Exploring Online Manifestations of Real-World Inequalities

Waleed Iqbal
w.iqbal@qmul.ac.uk
Queen Mary University of London

Gareth Tyson
gtyson@ust.hk
Hong Kong University of Science and Technology

Vahid Ghafouri
vahid.ghafouri@imdea.org
IMDEA Networks

Guillermo Suarez-Tangil
guillermo.suarez-tangil@imdea.org
IMDEA Networks

Ignacio Castro
i.castro@qmul.ac.uk
Queen Mary University of London

ABSTRACT

Socioeconomic gaps, particularly income inequality, affect crime and public opinion. Although official data sources can identify these patterns of income-based social disparity, a fundamental question remains: Can similar social inequalities be found using abundant online user activity? We explore two sub-questions. (*i*) How does a neighbourhood's income affect crime discussion in a geographical neighbourhood? (*ii*) Can user-generated data predict a neighbourhood's income? To answer these questions, we collect 2.5 million Nextdoor posts from 67608 USA and UK neighbourhoods between November 2020 and September 2021. We use official USA and UK data sources for crime and income information.

CCS CONCEPTS

• **Social and professional topics** → **Geographic characteristics**;

ACM Reference Format:
Waleed Iqbal, Gareth Tyson, Vahid Ghafouri, Guillermo Suarez-Tangil, and Ignacio Castro. 2022. Exploring Online Manifestations of Real-World Inequalities. In *Proceedings of the 22nd ACM Internet Measurement Conference (IMC '22), October 25–27, 2022, Nice, France.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3517745.3563027

1 DATA COLLECTION AND METHODOLOGY

1.1 Nextdoor Data

NextDoor is a location-based social network themed around small neighbourhoods of residents. It allows users to post messages, which are seen by all users in their neighbours (often about hyper-local matters, e.g., littering). We have collected 2.2M posts from 64,283 neighbourhoods in the USA and 252K posts from 3325 neighbourhoods in the ten most populated cities in the UK. Our data covers the period from November 2020–September 2021. In this study, we refer to each registered user on Nextdoor as a neighbour.

1.2 Data Augmentation

Official Data Sources. We further gather datasets from official sources to augment our Nextdoor data. We use the USA Census data repository [5], Data USA API [1], UK's Office of National Statistics [2], and UK Metropolitan Police [3] data to get the crimes per 10000 people, population and income data of all USA and UK cities.

Identification of Crime-Related Posts. To compare crime discussion on Nextdoor with official statistics, we categorise all posts with official crime categories. Specifically, we tag if user posts discuss those crime using the pre-trained *msmarco-distilbert-base-v4* Sentence-BERT (S-BERT) model [4]. This model is used for perform similarity search between small string such as user search in social media posts. It returns a cosine similarity score between crime name and user post. We observe different cosine similarity scores by manually inspecting output of a random sample of 1000 posts. Our S-BERT labelled data, and manually tagged data agree with a Cohen's Kappa Score of 0.95 on cosine similarity score threshold of 0.7. We obtain 124763 UK crime posts and 543459 USA crime posts.

2 DATA CHARACTERISATION

Neighbourhood and Neighbour Activity. We start by presenting a brief overview of our data to understand better how representative our data is compared to official population statistics. 60%, 70%, 61%, and 76%, respectively, of the UK's total residents, posts, neighbourhoods, and neighbours are concentrated in London. The USA population is less skewed, and so is our Nextdoor data: the 20 most populated states (40% of all states) account for 70% of the posts, 72% of the neighbourhoods, and 69% of the neighbours. Furthermore, our data shows high positive correlation with official statistics of population.

Ratio of Registered Neighbours and Official Population. We rank neighbourhoods by median income from richest (Q1) to poorest (Q10). The number of registered users in the USA, UK, London, and New York City decreases as the median income decreases. We find fewer registered users in London and NYC's first quantile. We hypothesise these places have more daytime and tourist residents. We map Nextdoor neighbours with people in official population in USA, UK, NYC, and London and see how many people in population are represented by a single Nextdoor neighbour. We observe that in London and New York City is 5 to 11 people are mapped to one Nextdoor neighbour; in the USA, it's 5 to 12 people: First-quantile Nextdoor user mapping in the UK is four persons and rises dramatically to 20 people per Nextdoor neighbour.

