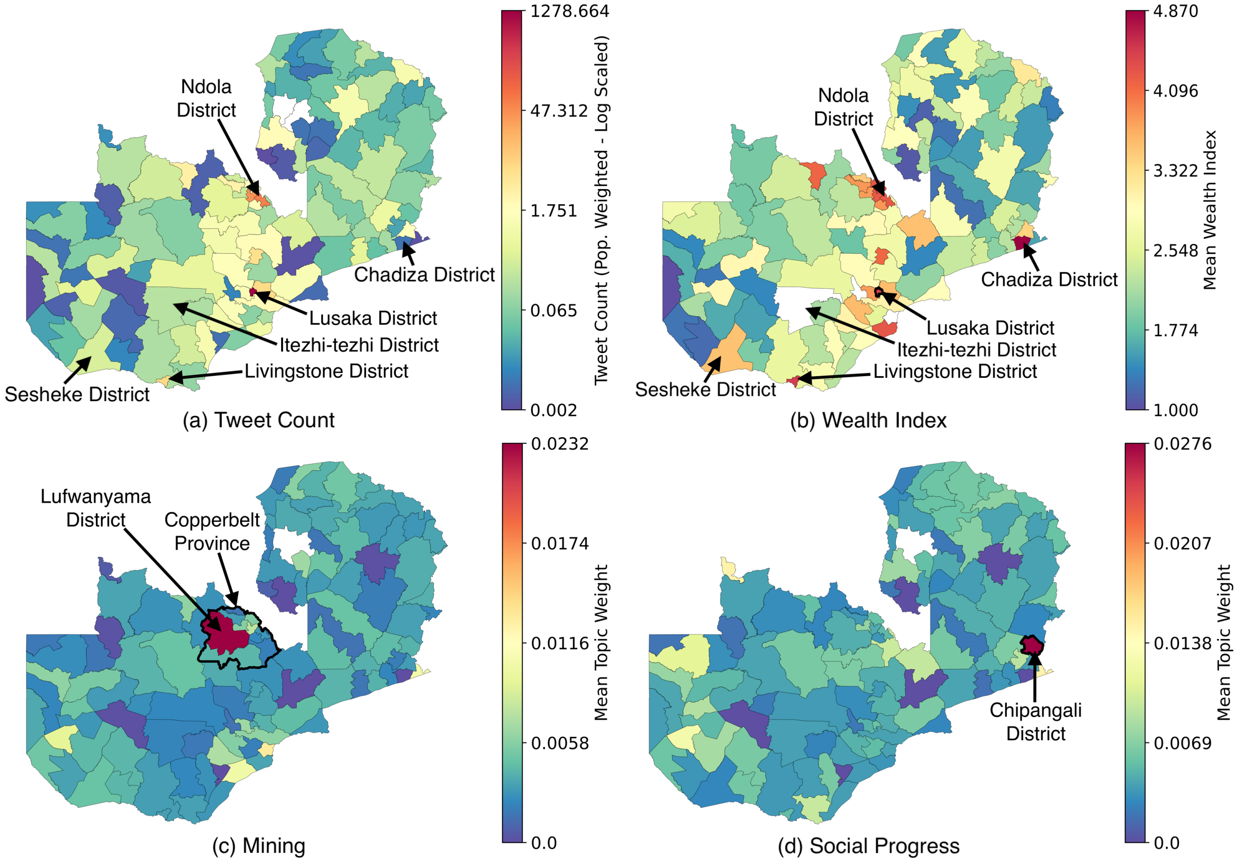
4.1.1. *Mapping Twitter Topics Spatially*

To analyze the topical signals in different provinces and districts, we averaged each topic weight per geographic region. We begin by using a district-level heat map to visually compare Twitter activity with the wealth index. We also map the distribution of specific topics such as mining and social progress, and qualitatively explore potential reasons why certain regions engage more intensively with these topics than others.

4.1.2. *Analyzing Topic Variation by Village-Wealth*

Development issues of interest differ between wealthier and less affluent groups. Social media topics reflect these differences through variations in topic salience across wealth groups. To capture these disparities, we construct a relative measure of tweet counts per topic by wealth tertile (high, medium, and low wealth), accounting for variations in absolute total tweet volume. We achieve this by standardizing the tweet count for each topic relative to all topics within each wealth tertile. We then compute the mean standardized tweet count for each of the seven development-related topics by wealth group. Given the unequal variance of topic distributions between rich and poor regions, we employ a Welch ANOVA test (Liu, 2015) with Games-Howell post hoc test to assess the significance of differences (Shingala and Rajyaguru, 2015). To further examine topic differences across wealth tertiles, we use the log-likelihood ratio to identify key distinctions. Finally, we sample tweets for qualitative analysis to contextualize these quantitative findings. The results are presented in §4.3.

4.2. Variation in Discourse Across Different Regions



(a) Tweet Count

(b) Wealth Index

(c) Mining

(d) Social Progress

Figure 2: (a) Displays the population-weighted tweet count distribution at the district level, where blue indicates lower concentrations and red indicates higher concentrations. (b) Illustrates wealth distribution by district. (c) Depicts the topic weight of the *mining* topic at the district level. (d) Represents the topic weight of the *social progress* topic at the district level.

Wealthier areas generally exhibit higher average tweet counts. As shown in Figure 2, tweet volume (a) correlates with wealth distribution (b) based on DHS data. Capital cities of Lusaka, Ndola, and Livingstone district, which are relatively wealthy, show dense tweet activity.

However, there are some districts with notable discrepancies between the surveyed wealth index and the Tweet count. District-level wealth estimates from survey data are not representative when they are sampled from a limited number of village clusters. In this case, Twitter data helps triangulate estimates and identify areas where validation from additional sources is needed. For example, the average wealth in Chadiza and Sesheke districts is derived from a single, wealthier village cluster, which may not accurately portray the overall wealth of the entire district. Triangulation with alternative estimations, such as small area estimation based on the 2015 Living Conditions Monitoring Survey and the 2010 census (de la Fuente et al., 2015), suggests that 80% of people in these districts are below the poverty line. As shown in plot (a), Twitter data seem to more accurately reflect the high poverty levels in these districts, indicated by the green and blue shading. In contrast, the DHS data in plot (b) appear to overestimate wealth in these areas, illustrated by the orange shading on the heat map. In cases where wealth data are not available, such as the Itezhi-tezhi shown in white in plot (b), Twitter data from plot (a) provide inferences about the likely wealth level of the district. The Twitter activity map indicates the Itezhi-tezhi district is relatively poor, which is supported by the Small Area Estimation showing a 70% headcount ratio.

We also examine the distribution of tweet topics (§3.2.3) based on their weights in each district (see §3.3.1). The Copperbelt Province, known for its mining industry, predominantly features tweets about *mining*, as shown in Figure 2(c). Tweets from these regions often highlight mining as a key source of employment and economic development for working-age individuals. In contrast, the Eastern and Central Provinces, particularly Chipangali District, frequently discuss *social progress* issues (Figure 2(d)). According to our partner organization (B Kabwela, July 19, 2022), the Innovations for Poverty Action, these conversations reflect long-standing deficiencies in social amenities and residents' expectations for improved basic services following government decentralization. The full distribution of development-related Twitter topics is provided in the Appendix A.

4.3. Variation in Discourse by Wealth

Significant differences exist in the topics villages tweet about based on wealth levels as shown in Figure 3. While the seven development-related topics (see Table 1) are, in general, discussed less frequently than the other 95 topics, especially in wealthier villages. Poorer villages tend to focus on concrete, localized issues within the seven development topics, contrary to their wealthier counterparts.

Overall, we find that middle-wealth and low-wealth villages focus on local policies and on-the-ground issues, whereas high-wealth villages engage more in macro-level national and, at times, regional issues. Poorer villages frequently discuss topics such as *election corruption*, *food systems*, and *mining*, while these issues receive the least attention in wealthier villages. Notably, in the wealthiest villages, public discourse-related topics, such as *social inequality*, are discussed more frequently relative to other topics.

The mean standardized tweet count statistically significantly differ across the three wealth groups in discussions on *election corruption*, *food systems*, *public health*, and *government policy*. For *mining* and *social progress*, significant differences were observed only for the wealthiest group. In contrast, *social inequality* did not show significant variation across wealth groups.

A closer examination of sampled tweets provides deeper context into how discussions vary across different wealth groups (see Appendix A for detailed analysis). Our analysis reveals the thematic and geographic patterns of Twitter discussions by wealth levels. In wealthier villages, discourse tends to be macro-level and globally oriented, often centered on Lusaka (the capital), national policy, and international comparisons. For instance, high-income village tweets on *election corruption* and *mining* mention Lusaka, the Copperbelt, and even the Kalahari Desert, indicating a broader geographic lens and interest in national or global systems. Similarly, their discussions on food systems and public health emphasize organized delivery systems, markets, research, and institutional healthcare, suggesting a higher degree of systemic awareness and access.

By contrast, middle- and low-income villages are more locally grounded, reflecting immediate, practical concerns and daily challenges. Middle-income tweets often connect local institutions such as hospitals and schools with regional identity, while still hinting at broader narratives such as #lifeinafrica. Low-income village discourse centers heavily on survival-related topics—food, electricity, spending, and loss—and frequently invokes NGOs like CARE Zambia, underlining both need and reliance on external support. Their language around election corruption, mining, and

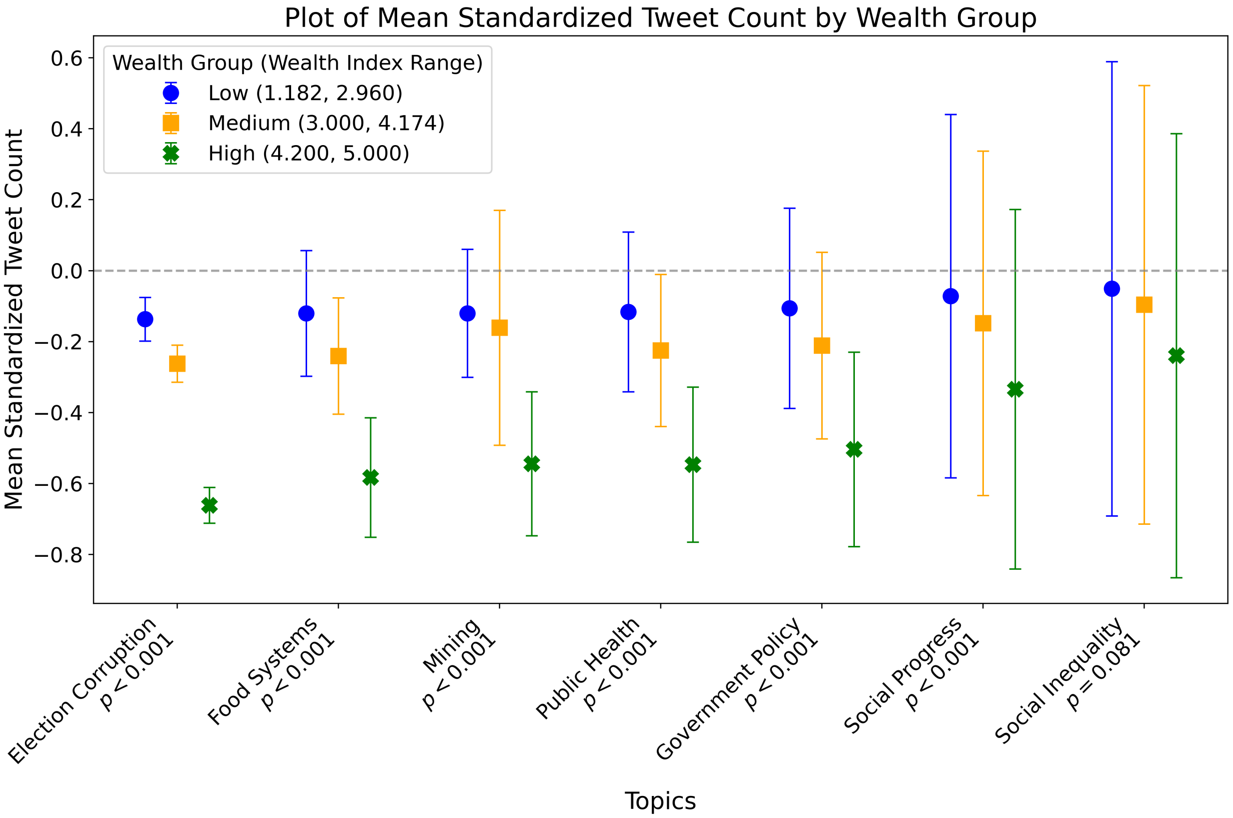


Figure 3: Mean standardized tweet count of 7 development topics with Welch ANOVA p value. The intervals represent ±1 SD.

government policy suggests an acute awareness of how political outcomes directly impact food systems, infrastructure, and local livelihoods.

5. RQ2: Wealth Prediction Using Twitter Topic Features

Recognizing that Twitter serves as an indicator of regional development, this section utilizes its topic features to predict wealth variation across villages.

5.1. Methods

We use machine learning to predict wealth based on Twitter topics and explore important topics. In addition, we compare the performance of Twitter data with other data sources and assess its added value when combined with them.

5.1.1. Features

To predict ground-truth wealth from the DHS surveys, we use the count of top topics (see §3.3.2) as features in our model. We varied the number of topics used as input features, aiming to create parsimonious models that balance complexity and predictive accuracy. The first model uses the seven development-focused topics identified earlier. The second model includes 14 topics: the seven development topics, plus seven uncorrelated topics. To reduce dimensionality in prediction and inference, we only retain topics with an absolute correlation of less than 0.6, Including highly correlated features can lead to the identifiability issue in model estimation, unstable predictions, and an increase in cost. Last, we use all 103 topics. Any clusters with missing Twitter data (n=222) were removed from the dataset for a final analytic sample of n=313. Later, we test a number of different imputation techniques in §6.

Table 2
Comparison of Machine Learning Model Architectures for Predicting Village-Level Wealth Using Twitter Topic Features

Architecture	Best Algorithm	R^2	mean val MSE	test MSE	test MAE
7 Development Topics	DT	0.610 ± 0.018	0.767 ± 0.006	0.421 ± 0.019	0.514 ± 0.011
14 Uncorrelated Topics	XGB	0.585 ± 0.009	0.682 ± 0.003	0.448 ± 0.009	0.513 ± 0.008
All 103 Topics	LGBM	0.646 ± 0.016	0.648 ± 0.006	0.382 ± 0.018	0.475 ± 0.013

Note: Sample size = 313. We show the performance of decision trees (DT), XGBoost (XGB), and LightGBM (LGBM) models across 30 iterations, evaluated by R^2, mean validation MSE, test MSE, and test MAE.

