Lady and the Tramp Nextdoor:
Online Manifestations of Real-World Inequalities in the Nextdoor Social Network

Waleed Iqbal[1], Vahid Ghafouri[2, 3], Gareth Tyson[1, 4],
Guillermo Suarez-Tangil[2], Ignacio Castro[1]

[1]Queen Mary University of London,
[2]IMDEA Networks Institute,
[3]Universidad Carlos III de Madrid,
[4]Hong Kong University of Science and Technology (GZ)
{w.iqbal, i.castro}@qmul.ac.uk, {vahid.ghafouri, guillermo.suarez-tangil}@imdea.org, gtyson@ust.hk

Abstract

From health to education, income impacts a huge range of life choices. Earlier research has leveraged data from online social networks to study precisely this impact. In this paper, we ask the opposite question: do different levels of income result in different online behaviors? We demonstrate it does. We present the first large-scale study of Nextdoor, a popular location-based social network. We collect 2.6 Million posts from 64,283 neighborhoods in the United States and 3,325 neighborhoods in the United Kingdom, to examine whether online discourse reflects the income and income inequality of a neighborhood. We show that posts from neighborhoods with different incomes indeed differ, *e.g.* richer neighborhoods have a more positive sentiment and discuss crimes more, even though their actual crime rates are much lower. We then show that user-generated content can predict both income and inequality. We train multiple machine learning models and predict both income (R^2=0.841) and inequality (R^2=0.77).

Introduction

Income is a critical factor in many aspects of life and a large body of research has examined the effect of income on how individuals interact with the real world (Wu 2012; Dekker 2007; Hays and Kogl 2007). In this paper, we address the opposite question: do income differences result in differences in online discourse? We hypothesize that this is the case (Bernstein 1960) and that these differences can be leveraged to infer the economic context of a person.

To test our hypothesis we need user data on online discourse and income. Online interaction data is abundant, but inferring the income of online users is challenging. Mobility data and location-based applications such as Foursquare can be used to infer the residential location of a user and then using official statistics associate, the user with the level of income of the user's neighborhood (Aggarwal, Almeida, and Kumaraguru 2013; Chorley et al. 2016). This presents two challenges: (*i*) the residential location of the user needs to be inferred using heuristics, and (*ii*) the data might inform of the user's location but will have limited vantage on the online discourse of the user.

Alternatively, online discourse data (*e.g.* from Twitter or Mastodon) can be used to study the content posted by online users and develop heuristics to infer their income, *e.g.* using the expected level of income for a given professional occupation (Preoţiuc-Pietro, Lampos, and Aletras 2015; Aletras and Chamberlain 2018). However, this implies that (*i*) income inferences rely on heuristics, and (*ii*) there might be selection bias as not all users might post content that can be used to infer their income.

In this work, we overcome those limitations by collecting and analyzing a previously unstudied location-based social network with over 10 million users, *Nextdoor* (Nextdoor 2021). Nextdoor users interact within closed social networks of neighbors, *i.e.* users registered in the same neighborhood. To ensure that users register in the location where they reside, new users have to validate their home addresses using regular (snail) mail. This allows us to associate neighbors with the median income of their neighborhood and study whether differences in income are reflected in their online discourse.

We present the first large-scale dataset and analysis of Nextdoor. We collect 2.6 Million posts from 64,283 neighborhoods in the United States (USA) and 3,325 neighborhoods in the ten most populous cities in the United Kingdom (UK) between November 2020 and September 2021. We augment this with official statistics on population, income, and crime for both countries and at the same level of geographical granularity as the Nextdoor neighborhoods.

We use our augmented dataset to study whether differences in income indeed result in online discourse differences with the following Research Questions (RQs):

- **RQ1:** Do neighborhoods with different income levels differ in how they discuss crime and the sentiment of their posts?

- **RQ2:** Do neighborhoods with different levels of income inequality differ in how they discuss crime and on the sentiment of their posts?

- **RQ3:** Can Nextdoor discussions be used to predict the income and income inequality of a user?

While answering these questions, we conduct a neighborhood-level analysis of our Nextdoor dataset. We examine how neighborhoods with different *income*

levels exhibit different traits in the online user-generated content. We identify strikingly clear differences. We find that richer neighborhoods seem more concerned about crime: The 20% richest neighborhoods discuss crime 1.6x and 1.47x more than the 80% poorest counterparts in USA and UK, respectively. This is the case even though the incidence of crime is 1.24x and 1.36x higher in poorer neighborhoods in USA and UK. We also find that richer neighborhoods seems to post more positive content: the sentiment of the text in the posts of the 20% richest neighborhoods is 1.37x and 1.42x more positive than in the 80% poorest neighborhoods in USA and UK.

We then examine *income inequality*. We determine whether a neighborhood is more or less unequal by comparing its median income with the income of the neighborhoods in its vicinity. We then look for differences in the content posted between neighborhoods with different levels of inequality in their vicinity. Again, we find clear differences between neighborhoods depending on the level of income inequality. The richest neighborhoods with the most *equal* vicinity, discuss crime more than any other neighborhood, in proportion to the official crime reported. We also find that the richest neighborhoods with the most *equal* vicinity have a more positive sentiment than any other neighborhood. The opposite is true for the poorest neighborhoods, where the neighborhoods with the most equal vicinities have the lowest sentiment of all.

We then wonder whether we can *predict* a neighborhood's income based on the text posted. We show that this is indeed the case. We predict the income of a neighborhood exclusively using text features obtained from the posts (up to 0.69 R^2 with a Lasso regressor). We then experiment with several machine learning models to predict income and inequality using features extracted from Nextdoor. We find that we can predict a neighborhood's income and its surroundings' inequality with even a higher degree of accuracy (income best R^2: 0.841, inequality best R^2: 0.77). We find this concerning, as it could facilitate income-based discrimination and algorithmic surveillance (Zuboff 2015).

This paper makes the following contributions:

- We conduct the first large-scale quantitative analysis of Nextdoor with 2.6 Million posts from 64,283 and 3,325 USA and UK neighborhoods, respectively.

- We show that the online content generated by the users reveals socioeconomic factors where a greater income of a neighborhood and its surroundings is associated with more crime-sensitive posting activity and more positive sentiment in the posts.

- We demonstrate that the features extracted from the user-generated content can predict both the income (R^2=0.841) and inequality (R^2=0.77) of the neighborhoods where they reside.

Data and Methodology

Nextdoor Data

Nextdoor primer. Nextdoor is a location-based social network with over 270K registered neighborhoods in 11 coun-

tries and over 10 million users (Nextdoor 2021). Nextdoor divides geographical areas into neighborhoods. Nextdoor assigns users to the *neighborhood* where they resides. To ensure that a user is a neighbor of a particular neighborhood, new users validate their home addresses via regular (snail) mail.

For each neighborhood, Nextdoor creates a dedicated forum where users can post and interact (*e.g.* reply, and react to each other's posts). Users exclusively interact with their neighbors, *i.e.* the users of the neighborhood they are associated with. As a result, the data from a neighborhood exclusively includes the posts of the users that have validated their location in that geographical area.

In the rest of this paper, we use the term *neighborhood* to refer to the specific areas into which Nextdoor divides a region and *neighbor* to refer to the Nextdoor user registered as living in a neighborhood.