IMC '22, October 25–27, 2022, Nice, France

ACM ISBN 978-1-4503-9259-4/22/10…$15.00
https://doi.org/10.1145/3517745.3563027

3 IMPACT OF INCOME ON ONLINE USER CRIME DISCUSSION

Income Affects Crime Discussion in Online Social Interactions: Figure 1 displays the crime rate per 10000 officially reported, as well as the rate discussed in neighbour posts over Nextdoor in the USA. We breakdown crimes into three categories: (*i*) Drugs and Order; (*ii*) Theft and Property Damage; and (*iii*) Weapons and Violent Crimes. We observe that the crime rate rises with decreasing income levels. We observe similar trends in crime rates in the UK, NYC, and London. We find that Non-violent crimes are discussed more on Nextdoor than firearms and violent crimes in all neighbourhoods.

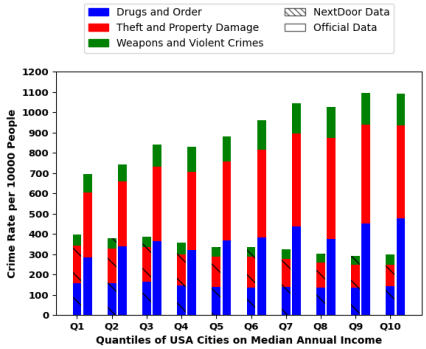

Figure 1: Distribution of official crime rate and crime discussion rate over Nextdoor per 10000 people in USA.

Our analyses from Figure 2 also show that the official crime rate and crime discussion are affected by income on Nextdoor. All violent crimes show a moderate negative correlation.

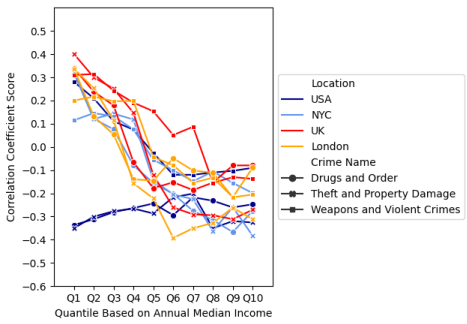

Figure 2: Correlation between crimes per 10000 people reported officially and discussed over Nextdoor.

4 PREDICTING NEIGHBOURHOOD INCOME.

Given the high Correlation of income with socioeconomic factors and Nextdoor, we hypothesise that we can predict a neighbourhood's income.

Features for Prediction Model. The target feature of our model is income of a neighbourhood. We use several input features, including neighbours, Nextdoor crime discussion, official crimes per 10000 people, the area's population, average post length, the average number of comments, median Age of population, and officially reported households. We experiment with several models: linear regression, random forest, SVR, MLP, and decision trees with 10-fold cross-validation. After training, we employ the coefficient of determination (R^2) value and root mean squared error to assess the model.

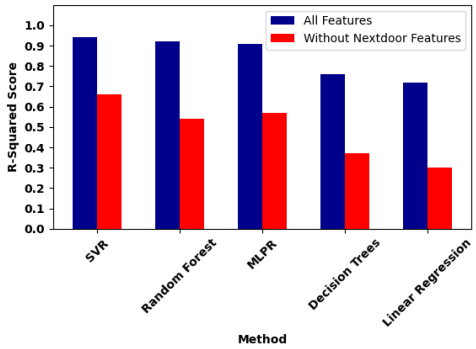

Figure 3: Evaluation for regression methods on Nextdoor data and Official data features.

Model Evaluation. The performance of our models is presented in Figure 3. Predictions using the NextDoor data obtain high accuracy. Interestingly, we also see that that official statistics without online user activity are insufficient to predict a neighbourhood's income.

5 CONCLUSION AND FUTURE WORK

We examined how crime reporting and conversation vary by neighbourhood. We found that crime is connected with a neighbourhood's income and influences crime conversation. We also tried to anticipate a neighbourhood's income using user conversation, meta-data, and official statistics. We found that combining official data with Nextdoor data can correctly forecast a neighbourhood's income. In future work, we will examine (*i*) how income influences a neighbourhood's sentiment, and (*ii*) whether online user activity can solely predict income.

6 ACKNOWLEDGEMENTS

This work is supported by EPSRC REPHRAIN "Moderation in Decentralised Social Networks" under grants EP/S033564/1 and EP/W032473/1, and the "Ramon y Cajal" Fellowship RYC-2020-029401-I.

REFERENCES

[1] DataUSA. 2022. *Data USA API.* (2022). https://datausa.io/about/api/
[2] UK ONS. 2022. *Office of National Statistics, UK.* (2022). https://www.ons.gov.uk/searchdata?q=explorable%20datasets
[3] UK Metropolitan Police. 2022. *UK Metropolitan Police.* (2022). https://www.met.police.uk/sd/stats-and-data/
[4] Nils Reimers and Iryna Gurevych. 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. https://arxiv.org/abs/1908.10084
[5] UScensus. 2022. *US Census Bureau Release.* (2022). https://www.census.gov/data.html