5.1.2. ML Approach

In our ML analysis, we investigate the performance of a wide range of regression models: regularized linear regressions, random forest (RF), decision tree (DT), light gradient-boosting machine (LightGBM), extreme gradient boosting (XGBoost), and multi-layer perception (MLP). We set an 80/20 random train/test split. We used Sklearn's standard scaler to scale data before training (Pedregosa et al., 2011). Using 5-fold cross-validation, we tune several hyperparameters for each model using the Tree-Structure Parzen estimator in Optuna with 50 iterations. We minimize the average mean squared error (MSE) across folds in the training set, which we call the mean validation MSE (mean val MSE). After training, we evaluate our model on the out-of-sample test set using the out-of-sample R^2, test MSE, and test mean absolute error (MAE) as our traditional ML metrics. In the case of gradient boosting algorithms and MLP, we use early stopping to prevent overfitting in the number of boosting rounds and epochs respectively. Table 2 shows the details of our trained models across 30 random seeds.

5.1.3. Explainabilty

Because we aim to assess how well social media language models can explain wealth distribution, we must also interpret the outputs of these models. As many machine learning models operate as black boxes, we apply post-hoc explainable AI (XAI) techniques to interpret their predictions of wealth values (Huang and Huang, 2023; Shoemaker et al., 2023). Specifically, we use Shapley values, which are rooted in game theory, to rank the most important features (Lundberg and Lee, 2017). We then analyze the partial dependence plots (PDPs) of the top features in each model using the PDP function from Scikit-learn (Pedregosa et al., 2011). PDPs visualize the relationship between a feature in a model and the target response while controlling for the average effect of the other model predictors (Moosbauer et al., 2021; Friedman, 2001).

5.2. Can we build predictive models using language features from topic modeling?

A parsimonious model not only demonstrates sound predictive performance but also offer greater interpretability. We compare three machine learning models using Twitter topic features for wealth prediction as shown in Table 3. The minimalist model with seven development topics achieves an R^2 of 61.0%, outperforming the model based on 14 uncorrelated topics and only slightly trailing the full model with all 103 topics. Our findings demonstrate that using linguistic features informed by domain knowledge (Canaydin et al., 2024; Lampos et al., 2017) can enhance model accuracy without sacrificing predictive power. In our case, constructing topic models from word embeddings provides the added benefit of generating features that are also explainable and contextualized.

We further investigate the most influential Twitter-derived features contributing to wealth prediction by analyzing their importance and direction of effect, as shown in Figure 4, which presents a SHAP beeswarm plot for the decision tree model.[7] Each point in the beeswarm plot represents a village cluster, with color indicating the value of the feature for that observation—where red represents a higher feature value and blue represents a lower feature value. Examining both the magnitude and direction of SHAP values allows us to understand not only which features are most influential, but also how their increases or decreases affect the predicted wealth outcomes. Topics such as *social inequality*, *public health challenges*, and *government policy* emerge as the most predictive features, with higher values (red points) associated with positive SHAP values, indicating that greater discussion of these topics tends to increase the model's wealth predictions. Conversely, topics like *food systems* and *election corruption* display a negative association, where higher values (red points) often correspond to lower predicted wealth. Interestingly, the *mining* topic shows minimal overall contribution to predictions, suggesting that, despite its relevance in certain regions, mining discourse

[7] Note that, as our top-performing model was a decision tree regressor, the feature contributions are not linear.

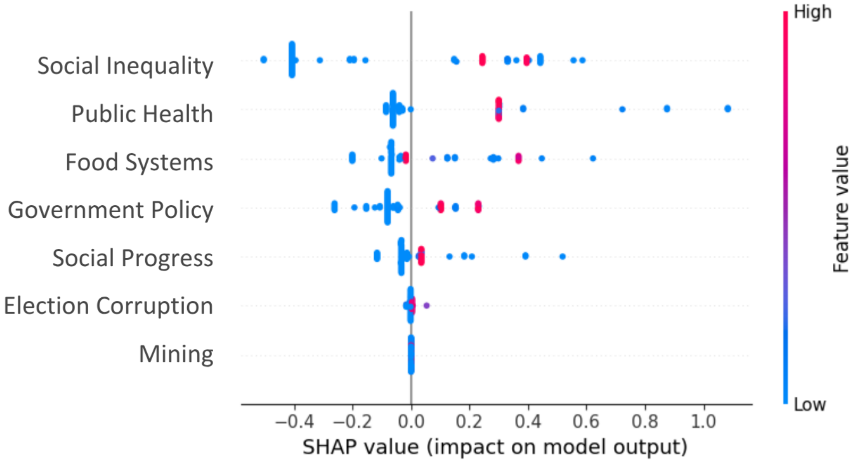
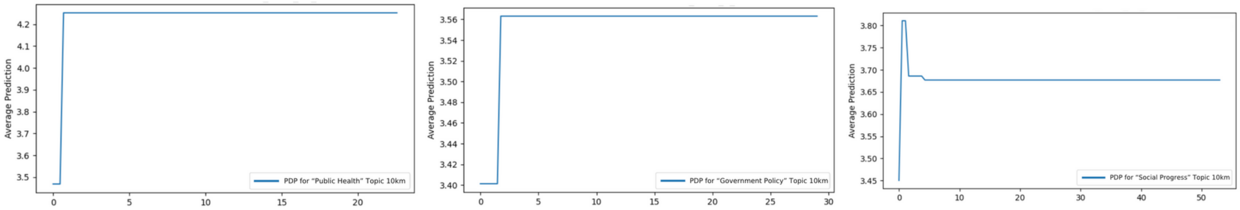


Figure 4: This figure shows a SHAP beeswarm plot highlighting the importance of seven development-related Twitter topics in predicting village-level wealth. Topics like social inequality and public health are linked to higher wealth, while food systems and election corruption are associated with lower wealth. These patterns suggest that wealthier villages engage in broader policy discourse, while poorer ones focus on immediate local concerns. The plot also shows that a small, curated set of topics improves interpretability.

is not a consistent signal of village-level wealth differences. This analysis underscores not only the uneven salience of development topics across socioeconomic contexts, but also how specific types of discourse systematically push wealth predictions higher or lower. In contrast to the interpretable patterns observed with the curated development-related topics, the most influential features in the full model using all 103 topics are less intuitively linked to conventional development indicators, complicating interpretation (see Appendix B for further details).

To understand how these features affect predictions on average, we examine partial dependence plots (PDPs) in Figure 5. These plots show the marginal effect of individual topics on the predicted wealth index, holding other variables constant. For example, Figure 5a shows that increased mentions of *public health* are associated with a consistent increase in wealth predictions. Similarly, *government policy* discourse (Figure 5b) has a steady positive effect. However, the *social progress* topic (Figure 5c) has a nonlinear influence—exerting a strong positive effect at low frequencies but plateauing as tweet counts rise. These results indicate that certain topics act as strong signals of wealth when they appear even infrequently, whereas others must reach a threshold before influencing predictions as presented in Figure 9 in the Appendix B.[8]



(a) Public Health (b) Government Policy (c) Social Progress

Figure 5: The partial dependency plots for the *Public Health*, *Government Policy*, and *Social Progress* tweet count within 10km from the best 7 development topic wealth index (WI) prediction model is presented. The x-axis is the tweet count while the y-axis is the average WI prediction. Here we see a single jump in predicted wealth as tweet counts increase.

[8] Further discussions with local experts and qualitative analyses could shed light on the contextual meanings behind these patterns and their differing contributions to the model.

Table 3
Predictive Performance of Models Using Different Spatial Features for Village-Level Wealth Estimation

Model	R^2	mean val MSE	test MSE	test MAE	Number of Features
Dist. to POIs	0.727 ± 0.010	0.409 ± 0.003	0.294 ± 0.011	0.437 ± 0.015	178
Internet Data	0.724 ± 0.018	0.428 ± 0.004	0.298 ± 0.019	0.430 ± 0.014	10
Satellite Imagery (Baseline)	0.705 ± 0.003	0.403 ± 0.002	0.319 ± 0.003	0.447 ± 0.004	768
Twitter (All 103 Topics)	0.646 ± 0.016	0.648 ± 0.006	0.382 ± 0.018	0.475 ± 0.013	103
Twitter (7 Development Topics)	0.610 ± 0.018	0.767 ± 0.006	0.421 ± 0.019	0.514 ± 0.011	7
Nighttime Lights	0.602 ± 0.012	0.538 ± 0.007	0.429 ± 0.013	0.484 ± 0.011	1
Building Density	0.520 ± 0.057	0.721 ± 0.028	0.518 ± 0.062	0.560 ± 0.050	1
Building Perimeter	0.519 ± 0.032	0.665 ± 0.008	0.519 ± 0.034	0.560 ± 0.027	1
Building Area	0.505 ± 0.006	0.672 ± 0.005	0.534 ± 0.006	0.571 ± 0.005	1
Structural Features	0.388 ± 0.079	0.827 ± 0.070	0.660 ± 0.085	0.655 ± 0.068	3
NDVI	0.186 ± 0.030	1.011 ± 0.013	0.879 ± 0.033	0.803 ± 0.031	1

Note: Sample size = 313. We compare R^2, mean validation MSE, test MSE, and test MAE across models built on Twitter topic features, satellite imagery, nighttime lights, POI distances, building footprint features, and NDVI, evaluated across 30 iterations. The Baseline model uses satellite image embeddings from a pre-trained Vision Transformer fine-tuned on nighttime lights, an established approach within the literature.

5.3. How Can Twitter Linguistic Features Boost Prediction Models?

While our study focuses exclusively on Twitter linguistic features, this section highlights the value of language models relative to other spatial feature sets. Twitter-based predictions perform comparably to other modalities, while also offering rich textual insights that shed light on development priorities across regions. Incorporating explainable language features from Twitter also boosts the predictive performance of models that rely on other spatial features. In this section, we compare several commonly used feature sets with Twitter features in §5.3.1 and then show how Twitter language features can enhance models using traditional features in §5.3.2.

5.3.1. The Twitter Features in Comparison to Other Traditionally Used Features

Satellite imagery, phone and connectivity data, point of interest (POI), and building footprint data have been used in the literature for wealth prediction (Jean et al., 2016; Blumenstock et al., 2015; Engstrom et al., 2021; Steele et al., 2017; Hu et al., 2022; Lee and Braithwaite, 2022; Pokhriyal et al., 2020; Reyes et al., 2023; Tang et al., 2018; Zhao et al., 2019; Jung et al., 2025a). Twitter models are comparable to other traditional sources of data in our case (see the Appendix C for additional data sources and extracted features). We compare our Twitter models by training models on our dataset using various features identified in the literature and a state-of-the-art baseline approach. When using all topic features, R^2 values of Twitter data (64.6%) outperform those of the nighttime lights (60.2%). When using seven development topics, it explains approximately 61.0% of the variation in wealth and is comparable to wealth prediction by nighttime lights. Similar to Twitter data, nighttime lights and internet data indicate access to electricity and information and communication-related infrastructure.

More advanced methods outperform simple nighttime lights or internet data models. As a baseline, we evaluate a leading approach that fine-tunes a pre-trained Vision Transformer (ViT) on nighttime lights data (Jung et al., 2025b), achieving 6% higher performance than the all-topic features model (R^2 = 70.5% vs. 64.6%). However, satellite-based embeddings lack human interpretability and require additional eXplainable AI techniques. In contrast, Twitter topic models offer a more transparent, intuitive alternative.

The performance of Twitter features exceed various building footprint features, such as the average building area, density, and structural features, e.g., orientation, shape, and compactness (R^2 ranging from 38.8% to 52.0%). Areas with dense, well-planned, and larger building footprints often indicate higher economic activity and better infrastructure. It is also substantially better than the Normalized Difference Vegetation Index (NDVI) or vegetation health index (R^2=18.6%). The NDVI is known to have an inverse relation with poverty in Africa (Sedda et al., 2015) as vegetation levels reflect agricultural productivity and environmental quality (e.g., degradation). While the predictive power of Twitter topics are lower than satellite images or distance to POIs (R^2=72.7%), they are more parsimonious than satellite embeddings and distance to POI features (see Table 3).

5.3.2. *Twitter Features Boost Predictive Models Using Traditional Features*

Beyond a single feature model, combining Twitter features with other data sources boosts prediction accuracy and interoperability (see Figure 6 and Table 5 in Appendix D). When using the NDVI, a feature that can be easily computed from Landsat-8 imagery, for instance, adding the Twitter feature improves its prediction by 43.5 percentage points. One reason for this might be that linguistic features can be more granular than NVDI features since NVDI can be "limited in certain phenological (seasonal changes) phases," and can be affected by atmospheric interference (NAX Solutions, 2022). In the literature, Twitter data and nighttime lights (NTL) are often combined for wealth prediction (Indaco, 2020; Kondmann et al., 2020). While NTL is a more consistently available and commonly used feature than tweets, prediction accuracy still improves by 5.4 percentage points when introducing all topics to the NTL predictive model. Introducing the Twitter features also increases the explainability of the model with the mixed features. Satellite embeddings are not only computationally costly to extract but also do not have any intrinsic meanings.

Additionally, even when combining all modalities, Twitter top topic count features still contribute to wealth prediction (see the Appendix D for more details). As this study focuses on Twitter features, a detailed multimodal analysis is beyond the scope of this paper. Nonetheless, our initial results indicate that Twitter features retain predictive power even when combined with other modalities.