Neighborhoods. We collect 2,201,051 posts from 64,283 USA neighborhoods and 351,894 posts from all the 3,325 neighborhoods in the 10 UK cities with the largest population. UK cities include London, Birmingham, Liverpool, Sheffield, Bristol, Glasgow, Edinburgh, Leeds, Manchester, and Bradford. Due to data scraping limitations, We have incomplete data for 13 USA states (see the percentage of neighborhoods for which we collected data in parenthesis): Texas (90.69%), California (88.6%), Georgia (80.1%), Florida (84.43%), Alaska (91.17%), Washington (78.6%), Wisconsin (85.81%), Virginia (80.51%), Alabama (82.62%), Nevada (88.02%), New Jersey (78.4%), Louisiana (87.23%), and New Mexico (91.1%). Overall our data includes 76.2% of neighborhoods in the USA, covering 65.8% of the population in the USA. For the 10 UK cities, our data is complete, *i.e.* the data collected covers all their neighborhoods. Our data includes almost one full year (November 2020 – September 2021). Note that we count Washington, District of Columbia (DC), as a state because of its autonomous status.

We obtain a list of neighborhood names directly from the Nextdoor website for the USA (Nextdoor 2022). For the UK, a list of neighborhoods does not exist on Nextdoor. Instead, we collect the list of UK neighborhoods by crawling the map integrated into Nextdoor. Due to the resource-intensive nature of this process, we collect data for the neighborhoods of the ten most populous UK cities.

Neighborhood location. For each neighborhood, we obtain neighborhood names and locations (*i.e.* latitude and longitude). For the UK, we obtain the latitude and longitude of a neighborhood while crawling the Nextdoor map. For the USA, we employ the GeoPy API to map the coordinates of a neighborhood to its zip code (GeoPy 2022). We then map the coordinates of each neighborhood into the lowest geographical granularity at which official statistical data is reported: Zip code for USA, and Lower Layer Super Output Area (LSOA) in the UK, a small geographic area used for UK's statistical reporting. To map a UK postcode to its LSOA, we use official data (ONS 2018).

Neighbors and posts. We then iterate through each neighborhood to collect posts written by its neighbors. In total,

Attributes	USA	UK	Total
Posts	2,201,051	351,894	2,602,045
Neighborhoods	64,283	3,325	67,608
Cities	5,849	10	5,859
zip code(USA)/LSOA(UK)	30872	2512	33284
Comments	17,421,050	2,246,814	19,667,864
Neighbors	6,6480,730	1,744,948	68,225,678

Table 1: Nextdoor dataset.

we obtain over 2.6M posts from 67,608 neighbors. Table 1 shows the main attributes of the data.

Feature Engineering

Crime data, population, and income. For each neighborhood, we obtain socioeconomic data at the zip code and LSOA level by querying governmental databases of official statistics. For the USA zip codes, we obtain the population and median annual income from the latest Census (Census 2022). We obtain crime per 10,000 people data from FBI's Crime Data Explorer (FBI 2022). For the UK LSOAs, we collect the median annual income and population data from the UK's Office of National Statistics (UK 2022) using its latest Census update (2021) and crimes per 10,000 people from the UK Metropolitan Police (Police 2022).

Inequality. Income inequality data is unavailable at the neighbor-level. Instead, we calculate inequality at the neighborhood level: we measure the inequality between the group of neighborhoods that are in the same vicinity. As an inequality metric we compute the Atkinson Index for each neighborhood (Atkinson, Micklewright, and Micklewright 1992):

$$A(y_1, ..., y_n) = 1 - \frac{1}{\mu}(\frac{1}{N}\sum_{i=1}^{N} y_i) \qquad (1)$$

where A is the Atkinson index of neighborhood i, y_i is its income, N is the set of nearby neighborhoods (including i), μ is the mean income of the N neighborhoods. Note that the Atkinson index uses an inequality aversion parameter which we equal to zero, to avoid any assumptions about the impact of inequality. A set of N neighborhoods with identical income will have an Atkinson Index equal to zero. The greater the difference between the incomes, the closer to 1 the index is.

We then compute each neighborhood's set of N nearby neighborhoods. To define the vicinity of a neighborhood, we identify the minimum radio that still renders at least another nearby neighborhood for each target neighborhood. We compute the distance between each pair of neighborhoods with the Haversine formula (Inman 1849). We find that a minimum radius of 24.92 and 2.97 miles for the USA and UK is necessary to have non-empty surroundings for all neighborhoods. We round up these numbers and use a radio of 25 and 3 miles for the USA and UK, respectively. We also discover 418 USA neighborhoods more than 100 miles from the nearest neighborhood. We verify that they are, in fact, in remote, isolated areas. Furthermore, due to the incomplete scraping of USA neighborhoods, many remote, isolated neighborhoods are not in our dataset.

Crime-related posts. To assess how sensitive to crime a neighborhood is, we identify the posts discussing crime and compare the number with the actual crime reported in that neighborhood. We focus on crime because it is the only topic for which we found official statistics with rich and detailed geolocated data for both countries. Additionally, manual inspection revealed that Nextdoor users commonly discuss crime in their neighborhoods.

We label each post according to whether it discusses crime and the type of crime using the Semantic Search application of the pre-trained *msmarco-distilbert-base-v4* Sentence-BERT (S-BERT) model (Reimers and Gurevych 2019). This model uses siamese and triplet network architectures to generate semantically significant sentence embedding. We then compute the cosine similarity between the embedding of a post and the description of the official crime category. We obtain the official categories of a crime from the FBI's Crime Data Explorer (FBI 2022), and the UK Metropolitan Police's Crime (Police 2022). We note that the categories of the same type of crime vary between the two countries. In particular, the USA definition of violent crimes is narrower than in the UK. While this does not affect within-country comparisons, we caution about the USA–UK comparisons. For consistency, we only consider categories in both countries' statistics. Additionally, we group crime categories into three major categories; (i) Drugs and Order, (ii) Theft and Property Damage, and (iii) Weapons and Violent Crimes. See the complete list of crime categories in Table 6.

We consider that a post discusses a crime whenever the similarity is greater than a threshold. To determine the threshold, our native English-speaking human annotator takes a random sample of 2,000 posts, 1,000 from each country from S-BERT tagged data, and manually annotates them as crime or no-crime discussions. When a post discusses a crime, our annotator also annotates the type of crime discussed. We repeat this process for several thresholds and use Cohen's Kappa score (K) (McHugh 2012) between the S-BERT and our manual annotation as the criteria for the threshold. Table 2 reports the thresholds and the corresponding K. We choose 0.7 as our threshold, which maximizes K for the USA and UK (K=0.971, K=0.944 in the USA and UK, respectively). We then use this 0.7 threshold to label 543,459 USA posts and 124,763 UK posts as crime-discussing.

Cosine Similarity	K (USA)	K (UK)
0.5	0.875	0.847
0.6	0.948	0.895
0.7	0.971	0.944
0.8	0.939	0.917
0.9	0.926	0.891

Table 2: Cosine similarity thresholds and their respective Cohen's Kappa Score (K).