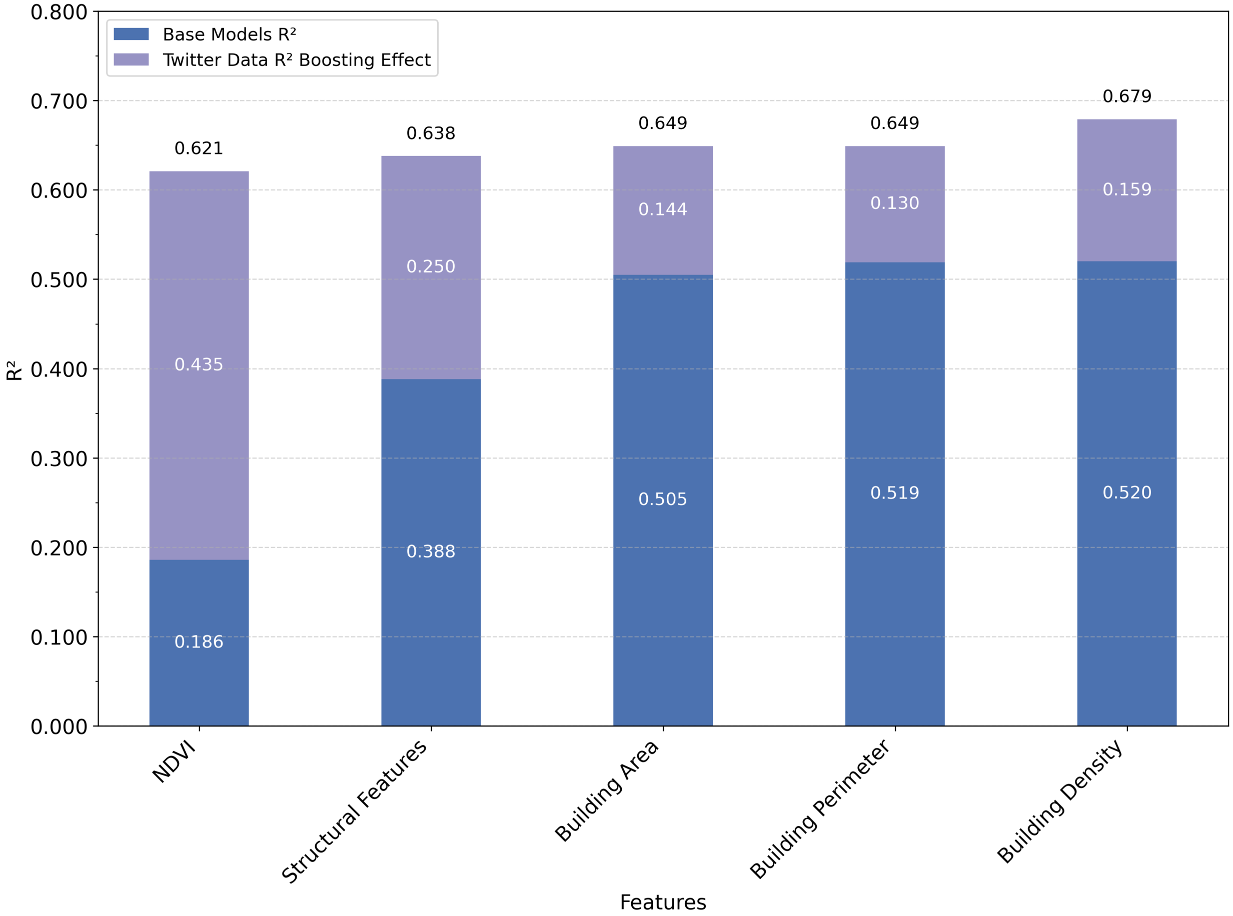


Figure 6: Top 5 Mean R² Boosts from Adding All Twitter Topic Features to Traditional Spatial Wealth Prediction Models

6. RQ3: Comparison of Imputation Methods to Handle Simulated Missingness

A common limitation of geolocated social media archives is missing data (Relia et al., 2018; Rehman et al., 2018). Ignoring missing data can result in biased inferences. Therefore, it is crucial to rigorously perform imputation,

Table 4
Imputation Simulation Study Using the Seven Development Topics

Imputation Method	Best Algorithm	R^2	mean val MSE	test MSE	test MAE
N = 313, Simulated MAR 10%					
Fill 0	XGB	0.545 ± 0.030	0.756 ± 0.012	0.491 ± 0.032	0.553 ± 0.021
Median	XGB	0.546 ± 0.032	0.757 ± 0.013	0.490 ± 0.034	0.551 ± 0.022
Kriging	RF	0.528 ± 0.043	0.769 ± 0.014	0.509 ± 0.047	0.553 ± 0.026
Rounded Kriging[a]	DT	0.586 ± 0.067	0.782 ± 0.017	0.447 ± 0.072	0.530 ± 0.044
Optimized Kriging	DT	0.520 ± 0.084	0.779 ± 0.018	0.518 ± 0.090	0.565 ± 0.050
N = 313, Simulated MAR 20%					
Fill 0	DT	0.501 ± 0.061	0.810 ± 0.037	0.538 ± 0.066	0.573 ± 0.034
Median	DT	0.496 ± 0.068	0.810 ± 0.037	0.543 ± 0.073	0.576 ± 0.040
Kriging	LGBM	0.537 ± 0.032	0.733 ± 0.018	0.499 ± 0.034	0.541 ± 0.022
Rounded Kriging[a]	DT	0.546 ± 0.080	0.786 ± 0.022	0.490 ± 0.086	0.550 ± 0.048
Optimized Kriging	LGBM	0.536 ± 0.032	0.737 ± 0.018	0.501 ± 0.034	0.545 ± 0.020
N = 313, Simulated MAR 30%					
Fill 0	XGB	0.518 ± 0.034	0.770 ± 0.018	0.520 ± 0.036	0.573 ± 0.023
Median	LGBM	0.529 ± 0.032	0.765 ± 0.017	0.508 ± 0.035	0.561 ± 0.022
Kriging	LGBM	0.524 ± 0.032	0.737 ± 0.016	0.514 ± 0.034	0.549 ± 0.022
Rounded Kriging[a]	DT	0.541 ± 0.079	0.787 ± 0.022	0.495 ± 0.085	0.552 ± 0.044
Optimized Kriging	RF	0.530 ± 0.041	0.762 ± 0.020	0.508 ± 0.044	0.552 ± 0.027
N = 313, Simulated MAR 40%					
Fill 0	DT	0.474 ± 0.053	0.829 ± 0.041	0.568 ± 0.057	0.597 ± 0.033
Median	DT	0.471 ± 0.060	0.828 ± 0.041	0.571 ± 0.065	0.598 ± 0.037
Kriging	XGB	0.510 ± 0.024	0.755 ± 0.019	0.529 ± 0.026	0.560 ± 0.016
Rounded Kriging[a]	RF	0.515 ± 0.057	0.778 ± 0.020	0.523 ± 0.062	0.562 ± 0.031
Optimized Kriging	RF	0.490 ± 0.057	0.773 ± 0.022	0.551 ± 0.062	0.576 ± 0.035
N = 313, Simulated MAR 50%					
Fill 0	RF	0.462 ± 0.051	0.809 ± 0.030	0.580 ± 0.055	0.603 ± 0.032
Median	XGB	0.448 ± 0.045	0.810 ± 0.026	0.596 ± 0.049	0.616 ± 0.028
Kriging	RF	0.480 ± 0.033	0.787 ± 0.023	0.562 ± 0.036	0.578 ± 0.024
Rounded Kriging[a]	LGBM	0.544 ± 0.035	0.752 ± 0.017	0.492 ± 0.038	0.544 ± 0.020
Optimized Kriging	RF	0.465 ± 0.059	0.777 ± 0.021	0.577 ± 0.064	0.588 ± 0.036

Note: Simulations were conducted with a sample size of 313, repeated 30 times across different random seeds for each missingness level. Rounded kriging rounds at .5 while optimized kriging computationally identifies the best rounding threshold. [a] indicates the best performer within each experiment.

which refers to the replacement of missing data by their estimates (Rehman et al., 2018). In order to simulate missing completely at random, we use our complete sample of 313 villages with Twitter data. [9] We first investigate a number of different imputation methods, showing that kriging is the top method.

6.1. Comparing Imputation Methods

This section presents the findings of several imputation methods for our topic tweet count data. We select three primary imputation methods for comparison. The simplest approach is to fill in all missing values with 0, which assumes that all missingness represents the absence of tweets. Another approach is median imputation for count data. The final approach is to use spatial interpolation using Gaussian-based kriging estimations. Since kriging estimations provide continuous estimations, we test the raw kriging estimates, rounded estimates, and optimized rounding threshold estimates. To simulate missingness, we take our dataset of valid Twitter data which we used in our prior prediction models (n=313). We then randomly select a percentage of cells to be missing going from 10%-50%. We simulate each

[9] In an attempt to impute Twitter for villages without Twitter, we assess whether the remaining 222 of the 535 villages had genuinely zero Tweets. We leveraged other multimodal features to train a classifier aimed at predicting the presence of any of the seven development-related tweet topics. When inferencing on the 222 villages, the model predicted no villages as having any of the 7 development tweet topics.

30 times using 30 different random seeds to ensure that the random selection of missingness does not influence our results.

An efficient approach to impute social media data is to take into account spatial autocorrelation by kriging. Kriging is a widely used surrogate model in spatial statistics, machine learning, and computer simulations (Stein, 1999; Park and Apley, 2018; Santner et al., 2018). By imposing a Gaussian process assumption and defining a covariance function/semivariogram based on spatial locations, kriging provides spatial interpolation and performs prediction at unsampled locations, using the best linear unbiased predictors (BLUP). In addition to BLUPs, the kriging predictions follow a normal distribution with a closed-form expression of variance that quantifies the prediction uncertainty.

In this study, an unknown constant mean function is assumed in kriging and the covariance function is assumed to be the J-Bessel function. The parameters in kriging, including the mean and the correlation parameters defined in the covariance functions, are estimated by maximizing the likelihood function.

Our findings (Table 4) show that the rounded kriging estimates are the best form of imputation regardless of the missing percentage. As expected, we generally see a drop in R^2 as the percentage of missing increases. Compared to the full valid R^2 of 0.610, rounded kriging estimates lose about 0.095 in R^2 in the worst simulation (40% missing). By comparison, fill 0 loses 0.148 R^2 in the worst simulation (50% missing) while median imputation loses 0.162 in the worst simulation (50% missing). Consistent findings are found in studies with 14 topics and 103 topics. The detailed analysis results can be found in Appendix E.

6.2. Using Kriging Interpolation to Enhance Model with Guided Adaptive Sampling

The proposed kriging imputation can be further extended to perform guided adaptive sampling of social media data to enhance the prediction efficiency. The numerical comparisons illustrate the efficiency of kriging imputation by borrowing the spatial information across the existing observations. In addition to imputing the missing data, kriging also provides the imputation uncertainty based on the prediction variance derived from Gaussian process models. This imputation uncertainty effectively guides a sequential sampling procedure, which is known as active learning or Bayesian optimization (Huang et al., 2006; Zhigljavsky and Žilinskas, 2008; Snoek et al., 2012). The best-known Bayesian optimization strategy is to add new observations sequentially based on the expected improvement (Jones et al., 1998). It provides a closed-form expression based on kriging predictor and prediction uncertainty to achieve a balance between extrapolation and interpolation. The illustrated kriging imputation is based on the stationarity assumption, in which the variance remains constant across the region. However, in practice, spatial heterogeneity is common. To address this issue, the proposed framework could be further extended to handle non-stationarity using approaches such as treed Gaussian processes (Gramacy, 2020).

7. Discussion

This study presents the utility of Twitter-derived topic features in filling data gaps and supporting decision making in information-short environments. After validating Twitter as a high resolution data source for development in the first study, we develop an intuitive wealth prediction model based on topic features in the second study and demonstrate a simulated imputation approach in the third study, to further enhance data availability. In doing so, we address the central challenges in developing countries where poverty assessments rely on aggregated statistics on coarse spatial scales (Jung, 2023) and are frequently hampered by missing or outdated census data (Carr-Hill, 2013; Kuffer et al., 2022). In what follows, we discuss how our findings add to the methodological, empirical, and practical use of social media data in development research. In general, our frameworks suggest that a language model can be effectively tailored to measure development needs, particularly in subnational contexts where data is unavailable or sparse. In the remainder of this paper, we first reflect on the contributions of this work in relation to the current literature (§7.1). Building on these contributions, we discuss design and policy (§7.2).

7.1. Contribution to Current Literature

This study contributes to the growing body of research on using alternative data sources for wealth prediction in low-resource settings. We address the three research questions outlined in the introduction by outlining our contributions to the current literature.

Twitter discourse reflects development issues that vary systematically by both geography and socioeconomic context. Poorer villages tend to focus on localized, immediate concerns, while wealthier villages engage in broader, national or global discussions, indicative of higher digital literacy and access to systemic discourse. This stratified communication landscape supports the use of Twitter as a citizen sensing tool, particularly for surfacing grassroots development needs and informing spatial data infrastructure (Barreneche and Lombana-Bermudez, 2023; Elwood, 2008). These patterns align with prior studies in developed countries, where lower-income users concentrate on local issues and higher-income groups discuss more abstract, policy-oriented topics (Giorgi et al., 2023b; Preoţiuc-Pietro et al., 2015b), suggesting similar dynamics are present in developing contexts like Zambia.

Study I (§4) finds that Twitter features sensitively capture variation in development issues across geography, socioeconomic context, and time. We find that wealthier areas tend to engage in abstract, macro-level policy discussions, whereas poorer villages are more likely to focus on localized, immediate concerns. This pattern mirrors findings from prior studies in developed countries, where lower-income users tend to focus on individual wants (sports, entertainment), while higher-income users engage with broader themes (policy, technology, and corporation), as seen in the study (Giorgi et al., 2023a; Preoţiuc-Pietro et al., 2015b). The nuance in our findings, drawn from a developing country context, suggests a more needs-based focus among lower-income groups compared to populations elsewhere. Regional topic distributions also align with known economic activities. For example, mining-related discourse is more prominent in mineral-rich districts. In addition, shifts in Twitter discourse reflect major events, including notable spikes in March 2020 tied to the COVID-19 pandemic and geothermal development projects, highlighting the platform's sensitivity to evolving development-related topics over time. The time dimension appears particularly relevant for topics such as public health, aligning with literature that leverage high-frequency social media data to track discourse during health crises and disasters. These spatio-temporal focal points can provide policymakers with actionable foresight (Kayser and Bierwisch, 2016; Lemoine-Rodríguez et al., 2024; Grassi et al., 2023), allowing governments to detect and respond to salient citizen concerns raised through social media discourse in a timely fashion (Büscher et al., 2014; Shang et al., 2022). Overall, the spatio temporally stratified communication landscape supports the use of Twitter as a tool for sensing grassroots development needs.

In study II(§5), topic features selected through a mixed methods approach predict and explain poverty. While there is a growing body of literature leveraging novel sources of data for poverty estimation, to the best of our knowledge, no existing study has used georeferenced Twitter content alone as a poverty metric. We find that topic features derived from Twitter not only perform competitively on their own but enhance prediction accuracy when combined with aerial, geographic, and network attributes. Our results demonstrate that incorporating curated linguistic features from Twitter enhance predictive power of models that rely on more commonly used development proxies (e.g., satellite imagery) while also offering model explainability. Notably, parsimonious machine learning models using only seven thematically curated topics achieve, compared to machine-driven models using 103 topics, achieve a strong balance between simplicity and predictive performance as discussed in (Das et al., 2023a; Angelov et al., 2014; Castelli et al., 2019; Lustgarten et al., 2017). This making them especially suitable for deployment in resource-constrained settings in line with literature (Hagen et al., 2010; Canaydin et al., 2024).

Our mixed-method approach to select 7 development topics out of 103 topics, make them especially suitable for deployment in resource-constrained settings (Hagen et al., 2010; Canaydin et al., 2024) These models are not only interpretable but also computationally efficient, making them especially suitable for deployment in resource-constrained settings (Hagen et al., 2010; Canaydin et al., 2024). Furthermore, we demonstrate that incorporating curated linguistic features from Twitter can enhance the predictive accuracy of models based on traditionally used as development proxies (e.g., satellite imagery) while offering improved model explainability.

Lastly, findings in study III (§6) indicate that kriging, a spatial interpolation technique leveraging spaital structure, is effective for interpolating irregular Twitter data. Imputation is generally under-discussed, and existing literature uses Twitter data to impute survey responses, rather than being the subject of imputation itself. This study offers a unique contribution by showing that Gaussian-based kriging enhances the spatial completeness of social media data, consistently outperforming simpler imputation methods such as zero-fill and median imputation. Furthermore this spatial interpolation technique provides uncertainty estimates (Huang et al., 2006; Zhigljavsky and Žilinskas, 2008; Snoek et al., 2012) that can inform adaptive sampling, especially in contexts where survey or census data are costly or unavailable. In developing regions with limited spatial data infrastructure (Giorgi et al., 2023a; Livermore et al., 2022; Jiang et al., 2018; Bartelmeß et al., 2024; Abitbol and Morales, 2021; Levy Abitbol et al., 2018; Blasi et al., 2022; Preoţiuc-Pietro et al., 2015a), focusing data collection efforts in areas of high uncertainty can support more efficient and targeted sampling strategies.