Post sentiment. To assess how positive a neighborhood's

	Population		Posts		Neighborhoods		Neighbors	
	USA	UK	USA	UK	USA	UK	USA	UK
Population	1	1	0.83	0.96	0.93	0.96	0.96	0.97
Posts	0.83	0.96	1	1	0.91	0.97	0.84	0.89
Neighborhoods	0.93	0.96	0.91	0.97	1	1	0.93	0.95
Neighbors	0.96	0.97	0.84	0.89	0.93	0.95	1	1

Table 3: Correlation between posts, population, neighborhoods, neighbors, and official population.

posts is, we label each post's sentiment in our dataset with a pre-trained Valence Aware Dictionary and Sentiment Reasoner (VADER) model (Hutto and Gilbert 2014). VADER is a lexicon and rule-based sentiment analysis instrument that outperforms the typical human reader in its sensitivity to assumptions transmitted in web-based media (VADER's F1=0.96, Human F1=0.84) (Hutto and Gilbert 2014). VADER works well on content-aware text with emotions, such as social media text. As people use Nextdoor as a social network and Nextdoor itself encourages people to post using few words, VADER is appropriate for Nextdoor data.

Text embedding. To investigate whether the text posted generally differs between richer and poorer neighborhoods, we obtain semantic features of the posts via embedding. We preprocess every post (removing mentions, URLs, *etc.*) and discard neighborhoods with less than 10 posts. Our final dataset consists of 63,587 neighborhoods from USA and 3,282 neighborhoods from the UK. Then, we convert each post's text into a sentence embedding using the best-performing pre-trained sentence transformer model, "all-mpnet-base-v2" powered by *Hugging-Face* (Huggingface 2022). This model is tuned to map every sentence or short paragraph to a 768-dimensional vector space while preserving important text features. We then aggregate the sentence embedding of all the posts of a neighborhood. Similarly to (Arora, Liang, and Ma 2017), we employ an element-wise mean-pooling aggregation method.

We obtain a 768 dimensions (mean-pooled) feature per neighborhood. This large number is problematic due to the (*i*) Curse of Dimensionality effect (Aggarwal, Hinneburg, and Keim 2001), and (*ii*) some features would outweigh the effect of others if we were to simply concatenate them. We avoid these problems by reducing the textual embedding with Uniform Manifold Approximation and Projection (UMAP) (McInnes et al. 2018), the state-of-the-art method. We reduce the 768 dimensions to 5, as suggested in (Zhang et al. 2022; Jimenez Villalonga 2021).

Data Representativity

We observe that the 20 most populated USA states (40%) account for 70% of the posts, 72% of the neighborhoods, and 69% of the neighbors. The UK is more skewed, with London concentrating 70%, 61%, and 76% of the UK's posts, neighborhoods, and neighbors, respectively.

Methodology. To assess whether this reflects the residents' concentration in those areas, we look at the correlation between Nextdoor data attributes and the official population

distribution in Table 3 using Pearson correlation coefficient (Lee Rodgers and Nicewander 1988). We calculate the correlation at the neighborhood level (*e.g.* users in a neighborhood and the corresponding official population), except for neighborhoods, which we do at the city level. We find a high correlation for each attribute, with a maximum of 0.97 (USA neighbors to population) and a minimum of 0.83 (USA posts to population). These high numbers give us confidence that the data of Nextdoor is representative of the USA and UK populations. See the Appendix for more details on the distribution of Nextdoor activity and population across USA states and UK cities.

Income and representativity. To examine how well-represented neighborhoods of different income levels are, we look at the correlation between Nextdoor attributes and the underlying populations in Figure 1 (similarly to Table 3). We find a remarkably high correlation between the number of posts, neighbors, neighborhoods, and the underlying population for both the USA and the UK. Correlations range between 0.88–0.98 for the UK and 0.79–0.93 for the USA, with most correlations above 0.9. These correlations tend to be higher for richer deciles and are usually higher for the UK. Further investigation shows that in the USA, residents of rich neighborhoods are more likely to be registered in Nextdoor. We find that the USA ratio of population-to-users (at the neighborhood level) tends to be indeed higher for the
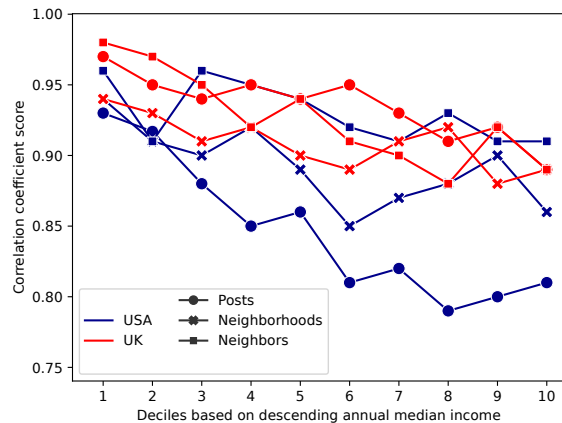

Figure 1: Correlation between posts, neighborhoods, neighbors, and official population over income deciles (from richer to poorer).

richest deciles (below 8) than for the poorer ones (below 18). See Figure 10 in Appendix for more details.

Throughout the rest of the paper, we compare the richest 20% and the poorest 80%. Since the richest deciles are better represented and the rest are aggregated together, this will help ensure that our results remain representative of those levels of income. For the USA, 29.87% neighborhoods fall in the richest 20%, and 70.13% are part of the poorest 80%. In the UK, 42.5% neighborhoods are in the richest 20%, and 57.5% are in the poorest 80%. We again caution regarding cross-country comparisons due to these differences.

Income and Online Discussions

With the above data, we now examine whether income differences result in different online behaviors. First, we study differences in crime discussions between richer and poorer neighborhoods. Then we compare the sentiment of posts for the same income categories.

Income and Crime Discussions

Richer neighborhoods seem to be more sensitive about crime. Figure 2 shows the official crime rate per 10,000 people and the crime-discussion rate per 10,000 Nextdoor neighbors for the 20% richest and 80% poorest neighborhoods in USA and UK. We find that neighbors in rich neighborhoods discuss crime more often than those in poorer ones. Interestingly, this is the case even though the actual crime rate is substantially higher in poorer neighborhoods. To further analyze this, we breakdown crimes into three types using the official crime categories: (*i*) Drugs and Order; (*ii*) Theft and Property Damage; and (*iii*) Weapons and Violent Crimes. See more details about considered crime categories in Table 6.

Non-violent crimes are discussed more than violent crimes on Nextdoor. We observe that non-violent crimes are dis-
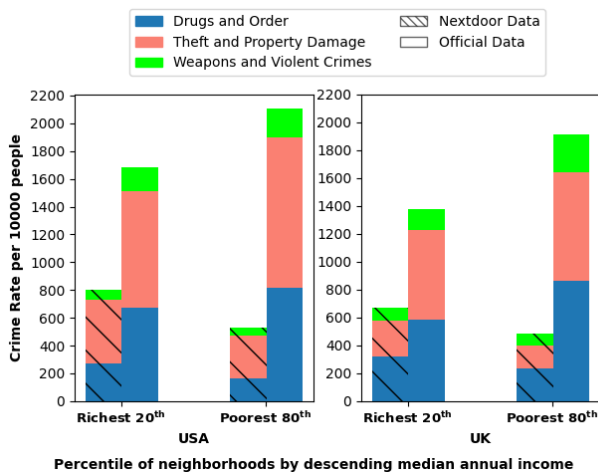

Figure 2: Official crime and crime discussion rates (per 10,000 people) in Nextdoor for the richest 20th and poorest 80th percentile neighborhoods.
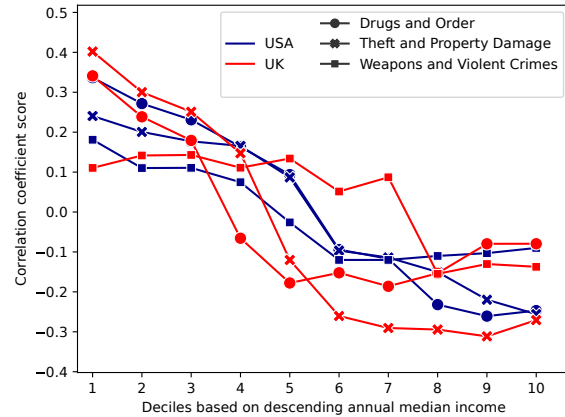

Figure 3: Correlation between crimes per 10000 people reported officially and discussed on Nextdoor over income deciles (from richer to poorer).

cussed more than weapons and violent crimes across all neighborhoods. To analyze these trends, we calculate the Pearson correlation coefficient between the rates of official crime rate and Nextdoor crime discussions. Figure 3 shows these correlations across crime types for the USA and UK from richest (left) to poorest percentiles (right). We observe a clear trend across all types of crime and countries: richer neighborhoods discuss these crimes proportionally to their actual occurrence. This changes abruptly as we move towards poorer neighborhoods, with the poorest neighborhoods discussing crime disproportionately to the actual crime occurrence.

Whereas most trends are very similar for both USA and UK, weapons and violent crimes show different behavior. These crimes are discussed proportionally more in the USA than in the UK for the richer neighborhoods, but the middle-income neighborhoods in UK tend to discuss this type of crime more than their USA counterparts. As we mentioned in Data and Methodology Section, the definitions of crime in the USA and the UK differ. This is particularly true for violent crime and possibly responsible for these differences.

Income and User Sentiment

As prior work indicates that income and self-reported positive sentiment are usually correlated (Easterlin 1974; Giorgi et al. 2021; Cui et al. 2022), we now look at the relationship between income and post sentiment. We calculate the median of the sentiment scores for the posts of the neighborhoods in the same income group and each % of the duration of the month. Figure 4 shows the median compound sentiment scores across all months in our dataset for the 20% richest and 80% poorest neighborhoods. Note that the higher the sentiment the more positive the text is.

Richer neighborhoods seem to be more positive. Our analyses show clear differences in user post sentiment across income levels. Posts from richer neighborhoods have more
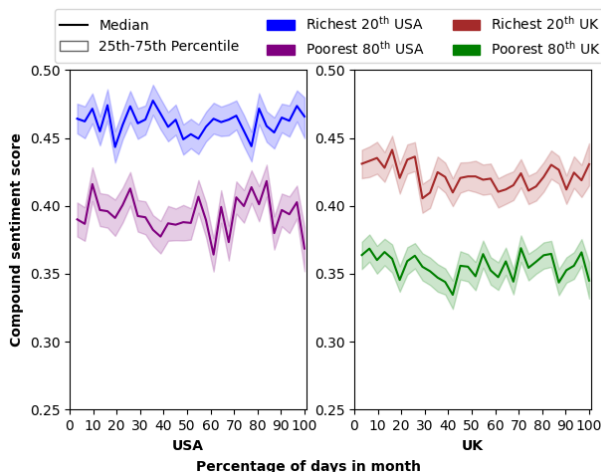
Figure 4: Median compound sentiment score for the richest 20% and poorest 80% neighborhoods.

positive sentiment both in the USA and UK, where the richest 20% have a clearly higher sentiment than the bottom 80%.

Income and Text Features

The previous analysis on crime discussion and post sentiment points to clear differences in online discourse depending on the income of a neighborhood. We now examine whether the text of the posts, in general, also reveals these differences. We employ the text embedding and dimensionality reduction described in Data and Methodology Section to study whether the content differs depending on the income of a neighborhood. Figures 5a and 5b depict the 2D projection of the embedding vectors per location in the USA and UK, respectively.

The dots, which represent locations, are color-coded in a "seismic" continuous colormap to portray the highest-income locations as dark red, and the lowest-income neighborhoods as dark blue. The closer the income is to the me-
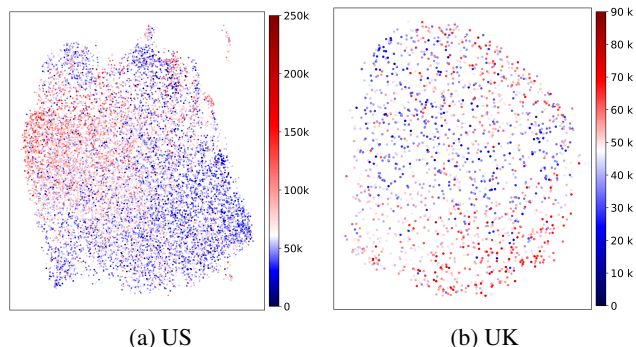

Figure 5: 2D projection of semantic features posted from every neighborhood. The "seismic" colormap represents the lowest-income neighborhood with the darkest blue and the highest-income with the darkest red.

dian income, the lighter the color is. The level of separation between the two colors indicates the level of online content polarization based on income. Since blue and red dots are visibly separated from each other in Figure 5, we confirm that the text posted by users in poor neighborhoods is semantically distinguishable from the texts generated in wealthier neighborhoods.

The income-based semantic differences in the user-generated text can stem from various qualities such as discussed topics. We showed that this is the case for income, where we had detailed and geolocated official data, but it could be true for others themes as well. As the semantic encoding provided by sentence transformers is black-box in its nature, understanding the underlying aspects of these semantic differences is beyond the scope of this paper.

Income Inequality and Online Discussions

In previous sections, We demonstrate how income manifests in online discussions. In this section, we explore whether income inequality also plays a role. Unfortunately, there is no official income inequality data available at the neighborhood level. Instead, we calculate how equal or unequal a neighborhood and its surroundings are. For each neighborhood, we identify other neighborhoods within its vicinity as described in Data and Methodology Section. We then calculate the Atkinson score for the set of neighborhoods across the surrounding area (including the target neighborhood). This gives us a metric of how unequal the vicinity is at the neighborhood level, *i.e.* it does not reveal how unequal neighborhoods are within themselves.

To see how income and inequality (as defined above) interact, we classify neighborhoods depending on their income and the inequality level of their surroundings. We again classify neighborhoods into the 20% richest and the 80% poorest. We also classify the vicinity of a neighborhood into the 20% most equal and the 80% most unequal. When combining income and inequality classifications, we identify 4 categories of neighborhoods:

- **RE**: Richest neighborhoods and most Equal vicinity.
- **RU**: Richest neighborhoods and most Unequal vicinity.
- **PE**: Poorest neighborhoods and most Equal vicinity.
- **PU**: Poorest neighborhoods and most Unequal vicinity.

Income Inequality and Crime Discussion

We investigate how crime is discussed in neighborhoods according to their income and the income inequality with their vicinity. We again find remarkably similar trends for the USA and UK. Figure 6 shows the crime discussed in Nextdoor and officially reported per 10,000 people for the 4 categories of income-inequality.