7.2. Policy and Design Implications: Wealth Prediction Pipeline Using Social Media Data

We extend our discussion to propose a wealth prediction pipeline that leverages social media data framed within the information infrastructure of *citizen sensors* (Goodchild, 2007) where citizens can "create and share information, forming vast sensing that allow for information in certain topics to be collected, stored, mined, and analyzed in a rapid manner." (Huang et al., 2022, P.2)

To make our findings widely applicable to other data sparse contexts, we suggest the pipeline in Figure 7 for collecting and analyzing social media data to predict wealth. This pipeline includes a number of points with a human-in-the-loop, illustrated in the top left corner in a dotted box. The human-in-the-loop starts with an exploratory and qualitative analysis of the topics (Crowley et al., 2013; Tsaneva et al., 2025), ideally in collaboration with citizens and local leaders and experts. This phase gives the opportunity to find topics most associated with development measures and explore the distribution of these topics over the geographic space. While collected Twitter data provides for sensing over long period of time, human coders "are much better at contextualizing and discriminating data (deciding what is interesting or important), filtering it across multiple modalities (reporting on topics of interest and importance), and capturing the resulting observations for future symbolic processing by machines or collectively with other humans." (Sheth, 2009, P.88)

Next, we can apply machine learning techniques to examine which linguistic topics are representative of ground-truth development data. The ground-truth data most responsible for explaining social media linguistic models can be analyzed using interpretable machine learning techniques, such as partial dependence plots.

Social media data may be incomplete in areas of our interest. To enhance the model efficiency, we would impute the missing data using kriging models which account for the spatial dependency. The imputation performance is examined by simulations of the existing social media data where the ground-truth data are available.

Finally, the predictive models can be trained. With the development of predictive models, this pipeline presents alternative development measures that can be aggregated at various administrative levels (Kuffer et al., 2022; Carr-Hill, 2013).

We envision this as an ongoing process, with continuous iterations of model retraining and exploratory analysis, akin to strategies proposed by Srikanth et al. (2021) for updating linguistic models in fast-evolving online discourse spaces. Such iterative engagement, while applying interpretable machine learning and exploratory analysis, is also important to maintain fair AI models (Awwad et al., 2020) for sustainable development (Lee et al., 2023; Briggs, 2021).

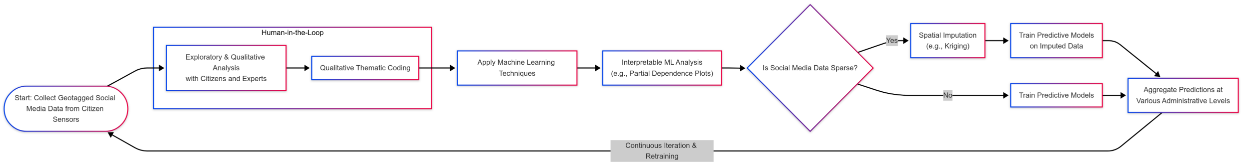

Human-in-the-Loop
Start: Collect Geotagged Social Media Data from Citizen Sensors
Exploratory & Qualitative Analysis with Citizens and Experts
Qualitative Thematic Coding
Apply Machine Learning Techniques
Interpretable ML Analysis (e.g., Partial Dependence Plots)
Is Social Media Data Sparse?
Yes
Spatial Imputation (e.g., Kriging)
Train Predictive Models on Imputed Data
No
Train Predictive Models
Aggregate Predictions at Various Administrative Levels
Continuous Iteration & Retraining

Figure 7: Proposed End-to-End Pipeline for Using Social Media Topic Features in Contextualized Wealth Prediction. This figure illustrates a human-in-the-loop workflow integrating exploratory analysis, interpretable machine learning, spatial imputation, and iterative model retraining to generate granular, citizen-centered development metrics from Twitter data

Collectively, the pipeline described above can offer insights into public discourse within online communities and how it relates to development measures. This aligns with Amartya Sen's capabilities view of development, in which people define their needs as they see fit (Sen, 2005), not as defined by other institutions. Our proposed pipeline embodies a concept of development conceived, measured, and planned by citizens (Ertiö and Bhagwatwar, 2017).

All this in turn would satisfy Sen's expanded view of development to include human capabilities (as citizens can engage in policy development), social and environmental transformation (as people can provide feedback about their local needs), and political empowerment (as citizens can see the effects of their engagement through communication with policymakers, or by seeing changes in public policy (Sachs, 2006; Sen, 2014).

7.3. Limitations

Our model relies on a single social media platform, Twitter using specifically geolocated English tweets. As in other studies that use social media data, particularly Twitter, we recognize that our results primarily reflect the perspectives and conditions of populations and communities represented in geolocated Twitter data. In particular, Twitter users may possess specific characteristics that differ from those of other social media sites. Factors such as age, class, and geography influence the choice of social media platforms (Poushter et al., 2018; Mark Graham and Medhat, 2014).

In Zambia, as in many African countries, Twitter users represent a relatively small and selective segment of the population, with known biases toward younger, urban, male, and English-speaking individuals who have reliable internet and smartphone access (Ragnedda, 2019). In addition, our findings primarily capture the perspectives of active Twitter users, who are disproportionately concentrated in well-to-do provinces, rather than the population as a whole. The demographics of this user base may also differ from those of other social media platforms, such as WhatsApp, Facebook, and TikTok that are increasingly popular in Africa (Geopoll, 2023). We therefore caution against overgeneralizing our findings to the broader Zambian population and emphasize that the availability of georeferenced tweets, as well as the geographic proximity of villages to these tweets, should be taken into account when interpreting results. To note, the lack of demographic representativeness is less problematic in our case, since our approach centers on the content of discourse rather than tweet volume or the assumption that Twitter users form a representative sample, as is common in other studies.

Viewed as a "virtual public square" (Boyd, 2010; Litt and Hargittai, 2016), Twitter enables the observation of emergent public concerns that are spatially tethered and thematically salient. Prior studies have demonstrated that public discourse on Twitter often tracks political events, economic pressures, and public health crises (Ngidi et al., 2016; Leetaru et al., 2013; Suk et al., 2023), making it a valuable, if imperfect, proxy for development-related sentiment. Our findings show systematic variation in topic salience by wealth level, aligning with previous research that suggests lower-income communities focus more on immediate, localized concerns, while higher-income areas discuss abstract policy themes (Preoţiuc-Pietro et al., 2015a; Giorgi et al., 2023a, 2022). We underscore the need for future work that triangulates social media signals with traditional surveys or qualitative fieldwork to ensure inclusion of underrepresented populations (Tufekci, 2014). Nevertheless, we argue that Twitter can complement conventional data sources by providing high-resolution and interpretable indicators of development priorities in data-scarce regions.

The reliance on a single platform such as Twitter is increasingly precarious in the "post-API" era, where access policies, paywalls, and data availability can change abruptly, which affects data-deficient areas and resource-limited stakeholders (Freelon, 2018). Future work could incorporate multiple social media platforms and user-contributed data (such as through structured data donations) to offset platform biases, mitigate access risks, and improve representativeness (Ohme et al., 2024). This requires acknowledging the balance between comprehensiveness and feasibility. In our future work, we aim to integrate cross-platform data and direct user contributions, using Kriging interpolation to guide targeted sampling and strengthen spatial coverage in underrepresented regions.

8. Conclusion

This study demonstrates the potential of Twitter-derived topic features to inform poverty estimation and policy design in data-scarce environments. By integrating qualitative analysis with topic modeling, we constructed a parsimonious and explainable language-based model that effectively predicts village-level wealth in Zambia. Importantly, we show that discourse patterns on social media reflect systematic differences across socioeconomic, geographic, and temporal contexts, with poorer communities focused on immediate survival needs and wealthier communities discussing broader policy issues. Our spatial interpolation approach, using kriging, further addresses missingness in social media datasets and enables more comprehensive spatial coverage. Together, these methodological advances support the development of scalable, interpretable, and adaptive tools for poverty monitoring, particularly where traditional data infrastructure is limited. As digital traces continue to expand, our work underscores the value of socially grounded, linguistically enriched models in enabling participatory development and guiding targeted interventions.

Future research may build upon our findings in two interconnected ways: first, by diversifying the range of social media platforms used, and by extending imputation techniques to further enhance prediction efficiency. The increasing restrictions on social media APIs (Freelon, 2018), as seen on Facebook, Twitter, and Reddit,[10] pose challenges for accessing social media data. Building language models based on the fusion of various social media platforms could, therefore, provide a more comprehensive understanding of citizens' needs in their local contexts. Innovative methodologies would be necessary, like suggestions by Garimella and Tyson (Garimella and Tyson, 2018) for WhatsApp data collection. Data collection processes can be guided by imputation uncertainty. Extending

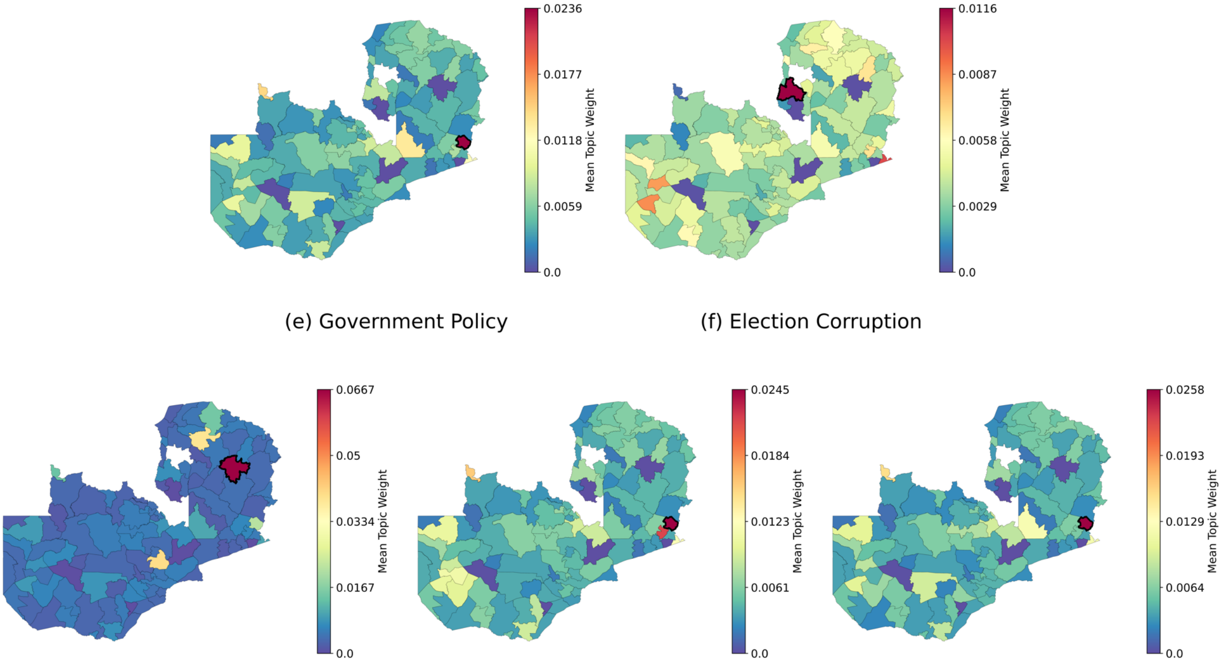
[10] Facebook `https://about.fb.com/news/2018/04/restricting-data-access/`, Twitter `https://techcrunch.com/2023/02/14/twitters-restrictive-api-may-leave-researchers-out-in-the-cold/`, and Reddit `https://www.reddit.com/r/redditdev/comments/14nbw6g/updated_rate_limits_going_into_effect_over_the/`

from the current kriging imputation, future research can further incorporate sequential sampling using the expected improvement method (Jones et al., 1998) to improve the accuracy, robustness, and efficiency of kriging imputation for Twitter data. Furthermore, researchers can extend kriging modeling and imputation to incorporate non-stationarity in Twitter data in which variation differs subnationally by areas within a country (Jiang et al., 2018). These analyses can offer a localized understanding of community needs with high spaito temporal and demographic resolution. This, in turn, enables governments to effectively respond citizen needs. In summary, our findings underscore the potential of social media as a tool for supporting more participatory and inclusive development.

Appendix

A. Distribution of Development Topics Across Zambia

Diverse development-specific topics emerge in other regions (Figure 8). The widespread discussion of *Food Systems* reflects the country's agricultural emphasis and pervasive food security issues. Chipangali District's focus on *Social Inequality* can be attributed to its location in a relatively underdeveloped province. These findings underscore the value of analyzing local Twitter content as it provides intuitive understanding of regional development characteristics on a subnational level.



(e) Government Policy

(f) Election Corruption

(g) Public Health

(h) Social Inequality

(i) Food Systems

Figure 8: District colors denote concentration, e.g., blue is lower concentration, red is higher concentration. (e) shows the topic weight of the *Government Policy* topic at the district level. (f) shows the topic weight of the *Election Corruption* topic at the district level. (g) shows the topic weight of the *Public Health* topic at the district level. (h) shows the topic weight of the *Social Inequality* topic at the district level. (i) shows the topic weight of the *Food Systems* topic at the district level.

The rest of the development-specific Twitter topic distributions are presented in Figure 8. Tweets about frustration with *Government Policy* in Figure 8(e) were most frequent in Chipangali District in the Eastern Province, while tweets about *Election Corruption* in Figure 8(f) appear most frequent in Mansa District in Luapula Province. Shiwang'andu District in Muchinga Province had the most frequent tweets about *Public Health-Related Challenges* in Figure 8(g), while Tweets were particularly about *Social Inequality* in Chipangali District in Figure 8(h). In Chipangali District, *Food Systems* were a frequent topic of discussion, as shown in Figure 8(i).

B. Partial Dependency and SHAP Plots

We present the PDP plots in Figures 9, 10, 11, 12 for the remaining topics from the best 7 development topics.

The only two topics that appeared somewhat relevant were: (1) *NGOs in Zambia* and (2) *cybersecurity*. The first topic centers on the effectiveness of NGOs operating across various development sectors in Zambia and can reasonably be linked to development discourse. In contrast, *cybersecurity* lacks an explicit connection to local or national policy and is less easily explained within traditional development frameworks (see Figure 13).

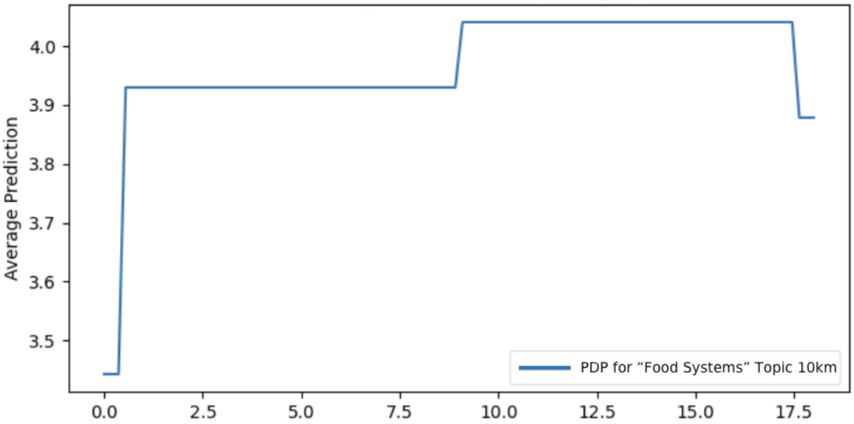


Figure 9: The partial dependency plot for the *Food Systems* tweet count within 10km from the best 7 development topic wealth index (WI) prediction model is presented. The x-axis is the tweet count while the y-axis is the average WI prediction. Here we see two jumps in predicted wealth as tweets increase but a drop in wealth for villages that tweet the most about *Food Systems*.