We observe that the Richest and Equal neighborhoods (RE), seem to be more crime-sensitive as crime is discussed there more often. While the difference is small (particularly in the UK), since the official rate is higher for the neighborhoods in more unequal vicinities (RU), residents indeed discuss crime proportionally more.

Exactly the opposite is true for the poorest neighborhoods. For the poorest neighborhoods, crime is less discussed when
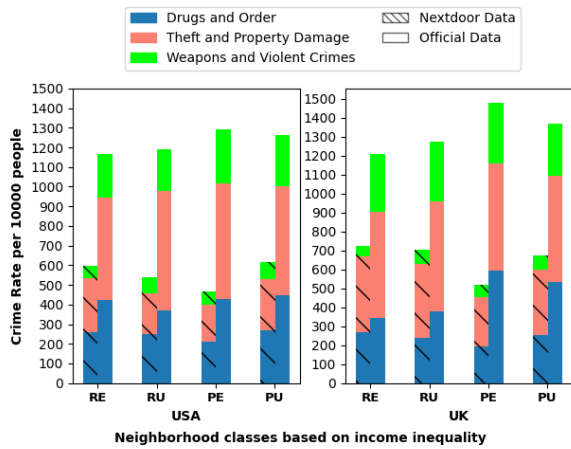
Figure 6: Crime discussion rate (Nextdoor) and official crime rate per 10,000 people.

the income of vicinity is more equal (PE) than when it is more unequal (PU). This is true even though the actual crime goes in the opposite direction: higher for more income-equal vicinities than for unequal ones in poorer neighborhoods. This indicates that the higher the income within a neighborhood or its surroundings, the more it is likely for crime to be discussed online.

We then calculate the Pearson correlation coefficient between the official crime rate and crime discussed in Nextdoor for each category in Figure 7. This figure confirms previously observed trends between crime rate and crime discussion for all categories except for "weapons and violent crimes" which follow the opposite trend in the richest neighborhoods, *i.e.* most equal neighborhoods discuss this type of crime proportionally less than the more unequal ones.
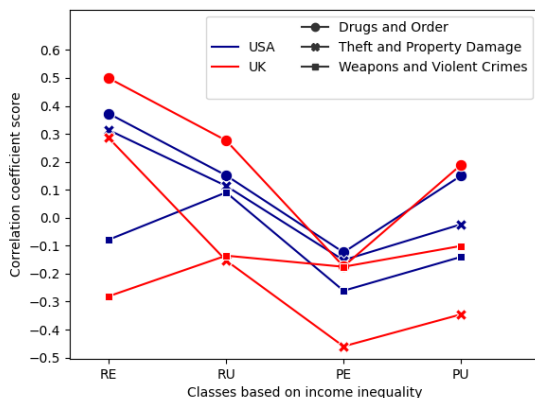

Figure 7: Correlation between crimes (per 10,000 people) discussed (Nextdoor) and officially reported in neighborhoods according to their income and inequality.
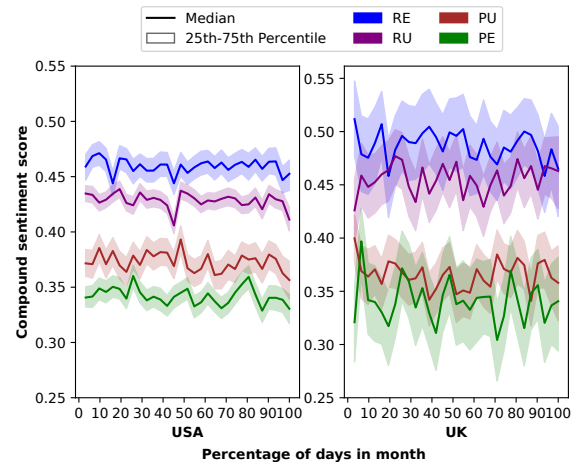

Figure 8: Median compound sentiment score of Nextdoor posts in neighborhoods with different income and inequality.

Income Inequality and User Sentiment

We now look at the relationship between income inequality and sentiment at the neighborhood level. We calculate the median of the sentiment scores for the posts of the neighborhoods in the same income inequality classes and each % of the duration of the month.

Figure 8 shows the median compound sentiment scores across all days of a month (in percentage) in our dataset for all income inequality classes. Note that the higher the sentiment the more positive the text is.

Richer neighborhoods again seem more positive: the richest neighborhoods tend to have a higher positive sentiment regardless of the level of inequality in their vicinity. Richer neighborhoods with more equal vicinities (RE) have a more positive sentiment. We observe the opposite in the poorest neighborhoods, more equal vicinities are associated with a lower sentiment (PE). These findings are again almost identical for both countries and again indicate that the higher the income within a neighborhood or its surroundings, the more positive the sentiment of its posts.

Evaluation and Validation

The previous sections demonstrate that neighborhoods with different income profiles exhibit clearly different textual traits. We now asses whether these differences are enough to predict whether posts are from a richer or poorer neighborhood. We show that commonly used models succeed in this prediction task.

Text can predict income differences. After splitting data into 70% train and 30% test sets, we use a Lasso regression to confirm that the text embeddings used in the Income and Text Features Section indeed differ depending on the income level. Table 4 shows the corresponding coefficient of determination (R^2) demonstrating that the text has a strong predictive power of the income of a neighborhood.

Prediction features. Encouraged by these findings, we de-

Country	Train R^2	Test R^2
USA	0.39	0.35
UK	0.69	0.59

Table 4: Determination coefficient (R^2) of a Lasso regression exclusively using sentence embeddings to predict the income of a neighborhood.

velop multiple features based on the analysis of the previous sections and investigate their predictive power. We first use UMAP for dimension reduction as prior work shows improvements in the performance (Zhang et al. 2022; Jimenez Villalonga 2021): we reduce 768-dimensional mean-pooled textual embeddings to 5 dimensions. For feature selection, we employ a Random Forest (RF) as this is the best-performing method (Chen et al. 2020). We compute the relative importance of each feature using the Gini index (Breiman et al. 2017). Table 5 presents the relative importance of each feature in our prediction tasks.

We observe that crime discussions are highly correlated with both income and inequality. This is particularly true for the non-violent crimes categories, which rank top in feature importance. We notice that the importance of these variables increases when we remove inequality from the features list while predicting income and vice versa.

Prediction Model Setup. Again we split our data into 70% train dataset and 30% test dataset for our models. We then train the regression models and use stratified 10-fold cross-validation as the evaluation approach. We implement Linear Regression (LR), Random Forest (RF), Support Vector Regressors (SVR), Multi-Layer Perceptrons-based Regressors (MLPR), Decision Trees, and Linear Regression (LR) models. We define two tasks: (*i*) predicting the annual median income of a neighborhood, and (*ii*) predicting the inequality (*i.e.* Atkinson index) of the vicinity of a neighborhood.
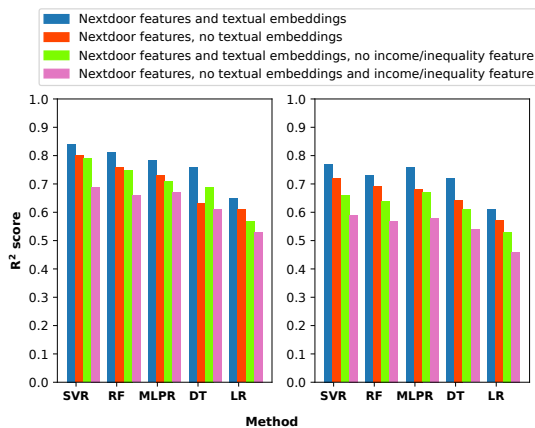
For each task, we train and test our models with and

Figure 9: Determination coefficient (R^2) for income and inequality predictions with Support Vector Regressor (SVR), Random Forest (RF), Multi-layer Perceptron Regressor (MLPR), Decision Trees (DT), and Linear Regression (LR).

without text features and with and without inequality (when predicting income) and income (when predicting inequality). We use the GridSearchCV method to optimize hyperparameters for each algorithm. Details of our best model parameters chosen by the GridSearchCV method are presented in Table 7 in the Appendix.

Model Evaluation. Our results show that the text posted online indeed reveals what is the income level and inequality of the area where they are posted. Figure 9 summarizes the performance of our models. The prediction of income and inequality with a combination of textual and other Nextdoor features (income best R^2: 0.841, inequality best R^2: 0.77) has the highest performance. We experiment training with fewer features and observe that even though the performance degrades, the R^2 still high goes up to 0.686 for income and 0.59 for inequality with the most restrictive set of features (all Nextdoor features, but no textual embeddings and income/inequality feature). The predicting power for inequality is slightly lower than for income (between 7 and 13% lower in the SVR model). This is natural as this prediction is inherently more difficult: a neighborhood in a highly unequal surrounding could be either a richer neighborhood surrounded by poorer ones, or the opposite, a poorer one surrounded by rich neighborhoods.

We are confident that our model is not overfitting our training data given the low Test Root Mean-Square Errors of 0.159 and 0.429 for income and inequality in the best-performing model (see Figures 11 and 12 in the Appendix).

Related Work

Much work has analyzed social networks from centralized social networking platforms such as Twitter and Reddit (Iqbal et al. 2022) to the increasingly popular Mastodon (Zia et al. 2023) and other Fediverse platforms (Bin Zia et al. 2022).

Previous research strove to infer the income of online users (Preoţiuc-Pietro, Lampos, and Aletras 2015; Aletras and Chamberlain 2018; Lampos et al. 2016; Abitbol, Karsai, and Fleury 2018). For instance, for users that mention their profession in Twitter, Preoctiuc et al. (Preoţiuc-Pietro et al. 2015; Aletras and Chamberlain 2018) approximate the income of those users with the median income of such profession. Then they predict the inferred user income with text features from their tweets (0.633 Pearson correlation). Differently, this work does not require heuristics: we use official statistics on the median income of the neighborhood where users reside. Moreover, our comprehensive data across USA and UK minimises possible sampling biases and successfully predicts neighborhood income ($R^2 = 0.841$).

Similarly to our work, research using mobility and location-based application data can map individuals to a neighborhood and its official median income (Moro et al. 2021; Tóth et al. 2021; Xu et al. 2019). Differently, this implies first inferring the residence based on the mobility patterns of a user. Additionally, there is limited online discourse that we could leverage for the purpose of this paper in mobility data and location-based apps such as Foursquare.

Feature	Feature Importance			
	Income	Inequality	Income, no inequality	Inequality, no income
Discussed-to-official crime (theft property damage) ratio	0.162	0.141	0.22	0.251
Discussed-to-official crime (drugs public order) ratio	0.138	0.125	0.181	0.192
Atkinson index	0.124	Target	N/A	Target
Median annual income	Target	0.120	Target	N/A
Mean-pooled (5 dimensions) sentence embeddings of posts	0.119	0.118	0.169	0.161
Registered Nextdoor users	0.112	0.093	0.143	0.136
Median readability score of posts	0.092	0.079	0.102	0.098
Median post length	0.078	0.061	0.098	0.091
Discussed-to-official crime (weapons violence) ratio	0.069	0.083	0.087	0.071

Table 5: Feature Importance (FI) of each feature in the Random Forest model to predict neighborhoods' income and income inequality with different feature sets. Note that "Target" refers to the explained variable.

Discussion: Limitations and Future Work

Bias. There is bias towards richer neighborhoods and official data only reports the median income of a neighborhood. Since we are unaware of the distribution of income within a neighboorhood, we cannot fully asses the magnitude of the bias: Nextdoor users might tend to be the richest within their neighborhoods. Therefore our results could refer to some extent to rich individuals living in neighborhoods of different incomes rather than to users with different income levels.

Data completeness. We aim to have a complete dataset for our future analysis. Currently, our data is incomplete for the USA (in 13 estates) and exclusively covers the 10 largest UK cities. This will allow us, to compare rural versus urban neighborhoods, for instance.

Topics. In our future work, we will analyze topics beyond crime. In this paper, we focused on crime because official statistics are rich and include the location of where a crime took place. Going forward we will study whether topics discussed differ across neighborhoods of different income levels. We also plan to focus again on topics for which there is geolocated data. In particular, we aim to look at how and whether politics manifest in the text posted by Nextdoor users and how politicians relate to the neighborhood they represent.

Income prediction. Finally, we want to explore further income prediction. Our future work will explore whether a general model can predict the income of individuals from their online discourse regardless of the platform where it is posted. We will experiment with models trained with Nextdoor data and predict the income level of users from other platforms (*e.g.* Twitter).

Conclusion

This paper provided the first large-scale analysis of Nextdoor. We showed that the online content generated by users reveals economic factors, including income and inequality. We collected 2.6 Million posts from 64,283 neighborhoods in the United States (USA) and 3,325 neighborhoods in the United Kingdom (UK) from an un-studied online social network, Nextdoor.

A unique feature of our Nextdoor dataset is that users and posts are associated to the neighborhood where they reside.

We then linked the neighborhood of the users with official socioeconomic (*e.g.* income, and crime) and showed that neighborhoods from different income levels post very different content. We found that the richest neighborhoods seem more sensitive about crime: crime is discussed more than in the poorest neighborhoods, even though the actual crime rates are higher in the latter. We also showed that the text posted by the richest neighborhoods seems to be more positive, with a consistently higher sentiment. We found that income inequality is also visible in online discourse and interacts with the income of the neighborhood. The richest neighborhoods with the most equal surroundings seemed to be more crime-sensitive and more positive, with posted text with higher sentiment than anyone else. The contrary was true for the poorest neighborhoods, where those with the most equal surroundings were less sensitive to crime, and their text had the lowest sentiment of all. Finally, we showed that the content generated by the users can predict their income and inequality levels. We trained multiple machine learning models with features extracted from Nextdoor and predicted the income ($R^2 = 0.841$) and inequality ($R^2 = 0.77$) of neighborhoods. Our work demonstrated that inferring socioeconomic factors from the text posted by users is feasible, presenting opportunities both for scientists and policymakers as well as for algorithmic surveillance.