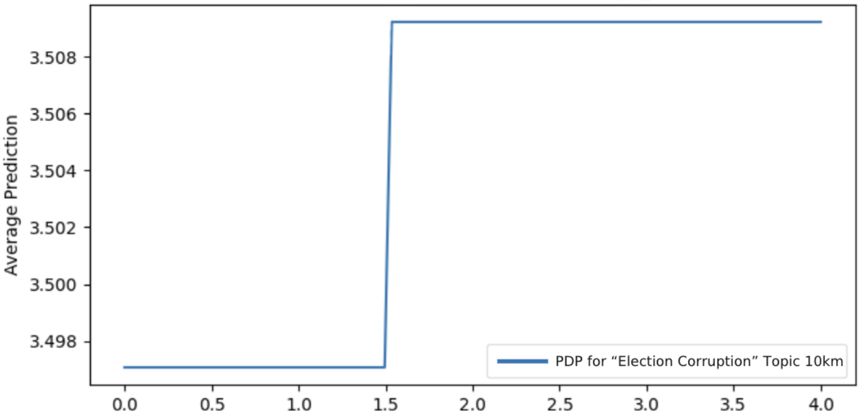


Figure 10: The partial dependency plot for the *Election Corruption* tweet count within 10km from the best 7 development topic wealth index (WI) prediction model is presented. The x-axis is the tweet count while the y-axis is the average WI prediction. Here we see a single jump in predicted wealth as tweet counts increase.

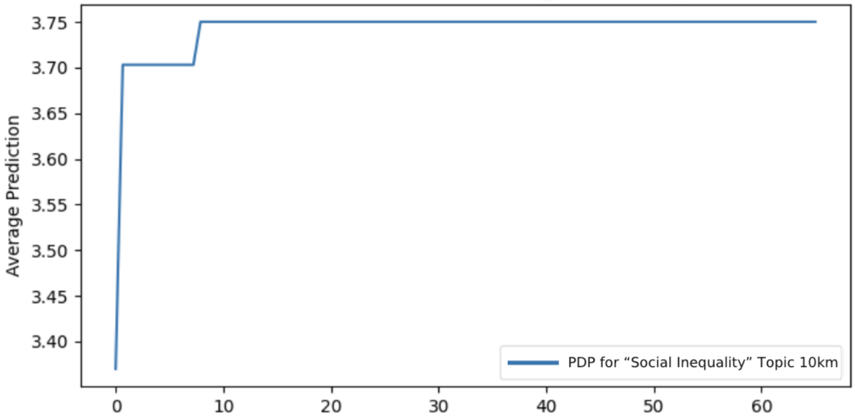


Figure 11: The partial dependency plot for the *Social Inequality* tweet count within 10km from the best 7 development topic wealth index (WI) prediction model is presented. The x-axis is the tweet count while the y-axis is the average WI prediction. Here we see two jumps in predicted wealth as counts increase.

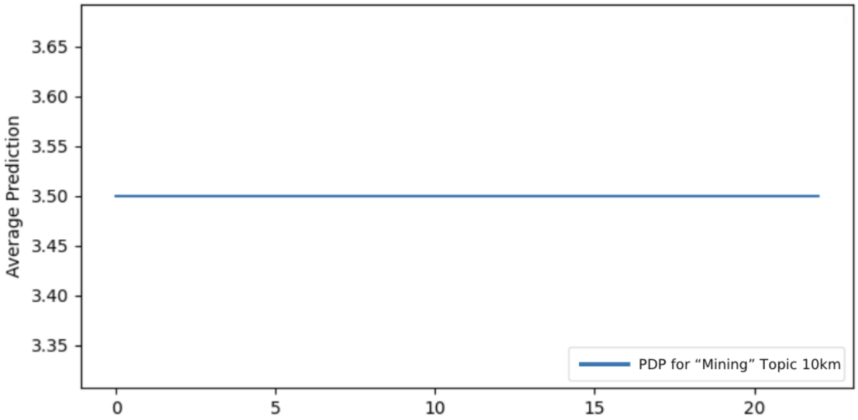


Figure 12: The partial dependency plot for the *Mining* tweet count within 10km from the best 7 development topic wealth index (WI) prediction model is presented. The x-axis is the tweet count while the y-axis is the average WI prediction. Here we see that changes in tweet count lead to no changes in wealth prediction.

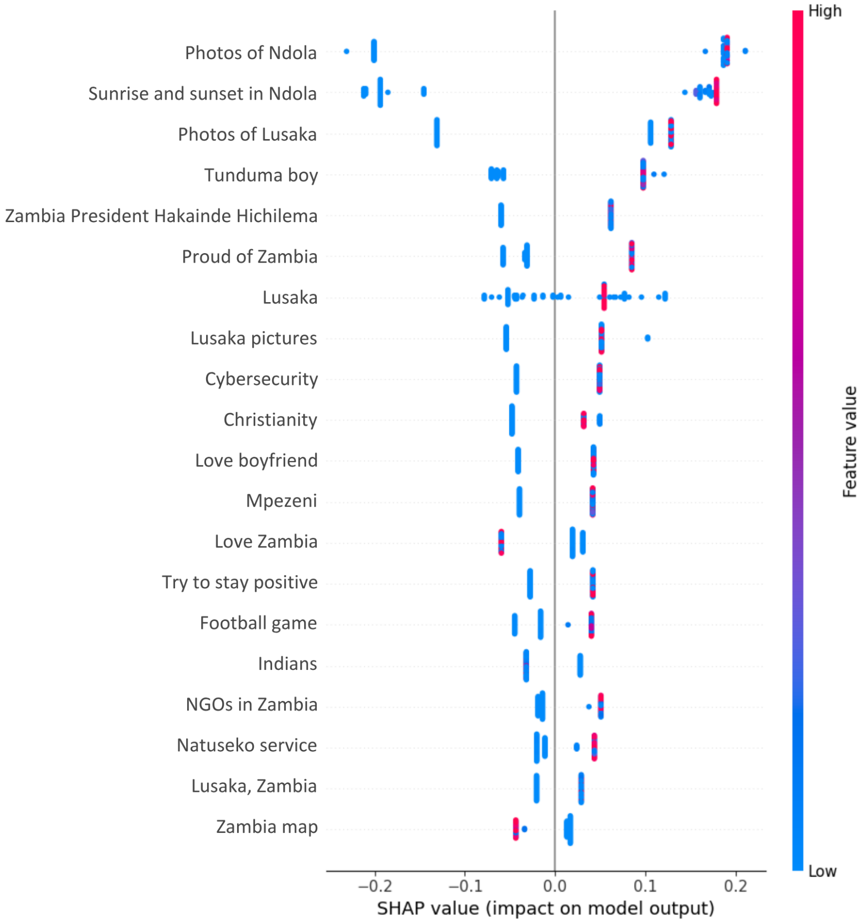


Figure 13: Beeswarm plot showing the top 20 features of the model built using all 103 topics. Most of the top topics in this model are not clearly linked to development or wealth in a conventional sense with two exceptions: (1) NGOs in Zambia; (2) Cybersecurity. The model presented in Figure 4 using only seven curated development topics is more interpretable and nearly as predictive, which supports the value of parsimonious, domain-informed modeling.

C. Other Spatial Data for Wealth Prediction

Spatial data such as satellite imagery, nighttime lights, point of interest data, internet data, vegetation data, and building footprint data have been used for wealth prediction other than social media data as described in section 5.3.1 in the main text. Satellite imagery has been used in five ways. They include i) extracting features from satellite imagery using pre-trained vision models (Ayush et al., 2020; Engstrom et al., 2021; Jung et al., 2025a), ii) obtaining image embeddings from daytime satellite imagery using vision models that had been fine-tuned on NTL data using transfer learning (Jean et al., 2016), iii) directly using NTL as a feature (Reyes et al., 2023; Jung et al., 2025a), iv) calculating Normalized Difference Vegetation Index (NDVI) using Landsat 8 or Sentinel-2 imagery (Reyes et al., 2023; Tang et al., 2018; Hu et al., 2022), and v) leveraging land use and land cover data (Zhao et al., 2019; Jung et al., 2025a; Hu et al., 2022). In addition, phone and connectivity data have been captured from call detail records (CDR) and internet connectivity sources (e.g., Facebook and Ookla) (Blumenstock et al., 2015; Steele et al., 2017; Pokhriyal et al., 2020). Point of Interest (POI) Data from OpenStreetMap, Google Maps, or HERE Maps have been used to measure the distance to POIs (eg., schools, hospitals, etc.) from village clusters or get a POI count within a buffer around the clusters (Muñetón-Santa and Manrique-Ruiz, 2023; Hu et al., 2022; Lee and Braithwaite, 2022; Reyes et al., 2023). Building footprint data have been used to analyze building density, average perimeter or area, and structural features such as shape index, compactness score, and angle entropy within a buffer around village clusters (Jung et al., 2025a; Engstrom et al., 2021).

D. Twitter with Other Spatial Data

Below, we present the R^2 boosting effects when adding Twitter data to other spatial data used for wealth prediction (see Appendix C for details on how other spatial data have been used in previous works). We present results for base models in Table 3 along with the results of base models combined with Twitter data in Table 5. The description of the features used in the models are as follows: Twitter top topic count represents all the 103 topics. Internet data features include total speedtest count and average internet metrics like upload speed, download speed, and latency for both mobile and fixed internet connections with distance to the closest mobile and fixed internet connection. Distance to POIs are distances from the cluster centroid to the nearest POI such as schools, hospitals, etc. Satellite imagery are embeddings retrieved from a pre-trained Vision Transformer (ViT) on ImageNet then fine-tuned using nighttime light data. The nighttime light feature uses the average radiance values within a 10km buffer around the village cluster. Building density is the count of buildings around the village cluster within the 10km buffer. Building perimeter is the average perimeter of the buildings around the village cluster within a 10km buffer. Building area is the average area of the buildings around the village cluster within a 10km buffer. Structural features are derived from OpenStreetMap building footprint data, in which we calculate the shape index (measures the uniformity of building shape), compactness score (measures the compactness of buildings in a neighborhood), and orientation (measures the orientation of buildings) of the buildings within a 10km buffer around the village cluster. All these structural features give insights into the uniformity of the neighborhood. Normalized Difference Vegetation Index (NDVI) is the measurement of vegetation which ranges from -1 to 1, where values closer to -1 mean no vegetation (e.g., water bodies), whereas values close to 0 mean bare soil and values close to 1 mean dense vegetation.

Additionally, Figure 14 shows Twitter topic features contribution in the multimodal model with all the other spatial data mentioned above. Twitter features like *mining* and *food systems* are some of the top features in the multimodal model. While *mining* was the least predictive feature in our seven-development-topic model (see §5.2), the dependence between different features might provide models that are both more predictive and more explainable.

Table 5
Predictive Performance of Base Models Enhanced with Twitter Topic Features

Base Model + Twitter (All 103 Topics)	R^2	mean val MSE	test MSE	test MAE	Twitter R^2 Boost
Internet Data	0.745 ± 0.015	0.427 ± 0.003	0.276 ± 0.016	0.416 ± 0.012	+0.021 ± 0.023
Dist. to POIs	0.730 ± 0.012	0.416 ± 0.005	0.292 ± 0.013	0.438 ± 0.013	+0.003 ± 0.016
Satellite Imagery	0.697 ± 0.015	0.379 ± 0.008	0.327 ± 0.017	0.456 ± 0.017	-0.008 ± 0.015
Building Density	0.679 ± 0.023	0.618 ± 0.006	0.346 ± 0.025	0.454 ± 0.017	+0.159 ± 0.061
Nighttime Lights	0.656 ± 0.015	0.570 ± 0.004	0.371 ± 0.016	0.462 ± 0.013	+0.054 ± 0.019
Building Perimeter	0.649 ± 0.012	0.596 ± 0.007	0.378 ± 0.013	0.476 ± 0.011	+0.130 ± 0.034
Building Area	0.649 ± 0.016	0.597 ± 0.007	0.379 ± 0.017	0.480 ± 0.014	+0.144 ± 0.017
Structural Features	0.638 ± 0.007	0.667 ± 0.008	0.390 ± 0.008	0.483 ± 0.006	+0.250 ± 0.079
NDVI	0.621 ± 0.020	0.666 ± 0.005	0.409 ± 0.022	0.482 ± 0.007	+0.435 ± 0.036

Note: Sample size = 313. We evaluate R², mean validation MSE, test MSE, and test MAE across 30 iterations to assess the impact of adding Twitter topic features to baseline models.

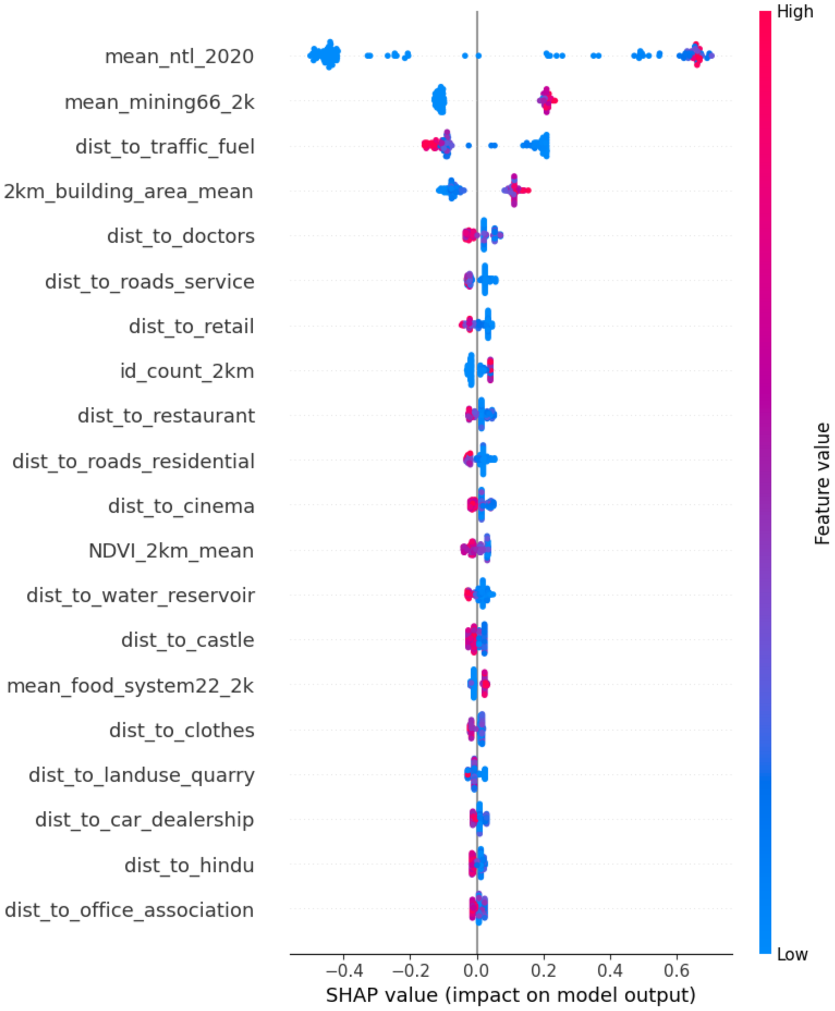


Figure 14: The beeswarm plot shows the top 20 features' contributions from a high-performing multimodal model ($R^2 = 0.785$) to predict wealth across 535 village clusters. Here we see the topic weights for *Mining* and *Food Systems* are important. In addition, the counts of any tweet (idcount 2km) are important.