Appendix

Official Crime Category	Aggregated Crime Category
Anti-social Behaviour	Drugs and Order
Drugs	
Public Order	
Bike Theft	Theft and Property Damage
Burglary	
Criminal Damage and Arson	
Robbery	
Shoplifting	
Theft from the Person	
Vehicle Theft	
Violent Crimes	Weapons and Violent Crimes
Murder	
Weapons	

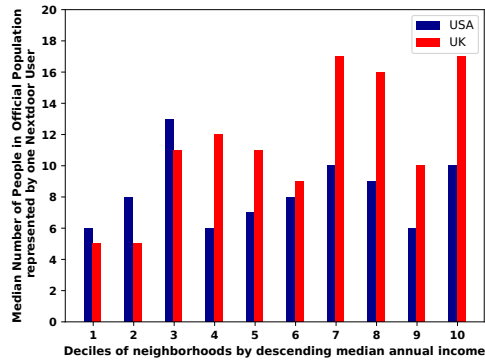Table 6: Crime categories from official crime databases.

Figure 10: Population-to-Nextdoor neighbor ratio over income deciles (from richest to poorest).
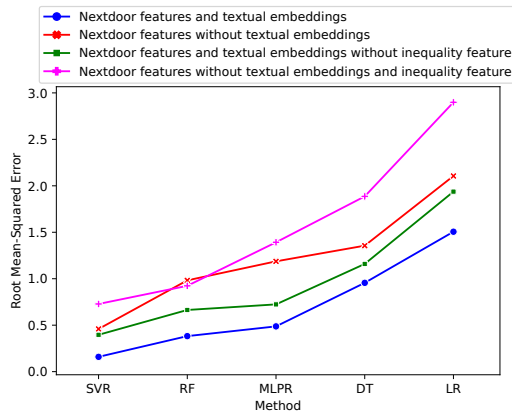

Figure 11: RMSE values for regression methods on Nextdoor and Official data features to predict income.
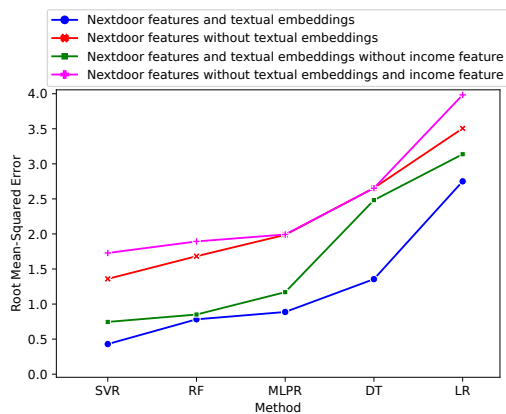

Figure 12: RMSE values for regression methods on Nextdoor data and Official data features to predict income inequality.

Model	Model Parameters
SVR	C:5, kernel: rbf, degree: 3, gamma: scale (income)
	C:10, kernel: rbf, degree: 2, gamma: scale (inequality)
	C:2, kernel: rbf, degree: 1, gamma: auto (income without inequality)
	C:3, kernel: rbf, degree: 1, gamma: scale (inequality without income)
Random Forest	max_depth: 3, max_features: sqrt, n_estimators: 20 (income)
	max_depth: 4, max_features: auto, n_estimators: 80 (inequality)
	max_depth: 5, max_features: auto, n_estimators: 150 (income without inequality)
	max_depth: 5, max_features: sqrt, n_estimators: 60 (inequality without income)
MLPR	hidden_layer_sizes': (120,80,40), max_iter: 5000, activation: relu, solver: adam, alpha: 0.01, learning_rate: adaptive (income)
	hidden_layer_sizes': (120,80,40), max_iter: 5000, activation: relu, solver: adam, alpha: 0.01, learning_rate: adaptive (inequality)
	hidden_layer_sizes': (150,100,50), max_iter: 5000, activation: tanh, solver: sgd, alpha: 0.05, learning_rate: adaptive (income without inequality)
	hidden_layer_sizes': (100,50,30), max_iter: 5000, activation: tanh, solver: sgd, alpha: 0.01, learning_rate: adaptive (inequality without income)
Decision Trees	splitter: best, max_depth: 7, min_leaf_sample: 0.4, max_leaf_node: 50, max_features: log2 (income)
	splitter: best, max_depth: 5, min_leaf_sample: 0.3, max_leaf_node: 40, max_features: log2 (inequality)
	splitter: best, max_depth: 7, min_leaf_sample: 0.1, max_leaf_node: 7, max_features: log2 (income without inequality)
	splitter: best, max_depth: 3, min_leaf_sample: 0.2, max_leaf_node: 40, max_features: auto (inequality without income)
Linear Regression	alpha=0.6, fit_intercept=True, max_iter=2000, tol=0.0005, selection='random' (income)
	alpha=0.5, fit_intercept=True, max_iter=2000, tol=0.0005, selection='random' (inequality)
	alpha=0.1, fit_intercept=True, max_iter=5000, tol=0.001, selection='random' (income without inequality)
	alpha=0.5, fit_intercept=True, max_iter=4000, tol=0.003, selection='random' (inequality without income)

Table 7: Best parameters for each model used to predict income and income inequality.

Ethical Statement

This research study has been approved by the Institutional Review Board (IRB) at the researchers' institution. The authors have no competing interests or funding that could undermine this research. We employ users' public post records from Nextdoor to study their conversations. Nextdoor data is public, as there is the expectation that strangers can view the posts (Townsend and Wallace 2016). Upon collection, we anonymize the data before use and store it in a secure silo. To prevent user identification, we aggregate our data and analyse at a neighborhood level. After aggregation, we discard any user-level information. Our work does not share or redistribute Nextdoor content, as per Nextdoor's Terms of Ser-

vice. Importantly, web crawling is legal for non-commercial research in the UK (IPO 2021) and the USA (TechCrunch 2022), where the data collection is performed.

From a **broader perspective**, our analysis demonstrates that predicting the income of users based on their online discourse is feasible. We believe this finding is an important contribution, particularly as this might further enable algorithmic surveillance (Zuboff 2015) by making easier to segment users based on their economic circumstances.

References

Abitbol, J. L.; Karsai, M.; and Fleury, E. 2018. Location, occupation, and semantics based socioeconomic status inference on twitter. In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, 1192–1199. IEEE.

Aggarwal, A.; Almeida, J.; and Kumaraguru, P. 2013. Detection of spam tipping behaviour on foursquare. In *Proceedings of the 22nd International Conference on World Wide Web*, 641–648.

Aggarwal, C. C.; Hinneburg, A.; and Keim, D. A. 2001. On the Surprising Behavior of Distance Metrics in High Dimensional Space. In Van den Bussche, J.; and Vianu, V., eds., *Database Theory — ICDT 2001*, 420–434. Berlin, Heidelberg: Springer Berlin Heidelberg.

Aletras, N.; and Chamberlain, B. P. 2018. Predicting twitter user socioeconomic attributes with network and language information. In *Proceedings of the 29th on Hypertext and Social Media*, 20–24.

Arora, S.; Liang, Y.; and Ma, T. 2017. A Simple but Tough-to-Beat Baseline for Sentence Embeddings. In *ICLR*.

Atkinson, A. B.; Micklewright, J.; and Micklewright, M. 1992. *Economic transformation in Eastern Europe and the distribution of income*. Cambridge University Press.

Bernstein, B. 1960. Language and social class. *The British journal of sociology*, 11(3): 271–276.