Table 6
Imputation Simulation Results for Wealth Prediction Using 14 Uncorrelated Twitter Topics

Imputation Method	Best Algorithm	R^2	mean val MSE	test MSE	test MAE
N = 313, Simulated MAR 10%					
Fill 0	XGB	0.566 ± 0.024	0.694 ± 0.013	0.468 ± 0.026	0.521 ± 0.015
Median	XGB	0.568 ± 0.022	0.698 ± 0.013	0.466 ± 0.024	0.523 ± 0.015
Kriging	XGB	0.574 ± 0.017	0.685 ± 0.011	0.460 ± 0.018	0.519 ± 0.014
Rounded Kriging[a]	LGBM	0.591 ± 0.019	0.679 ± 0.012	0.442 ± 0.020	0.506 ± 0.016
Optimized Kriging	MLP	0.479 ± 0.059	0.823 ± 0.033	0.562 ± 0.063	0.595 ± 0.044
N = 313, Simulated MAR 20%					
Fill 0	LGBM	0.554 ± 0.042	0.700 ± 0.016	0.481 ± 0.045	0.528 ± 0.027
Median	XGB	0.563 ± 0.026	0.705 ± 0.018	0.471 ± 0.028	0.531 ± 0.022
Kriging	LGBM	0.572 ± 0.033	0.682 ± 0.019	0.462 ± 0.036	0.517 ± 0.023
Rounded Kriging[a]	XGB	0.590 ± 0.017	0.693 ± 0.019	0.442 ± 0.018	0.511 ± 0.013
Optimized Kriging	LGBM	0.579 ± 0.043	0.683 ± 0.018	0.454 ± 0.046	0.516 ± 0.022
N = 313, Simulated MAR 30%					
Fill 0	RF	0.517 ± 0.059	0.718 ± 0.018	0.522 ± 0.064	0.547 ± 0.032
Median	RF	0.536 ± 0.054	0.726 ± 0.020	0.501 ± 0.058	0.540 ± 0.031
Kriging	LGBM	0.578 ± 0.028	0.690 ± 0.019	0.455 ± 0.030	0.512 ± 0.019
Rounded Kriging[a]	XGB	0.585 ± 0.018	0.699 ± 0.013	0.447 ± 0.019	0.514 ± 0.015
Optimized Kriging	LGBM	0.577 ± 0.032	0.685 ± 0.015	0.457 ± 0.035	0.514 ± 0.020
N = 313, Simulated MAR 40%					
Fill 0	XGB	0.516 ± 0.059	0.714 ± 0.022	0.522 ± 0.064	0.556 ± 0.038
Median	LGBM	0.512 ± 0.062	0.718 ± 0.031	0.527 ± 0.067	0.558 ± 0.041
Kriging	LGBM	0.568 ± 0.040	0.692 ± 0.022	0.466 ± 0.043	0.518 ± 0.024
Rounded Kriging[a]	LGBM	0.581 ± 0.023	0.696 ± 0.018	0.452 ± 0.024	0.517 ± 0.015
Optimized Kriging	LGBM	0.576 ± 0.028	0.692 ± 0.019	0.457 ± 0.030	0.519 ± 0.019
N = 313, Simulated MAR 50%					
Fill 0	XGB	0.513 ± 0.040	0.751 ± 0.026	0.526 ± 0.043	0.564 ± 0.023
Median	LGBM	0.513 ± 0.065	0.752 ± 0.037	0.525 ± 0.070	0.562 ± 0.043
Kriging	XGB	0.554 ± 0.023	0.718 ± 0.017	0.482 ± 0.025	0.534 ± 0.016
Rounded Kriging[a]	LGBM	0.592 ± 0.033	0.707 ± 0.016	0.440 ± 0.035	0.507 ± 0.024
Optimized Kriging	LGBM	0.576 ± 0.037	0.704 ± 0.020	0.458 ± 0.040	0.523 ± 0.024

Note: Simulations were conducted with a sample size of 313, repeated 30 times across different random seeds. This is a comparison of different imputation methods (zero-fill, median, kriging, rounded kriging, optimized kriging) under varying levels of simulated missingness, evaluated by model performance metrics. Rounded Kriging rounds at .5 while Optimized Kriging computationally identifies the best rounding threshold. [a] indicates the best performer within each experiment.

E. Kriging Appendix

Below, we present additional tables for kriging, as shown in Tables 6 and 7.

Table 7
Imputation Simulation Results for Wealth Prediction Using All 103 Twitter Topics

Imputation Method	Best Algorithm	R^2	mean val MSE	test MSE	test MAE
N = 313, Simulated MAR 10%					
Fill 0	LGBM	0.614 ± 0.030	0.646 ± 0.014	0.417 ± 0.032	0.501 ± 0.022
Median	LGBM	0.622 ± 0.034	0.648 ± 0.020	0.407 ± 0.037	0.493 ± 0.022
Kriging	RF	0.619 ± 0.023	0.667 ± 0.017	0.411 ± 0.024	0.488 ± 0.015
Rounded Kriging[a]	LGBM	0.633 ± 0.025	0.652 ± 0.012	0.396 ± 0.027	0.484 ± 0.016
Optimized Kriging	RF	0.622 ± 0.014	0.673 ± 0.012	0.407 ± 0.015	0.486 ± 0.010
N = 313, Simulated MAR 20%					
Fill 0	XGB	0.613 ± 0.025	0.640 ± 0.013	0.418 ± 0.027	0.504 ± 0.018
Median	XGB	0.609 ± 0.020	0.643 ± 0.017	0.422 ± 0.021	0.507 ± 0.014
Kriging	XGB	0.608 ± 0.018	0.649 ± 0.013	0.423 ± 0.019	0.507 ± 0.013
Rounded Kriging[a]	XGB	0.622 ± 0.016	0.651 ± 0.010	0.408 ± 0.017	0.496 ± 0.014
Optimized Kriging	LGBM	0.621 ± 0.028	0.656 ± 0.013	0.409 ± 0.031	0.495 ± 0.019
N = 313, Simulated MAR 30%					
Fill 0	XGB	0.620 ± 0.027	0.642 ± 0.014	0.410 ± 0.029	0.503 ± 0.021
Median	LGBM	0.601 ± 0.044	0.661 ± 0.024	0.431 ± 0.047	0.505 ± 0.030
Kriging	LGBM	0.597 ± 0.029	0.669 ± 0.015	0.435 ± 0.031	0.510 ± 0.021
Rounded Kriging[a]	LGBM	0.628 ± 0.025	0.665 ± 0.012	0.402 ± 0.027	0.488 ± 0.019
Optimized Kriging	XGB	0.610 ± 0.023	0.661 ± 0.011	0.421 ± 0.024	0.506 ± 0.018
N = 313, Simulated MAR 40%					
Fill 0[a]	RF	0.606 ± 0.028	0.667 ± 0.019	0.426 ± 0.030	0.511 ± 0.020
Median	RF	0.596 ± 0.039	0.686 ± 0.024	0.436 ± 0.042	0.515 ± 0.023
Kriging	RF	0.587 ± 0.029	0.718 ± 0.017	0.446 ± 0.032	0.508 ± 0.018
Rounded Kriging	XGB	0.601 ± 0.016	0.678 ± 0.009	0.430 ± 0.017	0.511 ± 0.013
Optimized Kriging	RF	0.597 ± 0.027	0.704 ± 0.015	0.435 ± 0.029	0.509 ± 0.016
N = 313, Simulated MAR 50%					
Fill 0	XGB	0.591 ± 0.035	0.655 ± 0.023	0.442 ± 0.037	0.523 ± 0.023
Median[a]	XGB	0.602 ± 0.031	0.668 ± 0.024	0.430 ± 0.033	0.516 ± 0.021
Kriging	LGBM	0.578 ± 0.024	0.692 ± 0.020	0.456 ± 0.026	0.525 ± 0.016
Rounded Kriging	RF	0.597 ± 0.018	0.726 ± 0.012	0.435 ± 0.019	0.504 ± 0.014
Optimized Kriging	XGB	0.584 ± 0.021	0.688 ± 0.012	0.449 ± 0.022	0.523 ± 0.013

Note: Simulations were conducted with a sample size of 313, repeated 30 times across different random seeds. The values represent the performance assessment of imputation methods under increasing missingness, using the full topic set to predict village-level wealth. Rounded kriging rounds at .5 while optimized kriging computationally identifies the best rounding threshold. [a] indicates the best performer within each experiment

References

Abitbol, J.L., Morales, A.J., 2021. Socioeconomic patterns of Twitter user activity. Entropy 23, 780. doi:10.3390/e23060780.

Acemoglu, D., García-Jimeno, C., Robinson, J.A., 2015. State capacity and economic development: A network approach. American economic review 105, 2364–2409. doi:10.1257/aer.20140044.

Alkire, S., Santos, M.E., 2010. Multidimensional poverty index. OPHI Research Brief , 1–8.

Angelov, D., 2020. Top2Vec: Distributed representations of topics. doi:10.48550/arXiv.2008.09470.

Angelov, P., Gu, X., Principe, J.C., 2014. pclass: An effective classifier for streaming examples, in: 2014 International Joint Conference on Neural Networks (IJCNN), IEEE. pp. 760–767. URL: https://ieeexplore.ieee.org/document/6776566.

Asyaky, M.S., Mandala, R., 2021. Improving the performance of hdbscan on short text clustering by using word embedding and umap, in: 2021 8th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA), pp. 1–6. doi:10.1109/ICAICTA53211.2021.9640285.

Awwad, Y., Fletcher, R., Frey, D., Gandhi, A., Najafian, M., Teodorescu, M., 2020. Exploring fairness in machine learning for international development. Technical Report. CITE MIT D-Lab.

Ayush, K., Uzkent, B., Burke, M., Lobell, D., Ermon, S., 2020. Generating interpretable poverty maps using object detection in satellite images, in: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence Organization, California. p. 4410–4416. doi:10.24963/ijcai.2020/608.

Barreneche, C., Lombana-Bermudez, A., 2023. Another infrastructure is possible: Grassroots citizen sensing and environmental data justice in colombia. International Journal of Communication (19328036) 17.

Bartelmeß, T., Schönfeld, M., Pfeffer, J., 2024. Exploring food poverty experiences in the German twitter-sphere. BMC Public Health 24, 1398. doi:10.1186/s12889-024-18926-8.

Besley, T., Persson, T., 2010. State capacity, conflict, and development. Econometrica 78, 1–34.

Best, M.L., Meng, A., 2015. Twitter democracy: Policy versus identity politics in three emerging African democracies, in: Proceedings of the Seventh International Conference on Information and Communication Technologies and Development, Association for Computing Machinery, New York, NY, USA. doi:10.1145/2737856.2738017.

Bird, K., 2019. Addressing spatial poverty traps. Chronic Poverty Advisory Network, London , 2doi:10.4324/9780203006214-33.

Blasi, S., Gobbo, E., Sedita, S., 2022. Smart cities and citizen engagement: Evidence from Twitter data analysis on Italian municipalities. Journal of Urban Management 11, 153–165. doi:10.1016/j.jum.2022.04.001.

Blei, D.M., Lafferty, J.D., 2006. Dynamic topic models, in: Proceedings of the 23rd International Conference on Machine Learning, ACM, New York, NY, USA. pp. 113–120. doi:10.1145/1143844.1143859.

Blumenstock, J., Cadamuro, G., On, R., 2015. Predicting poverty and wealth from mobile phone metadata. Science 350, 1073–1076. doi:10.1126/science.aac4420.

Boon-Itt, S., Skunkan, Y., et al., 2020. Public perception of the COVID-19 pandemic on Twitter: Sentiment analysis and topic modeling study. JMIR public health and surveillance 6, e21978. doi:10.2196/21978.

Boyd, D., 2010. Social network sites as networked publics: Affordances, dynamics, and implications, in: A networked self. Routledge, pp. 47–66. doi:10.4324/9780203876527-8.

Bozarth, L., Pal, J., 2019. Twitter discourse as a lens into politicians' interest in technology and development, in: Proceedings of the Tenth International Conference on Information and Communication Technologies and Development, Association for Computing Machinery, New York, NY, USA. doi:10.1145/3287098.3287129.

Breiman, L., Friedman, J., Olshen, R.A., Stone, C.J., 2017. Classification and regression trees. Routledge. doi:10.1201/9781315139470.

Briggs, M., 2021. Satdash: An interactive dashboard for assessing land damage in Nigeria and Mali, in: Proceedings of the 4th ACM SIGCAS Conference on Computing and Sustainable Societies, Association for Computing Machinery, New York, NY, USA. p. 100–114. doi:10.1145/3460112.3471949.

Bruckman, A., 2002. Studying the amateur artist: A perspective on disguising data collected in human subjects research on the Internet. Ethics and Information Technology 4, 217–231.

Bruns, A., Burgess, J., 2011. The use of Twitter hashtags in the formation of ad hoc publics, in: Proceedings of the 6th European consortium for political research (ECPR) general conference 2011, The European Consortium for Political Research (ECPR). pp. 1–9.

Büscher, M., Liegl, M., Thomas, V., 2014. Collective intelligence in crises, in: Social collective intelligence: Combining the powers of humans and machines to build a smarter society. Springer, pp. 243–265. doi:10.1007/978-3-319-08681-1_12.

Campello, R.J.G.B., Moulavi, D., Sander, J., 2013. Density-based clustering based on hierarchical density estimates, in: Pei, J., Tseng, V.S., Cao, L., Motoda, H., Xu, G. (Eds.), Advances in Knowledge Discovery and Data Mining, Springer. pp. 160–172. doi:10.1007/978-3-642-37456-2_14.

Canaydin, A., Fu, C., Balint, A., Khalil, M., Miller, C., Kazmi, H., 2024. Interpretable domain-informed and domain-agnostic features for supervised and unsupervised learning on building energy demand data. Applied Energy 360, 122741. doi:10.1016/j.apenergy.2024.122741.

Carr-Hill, R., 2013. Missing millions and measuring development progress. World Development 46, 30–44. doi:10.1016/j.worlddev.2012.12.017.

Castelli, M., Popovič, A., Vanneschi, L., Semenkin, E., 2019. Feature extraction and selection for parsimonious classifiers with multiobjective genetic programming. IEEE Transactions on Cybernetics 49, 2191–2203. URL: https://ieeexplore.ieee.org/document/8758127.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J.L., Blei, D.M., 2009. Reading tea leaves: How humans interpret topic models, in: Advances in neural information processing systems, pp. 288–296.

Chen, W., Su, Y., Yan, X., Wang, W.Y., 2020. KGPT: Knowledge-grounded pre-training for data-to-text generation, in: Webber, B., Cohn, T., He, Y., Liu, Y. (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online. pp. 8635–8648. doi:10.18653/v1/2020.emnlp-main.697.

Cingolani, L., 2018. The role of state capacity in development studies. Journal of Development Perspectives 2, 88–114. doi:10.5325/jdevepers.2.1-2.0088.

Crowley, D.N., Curry, E., Breslin, J.G., 2013. Closing the loop — from citizen sensing to citizen actuation, in: 2013 7th IEEE International Conference on Digital Ecosystems and Technologies (DEST), pp. 108–113. doi:10.1109/DEST.2013.6611338.

Das, A., Liu, H., Kovatchev, V., Lease, M., 2023a. The state of human-centered NLP technology for fact-checking. Information Processing Management 60, 103219. doi:10.1016/j.ipm.2022.103219.

Das, D., Islam, A.N., Haque, S.M.T., Vuorinen, J., Ahmed, S.I., 2023b. Understanding the strategies and practices of Facebook microcelebrities for engaging in sociopolitical discourses, in: Proceedings of the 2022 International Conference on Information and Communication Technologies and Development, Association for Computing Machinery, New York, NY, USA. doi:10.1145/3572334.3572368.

Dincecco, M., 2017. State capacity and economic development: Present and past. Cambridge University Press. doi:10.1017/9781108539913.

Egger, R., Yu, J., 2022. A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts. Frontiers in Sociology 7, 886498. doi:10.3389/fsoc.2022.886498.

Elmustafa, A., Rozi, E., He, Y., Mai, G., Ermon, S., Burke, M., Lobell, D., 2022. Understanding economic development in rural Africa using satellite imagery, building footprints and deep models, in: Proceedings of the 30th International Conference on Advances in Geographic Information Systems, Association for Computing Machinery, New York, NY, USA. URL: https://doi.org/10.1145/3557915.3561025, doi:10.1145/3557915.3561025.

Elwood, S., 2008. Grassroots groups as stakeholders in spatial data infrastructures: challenges and opportunities for local data development and sharing. International Journal of Geographical Information Science 22, 71–90. doi:10.1080/13658810701348971.

Engstrom, R., Hersh, J., Newhouse, D., 2021. Poverty from space: Using high resolution satellite imagery for estimating economic well-being. Published by Oxford University Press on behalf of the World Bank. doi:10.1596/40907.

Ertiö, T.P., Bhagwatwar, A., 2017. Citizens as planners: Harnessing information and values from the bottom-up. International Journal of Information Management 37, 111–113. doi:10.1016/j.ijinfomgt.2017.01.001.

Fatehkia, M., Coles, B., Ofli, F., Weber, I., 2020. The relative value of Facebook advertising data for poverty mapping, in: Proceedings of the International AAAI Conference on Web and Social Media, pp. 934–938. doi:10.1609/icwsm.v14i1.7361.

Freelon, D., 2018. Computational research in the post-API age. Political Communication 35, 665–668. doi:10.1080/10584609.2018.1477506.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Annals of Statistics , 1189–1232doi:10.1214/aos/1013203451.

de la Fuente, A., Murr, A., Rascón, E., 2015. Mapping Subnational Poverty in Zambia. World Bank. doi:10.1596/21783.

Gao, J., Zhang, Y.C., Zhou, T., 2019. Computational socioeconomics. Physics Reports 817, 1–104. doi:10.1016/j.physrep.2019.05.002.

Garimella, K., Tyson, G., 2018. Whatapp doc? A first look at WhatsApp public group data, in: Proceedings of the international AAAI conference on web and social media. doi:10.1609/icwsm.v12i1.14989.

Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E.L., Fei-Fei, L., 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, 13108–13113. doi:10.1073/pnas.1700035114.

Geopoll, 2023. Social media usage trends in Africa: Geopoll report. Social Media Usage Survey. Geopoll. Denver, Colorado.

Giorgi, S., Eichstaedt, J.C., Preoţiuc-Pietro, D., Gardner, J.R., Schwartz, H.A., Ungar, L.H., 2023a. Filling in the white space: Spatial interpolation with gaussian processes and social media data. Current research in ecological and social psychology 5, 100159. doi:10.1016/j.cresp.2023.100159.

Giorgi, S., Eichstaedt, J.C., Preoţiuc-Pietro, D., Gardner, J.R., Schwartz, H.A., Ungar, L.H., 2023b. Filling in the white space: Spatial interpolation with gaussian processes and social media data. Current Research in Ecological and Social Psychology 5, 100159. URL: https://www.sciencedirect.com/science/article/pii/S2666622723000722, doi:https://doi.org/10.1016/j.cresp.2023.100159.

Giorgi, S., Lynn, V.E., Gupta, K., Ahmed, F., Matz, S., Ungar, L.H., Schwartz, H.A., 2022. Correcting sociodemographic selection biases for population prediction from social media, in: Proceedings of the International AAAI Conference on Web and Social Media, pp. 228–240. doi:10.1609/icwsm.v16i1.19287.

Goodchild, M.F., 2007. Citizens as sensors: the world of volunteered geography. GeoJournal 69, 211–221. doi:10.1007/s10708-007-9111-y.

Goovaerts, P., 2005. Geostatistical analysis of disease data: Estimation of cancer mortality risk from empirical frequencies using poisson kriging. International Journal of Health Geographics 4, 1–33. doi:10.1186/1476-072x-4-31.

Gramacy, R., 2020. Surrogates: Gaussian process modeling, design and optimization for the applied sciences. Chapman Hall/CRC, Boca Raton, FL.

Grassi, L., Ciranni, M., Baglietto, P., Recchiuto, C.T., Maresca, M., Sgorbissa, A., 2023. Emergency management through information crowdsourcing. Information Processing Management 60, 103386. URL: https://www.sciencedirect.com/science/article/pii/S0306457323001231, doi:https://doi.org/10.1016/j.ipm.2023.103386.

Griffiths, T.L., Steyvers, M., Tenenbaum, J.B., 2007. Topics in semantic representation 114, 211–244. doi:10.1037/0033-295X.114.2.211. place: US Publisher: American Psychological Association.

Grootendorst, M., 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794 .

Hagen, E., Shroder, J., Lu, X., Teufert, J.F., 2010. Reverse engineered flood hazard mapping in Afghanistan: A parsimonious flood map model for developing countries. Quaternary International 226, 82–91. doi:10.1016/j.quaint.2009.11.021.

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

Harriet, S., Diattara, A., Traore, A., Hu, R., Zhang, D., Rundensteiner, E., Ba, C., 2024. Extracting Semantic Topics about Development in Africa from Social Media. URL: https://doi.org/10.36227/techrxiv.171837929.95002363/v1.

Hastie, T., Tibshirani, R., Friedman, J., et al., 2009. The elements of statistical learning: Data mining, inference, and prediction. doi:10.1111/j.1751-5823.2009.00095_18.x.

Hays, S., 1998. The Cultural Contradictions of Motherhood. Yale University Press.

Hoover, J., Dehghani, M., 2020. The big, the bad, and the ugly: Geographic estimation with flawed psychological data. Psychological Methods 25, 412.

Hu, S., Ge, Y., Liu, M., Ren, Z., Zhang, X., 2022. Village-level poverty identification using machine learning, high-resolution images, and geospatial data. International Journal of Applied Earth Observation and Geoinformation 107, 102694. doi:10.1016/j.jag.2022.102694.

Huang, A.A., Huang, S.Y., 2023. Increasing transparency in machine learning through bootstrap simulation and shapely additive explanations. PLoS One 18, e0281922. doi:10.1371/journal.pone.0281922.

Huang, D., Allen, T.T., Notz, W.I., Zeng, N., 2006. Global optimization of stochastic black-box systems via sequential kriging meta-models. Journal of Global Optimization 34, 441–466. doi:10.1007/s10898-005-2454-3.

Huang, X., Wang, S., Zhang, M., Hu, T., Hohl, A., She, B., Gong, X., Li, J., Liu, X., Gruebner, O., et al., 2022. Social media mining under the COVID-19 context: Progress, challenges, and opportunities. International Journal of Applied Earth Observation and Geoinformation 113, 102967. doi:10.1016/j.jag.2022.102967.

Indaco, A., 2020. From Twitter to GDP: Estimating economic activity from social media. Regional Science and Urban Economics 85, 103591. doi:10.1016/j.regsciurbeco.2020.103591.

Jaidka, K., Ahmed, S., 2015. The 2014 Indian general election on Twitter: An analysis of changing political traditions, in: Proceedings of the Seventh International Conference on Information and Communication Technologies and Development, Association for Computing Machinery, New York, NY, USA. doi:10.1145/2737856.2737889.

Jamali, M., Nejat, A., Ghosh, S., Jin, F., Cao, G., 2019. Social media data and post-disaster recovery. International Journal of Information Management 44, 25–37.

Jean, N., Burke, M., Xie, M., Davis, W.M., Lobell, D.B., Ermon, S., 2016. Combining satellite imagery and machine learning to predict poverty. Science 353, 790–794. doi:10.1126/science.aaf7894.

Jiang, Y., Li, Z., Ye, X., 2018. Understanding demographic and socioeconomic biases of geotagged Twitter users at the county level. Cartography and Geographic Information Science 46, 228–242. doi:10.1080/15230406.2018.1434834.

Jones, D.R., Schonlau, M., Welch, W.J., 1998. Efficient global optimization of expensive black-box functions. Journal of Global Optimization 13, 455–492.

Jongman, B., Wagemaker, J., Revilla Romero, B., Coughlan de Perez, E., 2015. Early flood detection for rapid humanitarian response: harnessing near real-time satellite and Twitter signals. ISPRS International Journal of Geo-Information 4, 2246–2266. doi:10.3390/ijgi4042246.

Jung, W., 2023. Mapping community development aid: Spatial analysis in Myanmar. World Development 164, 106124. doi:10.1016/j.worlddev.2022.106124.

Jung, W., Ghadimi, S., Ntarlagiannis, D., Kim, A.H., 2025a. Using artificial intelligence/machine learning to evaluate the distribution of community development aid across myanmar. Socio-Economic Planning Sciences 98, 102139. doi:10.1016/j.seps.2024.102139.

Jung, W., Sinha, A., Kim, A., Shah, V., Lu, Y., Lee, L., Ammari, T., 2025b. The last mile in remote sensing poverty prediction. ACM J. Comput. Sustain. Soc. 3. doi:10.1145/3724422.

Kayser, V., Bierwisch, A., 2016. Using Twitter for foresight: An opportunity? Futures 84, 50–63. URL: https://www.sciencedirect.com/science/article/pii/S0016328716302749, doi:https://doi.org/10.1016/j.futures.2016.09.006.

Killick, R., Fearnhead, P., Eckley, I.A., 2012. Optimal detection of changepoints with a linear computational cost. Journal of the American Statistical Association 107, 1590–1598. doi:10.1080/01621459.2012.737745.

Kondmann, L., Haeberle, M., Zhu, X.X., 2020. Combining Twitter and earth observation data for local poverty mapping, in: NeuRIPS Machine Learning for the Developing World Workshop, pp. 1–5.

Kraak, M.J., Ormeling, F., 2020. Cartography: Visualization of geospatial data. CRC Press. doi:10.1080/23729333.2021.1882050.

Kuffer, M., Owusu, M., Oliveira, L., Sliuzas, R., van Rijn, F., 2022. The missing millions in maps: Exploring causes of uncertainties in global gridded population datasets. ISPRS International Journal of Geo-Information 11, 403. doi:10.3390/ijgi11070403.

Lampos, V., Zou, B., Cox, I.J., 2017. Enhancing feature selection using word embeddings: The case of flu surveillance, in: Proceedings of the 26th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE. p. 695–704. doi:10.1145/3038912.3052622.

Lasri, K., Tonneau, M., Naushan, H., Malhotra, N., Farouq, I., Orozco-Olvera, V., Fraiberger, S., 2023. Large-scale demographic inference of social media users in a low-resource scenario, in: Proceedings of the International AAAI Conference on Web and Social Media, pp. 519–529. doi:10.1609/icwsm.v17i1.22165.

Lee, J., Song, H., Lee, D., Kim, S., Sim, J., Cha, M., Park, K.R., 2023. Machine learning driven aid classification for sustainable development, in: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI), pp. 6040–6048. doi:10.24963/ijcai.2023/670.

Lee, K., Braithwaite, J., 2022. High-resolution poverty maps in Sub-Saharan Africa. World Development 159, 106028. doi:10.1016/j.worlddev.2022.106028.

Leetaru, K., Wang, S., Cao, G., Padmanabhan, A., Shook, E., 2013. Mapping the global Twitter heartbeat: The geography of Twitter. First Monday doi:10.5210/fm.v18i5.4366.

Leidig, M., Teeuw, R.M., 2015. Quantifying and mapping global data poverty. PloS one 10, e0142076. doi:10.1371/journal.pone.0145591.

Lemoine-Rodríguez, R., Mast, J., Mühlbauer, M., Mandery, N., Biewer, C., Taubenböck, H., 2024. The voices of the displaced: Mobility and Twitter conversations of migrants of Ukraine in 2022. Information Processing Management 61, 103670. doi:10.1016/j.ipm.2024.103670.

Levy Abitbol, J., Karsai, M., Fleury, E., 2018. Location, occupation, and semantics based socioeconomic status inference on Twitter, in: 2018 IEEE International Conference on Data Mining Workshops (ICDMW), IEEE. p. 1192–1199. doi:10.1109/icdmw.2018.00171.

Li, B., Zhou, H., He, J., Wang, M., Yang, Y., Li, L., 2020. On the sentence embeddings from pre-trained language models, in: Webber, B., Cohn, T., He, Y., Liu, Y. (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online. pp. 9119–9130. doi:10.18653/v1/2020.emnlp-main.733.

Litt, E., Hargittai, E., 2016. The imagined audience on social network sites. Social Media+ Society 2. doi:10.1177/2056305116633482.

Little, R.J.A., Rubin, D.B., 2002. Statistical Analysis with Missing Data. Wiley.

Liu, H., 2015. Comparing Welch ANOVA, a Kruskal-Wallis test, and traditional ANOVA in case of heterogeneity of variance. Virginia Commonwealth University.

Livermore, M., Chowdhury, M., Baumgartner, G., Jeanlouis, J., 2022. Organizational social media use and community social capital: Disparities by poverty and racial composition. Journal of Poverty 27, 374–390. doi:10.1080/10875549.2022.2080030.

Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA. p. 4768–4777.

Lustgarten, J.L., Balasubramanian, J.B., Visweswaran, S., Gopalakrishnan, V., 2017. Learning parsimonious classification rules from gene expression data using bayesian networks with local structure. Data 2, 5. doi:10.3390/data2010005.