Bin Zia, H.; Raman, A.; Castro, I.; Hassan Anaobi, I.; De Cristofaro, E.; Sastry, N.; and Tyson, G. 2022. Toxicity in the decentralized web and the potential for model sharing. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 6(2): 1–25.

Breiman, L.; Friedman, J. H.; Olshen, R. A.; and Stone, C. J. 2017. *Classification and regression trees*. Routledge.

Census, U. 2022. *US Census Bureau Release*. https://www.census.gov/data.html. Accessed: 2022-07-10.

Chen, R.-C.; Dewi, C.; Huang, S.-W.; and Caraka, R. E. 2020. Selecting critical features for data classification based on machine learning methods. *Journal of Big Data*, 7(1): 1–26.

Chorley, M. J.; Rossi, L.; Tyson, G.; and Williams, M. J. 2016. Pub crawling at scale: tapping untappd to explore social drinking. In *Tenth International AAAI Conference on Web and Social Media*.

Cui, J.; Zhang, T.; Jaidka, K.; Pang, D.; Sherman, G.; Jakhetiya, V.; Ungar, L. H.; and Guntuku, S. C. 2022. Social Media Reveals Urban-Rural Differences in Stress across China. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, 114–124.

Dekker, K. 2007. Social capital, neighbourhood attachment and participation in distressed urban areas. A case study in The Hague and Utrecht, the Netherlands. *Housing Studies*, 22(3): 355–379.

Easterlin, R. A. 1974. Does economic growth improve the human lot? Some empirical evidence. In *Nations and households in economic growth*, 89–125. Elsevier.

FBI. 2022. *Federal Bureau of Investigation Crime Data Explorer*. https://cde.ucr.cjis.gov/LATEST/webapp/pages/home. Accessed: 2022-07-10.

GeoPy. 2022. *GeoPy*. https://geopy.readthedocs.io/en/stable/. Accessed: 2022-07-10.

Giorgi, S.; Guntuku, S. C.; Eichstaedt, J. C.; Pajot, C.; Schwartz, H. A.; and Ungar, L. H. 2021. Well-Being Depends on Social Comparison: Hierarchical Models of Twitter Language Suggest That Richer Neighbors Make You Less Happy. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, 1069–1074.

Hays, R. A.; and Kogl, A. M. 2007. Neighborhood attachment, social capital building, and political participation: A case study of low-and moderate-income residents of Waterloo, Iowa. *Journal of Urban Affairs*, 29(2): 181–205.

Huggingface. 2022. *all-mpnet-base-v2*. https://huggingface.co/sentence-transformers/all-mpnet-base-v2. Accessed: 2022-07-10.

Hutto, C.; and Gilbert, E. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, 216–225.

Inman, J. 1849. *Navigation and nautical astronomy, for the Use of British Seamen*. F. & J. Rivington.

IPO, U. 2021. *Exceptions to copyright*. https://www.gov.uk/guidance/exceptions-to-copyright#text-and-data-mining-for-non-commercial-research. Accessed: 2022-07-10.

Iqbal, W.; Arshad, M. H.; Tyson, G.; and Castro, I. 2022. Exploring Crowdsourced Content Moderation Through Lens of Reddit during COVID-19. In *Proceedings of the 17th Asian Internet Engineering Conference*, 26–35.

Jimenez Villalonga, F. 2021. Uncovering Correlations Between Two UMAP Hyperparameters and the Input Dataset. *KTH EECS Thesis Repository*, 1: 1–240.

Lampos, V.; Aletras, N.; Geyti, J. K.; Zou, B.; and Cox, I. J. 2016. Inferring the socioeconomic status of social media users based on behaviour and language. In *European conference on information retrieval*, 689–695. Springer.

Lee Rodgers, J.; and Nicewander, W. A. 1988. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1): 59–66.

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

McInnes, L.; Healy, J.; Saul, N.; and Großberger, L. 2018. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29).

Moro, E.; Calacci, D.; Dong, X.; and Pentland, A. 2021. Mobility patterns are associated with experienced income segregation in large US cities. *Nature communications*, 12(1): 1–10.

Nextdoor. 2021. *Nextdoor can bring communities together*. https://inews.co.uk/inews-lifestyle/nextdoor-app-neighbourhood-online-groups-780656. Accessed: 2022-07-10.

Nextdoor. 2022. *Nextdoor Neighborhoods*. https://nextdoor.com/find-neighborhood/. Accessed: 2022-07-10.

ONS, U. 2018. *Lower Layer Super Output Area (2001) to Lower Layer Super Output Area (2011) to Local Authority District (2011) Lookup in England and Wales*. https://www.data.gov.uk/dataset/afc2ed54-f1c5-44f3-b8bb-6454eb0153d0/lower-layer-super-output-area-2001-to-lower-layer-super-output-area-2011-to-local-authority-district-2011-lookup-in-england-and-wales. Accessed: 2022-07-10.

Police, U. M. 2022. *Stats and data*. https://www.met.police.uk/sd/stats-and-data/. Accessed: 2022-07-10.

Preoţiuc-Pietro, D.; Lampos, V.; and Aletras, N. 2015. An analysis of the user occupational class through Twitter content. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1754–1764.

Preoţiuc-Pietro, D.; Volkova, S.; Lampos, V.; Bachrach, Y.; and Aletras, N. 2015. Studying user income through language, behaviour and affect in social media. *PloS one*, 10(9): e0138717.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

TechCrunch. 2022. *Web scraping is legal, US appeals court reaffirms*. https://techcrunch.com/2022/04/18/web-scraping-legal-court/. Accessed: 2022-07-10.

Tóth, G.; Wachs, J.; Di Clemente, R.; Jakobi, Á.; Ságvári, B.; Kertész, J.; and Lengyel, B. 2021. Inequality is rising where social network segregation interacts with urban topology. *Nature communications*, 12(1): 1–9.

Townsend, L.; and Wallace, C. 2016. Social media research: A guide to ethics. *University of Aberdeen*, 1: 16.

UK, O. 2022. *Estimates of the population for the UK, England, Wales, Scotland and Northern Ireland*. https://www.ons.gov.uk/peoplepopulationandcommunit/populationandmigration/populationestimates/dataset/populationestimatesforukenglandandwalesscotlandand//northernireland. Accessed: 2022-07-10.

Wu, F. 2012. Neighborhood attachment, social participation, and willingness to stay in China's low-income communities. *Urban Affairs Review*, 48(4): 547–570.

Xu, Y.; Belyi, A.; Santi, P.; and Ratti, C. 2019. Quantifying segregation in an integrated urban physical-social space. *Journal of the Royal Society Interface*, 16(160): 20190536.

Zhang, Z.; Fang, M.; Chen, L.; and Namazi-Rad, M.-R. 2022. Is Neural Topic Modelling Better than Clustering? An Empirical Study on Clustering with Contextual Embeddings for Topics. *arXiv preprint arXiv:2204.09874*.

Zia, H. B.; He, J.; Raman, A.; Castro, I.; Sastry, N.; and Tyson, G. 2023. Flocking to mastodon: Tracking the great twitter migration. *arXiv preprint arXiv:2302.14294*.

Zuboff, S. 2015. Big other: surveillance capitalism and the prospects of an information civilization. *Journal of information technology*, 30(1): 75–89.