Lynn, T., Rosati, P., Nair, B., 2020. Calculated vs. ad hoc publics in the# brexit discourse on Twitter and the role of business actors. Information 11, 435. doi:10.3390/info11090435.

Mark Graham, Bernie Hogan, R.K.S., Medhat, A., 2014. Uneven geographies of user-generated information: Patterns of increasing informational poverty. Annals of the Association of American Geographers 104, 746–764. doi:10.1080/00045608.2014.910087.

McDonald, N., Schoenebeck, S., Forte, A., 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. Proc. ACM Hum.-Comput. Interact. 3. doi:10.1145/3359174.

McInnes, L., Healy, J., Melville, J., 2018. UMAP: Uniform manifold approximation and projection for dimension reduction arXiv:1802.03426.

Meena, A., Bhatia, V., Pal, J., 2020. Digital divine: Technology use by Indian spiritual sects, in: Proceedings of the 2020 International Conference on Information and Communication Technologies and Development, Association for Computing Machinery, New York, NY, USA. doi:10.1145/3392561.3394650.

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781 [cs] URL: http://arxiv.org/abs/1301.3781.

Mislove, A., Lehmann, S., Ahn, Y.Y., Onnela, J.P., Rosenquist, J., 2011. Understanding the demographics of Twitter users, in: Proceedings of the international AAAI conference on web and social media, pp. 554–557. doi:10.1609/icwsm.v5i1.14168.

Moosbauer, J., Herbinger, J., Casalicchio, G., Lindauer, M., Bischl, B., 2021. Explaining hyperparameter optimization via partial dependence plots. Advances in Neural Information Processing Systems 34, 2280–2291.

Muñetón-Santa, G., Manrique-Ruiz, L.C., 2023. Predicting multidimensional poverty with machine learning algorithms: an open data source approach using spatial data. Social Sciences 12, 296. doi:10.3390/socsci12050296.

Murray, J.S., 2018. Multiple imputation: a review of practical and theoretical findings. Statistical Science 33, 142–159.

Muñetón-Santa, G., Manrique-Ruiz, L.C., 2023. Predicting multidimensional poverty with machine learning algorithms: An open data source approach using spatial data. Social Sciences 12, 296. doi:10.3390/socsci12050296.

NAX Solutions, 2022. Ndvi: Pros and cons. URL: https://naxsolutions.com/en/agriculture-precision-dictionary/ndvi-pros-cons/, doi:10.1302/3114-210265. accessed: 2025-03-21.

Ngidi, N.D., Mtshixa, C., Diga, K., Mbarathi, N., May, J., 2016. 'Asijiki' and the capacity to aspire through social media: The feesmustfall movement as an anti-poverty activism in South Africa, in: Proceedings of the Eighth International Conference on Information and Communication Technologies and Development, Association for Computing Machinery, New York, NY, USA. doi:10.1145/2909609.2909654.

Nia, Z.M., Ahmadi, A., Bragazzi, N.L., Woldegerima, W.A., Mellado, B., Wu, J., Orbinski, J., Asgary, A., Kong, J.D., 2022. A cross-country analysis of macroeconomic responses to COVID-19 pandemic using Twitter sentiments. Plos one 17, e0272208. doi:10.1371/journal.pone.0272208.

Niu, T., Chen, Y., Yuan, Y., 2020. Measuring urban poverty using multi-source data and a random forest algorithm: A case study in Guangzhou. Sustainable Cities and Society 54, 102014. doi:10.1016/j.scs.2019.102014.

North, D.C., Weingast, B.R., 1989. Constitutions and commitment: The evolution of institutions governing public choice in seventeenth-century england. The Journal of Economic History 49, 803–832. doi:10.1017/s0022050700009451.

Ohme, J., Araujo, T., Boeschoten, L., Freelon, D., Ram, N., Reeves, B.B., Robinson, T.N., 2024. Digital trace data collection for social media effects research: Apis, data donation, and (screen) tracking. Communication Methods and Measures 18, 124–141.

Oshri, B., Hu, A., Adelson, P., Chen, X., Dupas, P., Weinstein, J., Burke, M., Lobell, D., Ermon, S., 2018. Infrastructure quality assessment in Africa using satellite imagery and deep learning, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Association for Computing Machinery, New York, NY, USA. p. 616–625. URL: https://doi.org/10.1145/3219819.3219924, doi:10.1145/3219819.3219924.

Osuagwu, U.L., Miner, C.A., Bhattarai, D., Mashige, K.P., Oloruntoba, R., Abu, E.K., Ekpenyong, B., Chikasirimobi, T.G., Goson, P.C., Ovenseri-Ogbomo, G.O., et al., 2021. Misinformation about COVID-19 in Sub-Saharan Africa: Evidence from a cross-sectional survey. Health security 19, 44–56. doi:10.1089/hs.2020.0202.

Pal, J., 2017. The technological self in India: From tech-savvy farmers to a selfie-tweeting prime minister, in: Proceedings of the Ninth International Conference on Information and Communication Technologies and Development, Association for Computing Machinery, New York, NY, USA. doi:10.1145/3136560.3136583.

Panda, A., Chakraborty, S., Raval, N., Zhang, H., Mohapatra, M., Akbar, S.Z., Pal, J., 2020. Affording extremes: Incivility, social media and democracy in the indian context, in: Proceedings of the 2020 International Conference on Information and Communication Technologies and Development, Association for Computing Machinery, New York, NY, USA. doi:10.1145/3392561.3394637.

Park, C., Apley, D., 2018. Patchwork kriging for large-scale Gaussian process regression. The Journal of Machine Learning Research 19, 269–311.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, , 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830.

Perez-Heydrich, C., Warren, J.L., Burgert, C.R., Emch, M.E., 2013. Guidelines on the Use of DHS GPS Data. Technical Report 8. ICF International. Calverton, Maryland, USA. URL: https://dhsprogram.com/pubs/pdf/SAR8/SAR8.pdf.

Pokhriyal, N., Zambrano, O., Linares, J., Hernández, H., 2020. Estimating and Forecasting Income Poverty and Inequality in Haiti Using Satellite Imagery and Mobile Phone Data. Inter-American Development Bank. doi:10.18235/0002466.

Poushter, J., Bishop, C., Chwe, H., 2018. Social media use continues to rise in developing countries but plateaus across developed ones. Pew Research Center 22, 2–19.

Preoţiuc-Pietro, D., Volkova, S., Lampos, V., Bachrach, Y., Aletras, N., 2015a. Studying user income through language, behaviour and affect in social media. PLOS ONE 10, e0138717. doi:10.1371/journal.pone.0138717.

Preoţiuc-Pietro, D., Volkova, S., Lampos, V., Bachrach, Y., Aletras, N., 2015b. Studying user income through language, behaviour and affect in social media. PloS one 10, e0138717.

Ragnedda, M., 2019. Reconceptualising the digital divide, in: Mutsvairo, B., Ragnedda, M. (Eds.), Mapping the Digital Divide in Africa: A Mediated Analysis. Amsterdam University Press, Amsterdam, pp. 27–43. doi:10.2307/j.ctvh4zj72.6.

Rehman, N.A., Relia, K., Chunara, R., 2018. Creating full individual-level location timelines from sparse social media data, in: Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Association for Computing Machinery, New York, NY, USA. p. 379–388. doi:10.1145/3274895.3274982.

Reimers, N., Gurevych, I., 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks doi:10.18653/v1/d19-1410.

Relia, K., Akbari, M., Duncan, D., Chunara, R., 2018. Socio-spatial self-organizing maps: Using social media to assess relevant geographies for exposure to social processes. Proc. ACM Hum.-Comput. Interact. 2. doi:10.1145/3274414.

Reyes, D., Antonio, R.J., Orden, A., Heinrich, A., Phoa, R., Bilal, S., Bonganay, G., Singson, M., 2023. Wealth index estimation using machine learning with environmental, demographics, remote sensing, and points of interest data, in: Proceedings of the 1st ACM SIGSPATIAL International Workshop on Geocomputational Analysis of Socio-Economic Data, ACM, New York, NY, USA. p. 12–19. doi:10.1145/3615892.3628478.

Robinson, J.A., Acemoglu, D., 2012. Why nations fail: The origins of power, prosperity and poverty. Profile London.

Röder, M., Both, A., Hinneburg, A., 2015. Exploring the Space of Topic Coherence Measures, in: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, ACM, New York, NY, USA. pp. 399–408. doi:10.1145/2684822.2685324.

Rowlands, J., 1995. Empowerment examined. Development in practice 5, 101–107. doi:10.1080/0961452951000157074.

Rowlands, J., 1997. Questioning empowerment: Working with women in Honduras. Oxfam.

Sachs, J.D., 2006. The end of poverty: Economic possibilities for our time. Penguin. doi:10.23962/10539/19811.

Saha, M., Varghese, D., Bartindale, T., Thilsted, S.H., Ahmed, S.I., Olivier, P., 2022. Towards sustainable ICTD in Bangladesh: Understanding the program and policy landscape and its implications for CSCW and HCI. Proc. ACM Hum.-Comput. Interact. 6. doi:10.1145/3512973.

Sahn, D.E., Stifel, D., 2003. Exploring alternative measures of welfare in the absence of expenditure data. Review of income and wealth 49, 463–489. doi:10.1111/j.0034-6586.2003.00100.x.

Santner, T.J., Williams, B.J., Notz, W.I., 2018. The Design and Analysis of Computer Experiments (Second Edition). Springer New York. doi:10.1007/978-1-4757-3799-8_1.

Savoia, A., Sen, K., 2015. Measurement, evolution, determinants, and consequences of state capacity: A review of recent research. Journal of economic surveys 29, 441–458. doi:10.1111/joes.12065.

Sedda, L., Tatem, A.J., Morley, D.W., Atkinson, P.M., Wardrop, N.A., Pezzulo, C., Sorichetta, A., Kuleszo, J., Rogers, D.J., 2015. Poverty, health and satellite-derived vegetation indices: their inter-spatial relationship in west africa. International health 7, 99–106. doi:10.1093/inthealth/ihv005.

Sen, A., 2005. Human rights and capabilities. Journal of human development 6, 151–166. doi:10.1080/14649880500120491.

Sen, A., 2014. Development as freedom (1999). The globalization and development reader: Perspectives on development and global change 525.

Shang, L., Zhang, Y., Youn, C., Wang, D., 2022. SAT-Geo: A social sensing based content-only approach to geolocating abnormal traffic events using syntax-based probabilistic learning. Information Processing Management 59, 102807. URL: https://www.sciencedirect.com/science/article/pii/S0306457321002843, doi:https://doi.org/10.1016/j.ipm.2021.102807.

Sheehan, E., Meng, C., Tan, M., Uzkent, B., Jean, N., Burke, M., Lobell, D., Ermon, S., 2019. Predicting economic development using geolocated wikipedia articles, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Association for Computing Machinery, New York, NY, USA. p. 2698–2706. doi:10.1145/3292500.3330784.

Sheth, A., 2009. Citizen sensing, social signals, and enriching human experience. IEEE Internet Computing 13, 87–92. doi:10.1109/MIC.2009.77.

Shingala, M.C., Rajyaguru, A., 2015. Comparison of post hoc tests for unequal variance. International Journal of New Technologies in Science and Engineering 2, 22–33.

Shoemaker, E., Malik, H., Narman, H., Chaudri, J., 2023. Explaining the unseen: Leveraging XAI to enhance the trustworthiness of black-box models in performance testing. Procedia Computer Science 224, 83–90. doi:10.1016/j.procs.2023.09.014.

Simon, T., Goldberg, A., Aharonson-Daniel, L., Leykin, D., Adini, B., 2014. Twitter in the cross fire—the use of social media in the Westgate Mall terror attack in Kenya. PLOS One 9, e104136. doi:10.1371/journal.pone.0104136.

Snoek, J., Larochelle, H., Adams, R.P., 2012. Practical Bayesian optimization of machine learning algorithms, in: Advances in Neural Information Processing Systems.

Srikanth, M., Liu, A., Adams-Cohen, N., Cao, J., Alvarez, R.M., Anandkumar, A., 2021. Dynamic social media monitoring for fast-evolving online discussions, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, Association for Computing Machinery, New York, NY, USA. p. 3576–3584. doi:10.1145/3447548.3467171.

Steele, J.E., Sundsøy, P.R., Pezzulo, C., Alegana, V.A., Bird, T.J., Blumenstock, J., Bjelland, J., Engø-Monsen, K., De Montjoye, Y.A., Iqbal, A.M., et al., 2017. Mapping poverty using mobile phone and satellite data. Journal of The Royal Society Interface 14, 20160690. doi:10.1098/rsif.2016.0690.

Stein, M.L., 1999. Interpolation of Spatial Data: Some Theory for Kriging. Springer: New York.

Suk, J., Zhang, Y., Yue, Z., Wang, R., Dong, X., Yang, D., Lian, R., 2023. When the Personal Becomes Political: Unpacking the Dynamics of Sexual Violence and Gender Justice Discourses Across Four Social Media Platforms. Communication Research 50, 610–632. doi:10.1177/

00936502231154146. publisher: SAGE Publications Inc.

Tang, B., Sun, Y., Liu, Y., Matteson, D.S., 2018. Dynamic poverty prediction with vegetation index, in: Workshop on Modeling and Decision-Making in the Spatiotemporal Domain, 32nd Conference on Neural Information Processing Systems (NIPS 2018).

Toyama, K., 2014. Teaching how to fish: lessons from information and communication technologies for international development. Journal of Marketing Management 30, 439–444. doi:10.1080/0267257x.2014.884621.

Toyama, K., 2015. Geek heresy: Rescuing social change from the cult of technology. Public Affairs.

Tsaneva, S., Dessì, D., Osborne, F., Sabou, M., 2025. Knowledge graph validation by integrating LLMs and human-in-the-loop. Information Processing Management 62, 104145. doi:10.1016/j.ipm.2025.104145.

Tufekci, Z., 2014. Big questions for social media big data: Representativeness, validity and other methodological pitfalls, in: Proceedings of the international AAAI conference on web and social media, pp. 505–514. doi:10.1609/icwsm.v8i1.14517.

Walk, E., Garimella, K., Christia, F., 2023. Displacement and return in the internet Era: Social media for monitoring migration decisions in Northern Syria. World Development 168, 106268. doi:10.1016/j.worlddev.2023.106268.

Yoshida, N., Takamatsu, S., Yoshimura, K., Aron, D., Chen, X., Malgioglio, S., Shivakumaran, S., Zhang, K., 2022. The Concept and Empirical Evidence of SWIFT Methodology. Technical Report. World Bank. doi:10.1596/38095.

Zhao, X., Yu, B., Liu, Y., Chen, Z., Li, Q., Wang, C., Wu, J., 2019. Estimation of poverty using random forest regression with multi-source data: A case study in Bangladesh. Remote Sensing 11, 375. doi:10.3390/rs11040375.

Zhigljavsky, A., Žilinskas, A., 2008. Stochastic Global Optimization. Springer: US. doi:10.1007/978-0-387-74740-8_3.

Řehůřek, R., Sojka, P., 2010. Software framework for topic modelling with large corpora, in: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, ELRA, Valletta, Malta. pp. 45–50.