The Human Labor Behind AI: Exploring the Geography of Invisible Data Work

Mykyta Solonko

Yale University

Yale

Introduction

The rise of microtask labor, in which individuals complete small tasks via digital platforms for minimal pay, has significantly contributed to AI development through data provision for machine learning models. Despite their essential role in tasks such as image labeling and text transcription, the human geography of this often-overlooked workforce remains largely uncharted. A deeper understanding of the distribution of microtask workers and the factors influencing it is crucial for grasping the implications of the invisible gig economy on the future of work. In this study, we use web traffic and macroeconomic data to reveal the relationships between COVID cases, GDP, population, and microtask labor traffic. Our findings shed light on the geographies of both microtask clients and workers, as well as the criteria that attract microtask platforms to specific countries.

Objectives

In this study, we aim to better understand the invisible microtask population by focusing on the human geography and key factors shaping the landscape of microtask labor. Our main objectives include:

- Investigating the geographic distribution of microtask workers and comparing it to the distribution of clients.
- Developing a model to predict microtask labor traffic based on relevant macroeconomic variables.
- Identifying potential country targets for microtask platforms and analyzing their common characteristics.

By addressing these goals, we hope to contribute valuable insights to the ongoing discourse surrounding the future of work and provide a deeper understanding of the microtask labor workforce.

Methodology

For traffic data, we will be using traffic aggregators Similarweb and Semrush, which provide us with country-specific traffic data for relevant platforms. Similarweb provides us with data aggregated for the last 3 months while Semrush gives us annual as well as monthly traffic data from as early as 2017. Here is the list of the platforms that are included in our dataset:

2Captcha	Appen	Clickworkers	Fiverr
Guru	Hive	Isahit	Kolotibablo
Microworkers	Amazon MTurk	Microsoft (UHRS)	Premise
Google (Raterhub)	Scale AI	Superannotate	Teemwork
Telus	Toloka	Upwork	Wirk

Table 1. Microtask Labor Companies

In order to find which countries' labor is targeted by such platforms as well as for the modeling section, we rely on various macroeconomic data provided by the World Bank, United Nations, and other sources. This data includes GDP information, daily COVID cases, life expectancy, human development index, and others.

Worker vs. Client Distribution

Some platforms in our dataset had separate subdomains for clients and workers, allowing us to see the differences in geographic distribution. These maps highlight the differences.

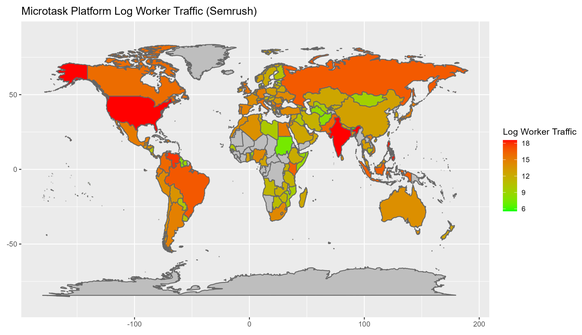

Figure 1. Worker Distribution

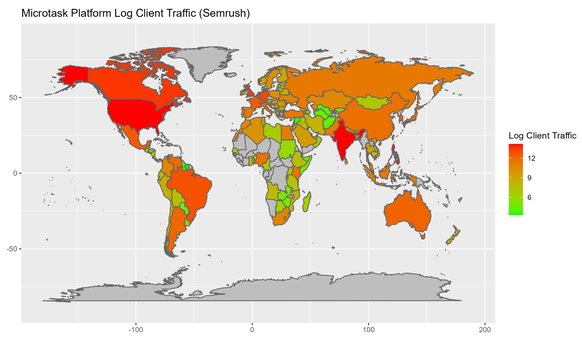

Figure 2. Client Distribution

While the worker and client distributions are related, clients are more likely to be from places like Mexico, Canada, the European Union, China, and Australia. India, Brazil, Russia, and Venezuela are most represented from the worker side.

Traffic Forecasting Model

We built a linear model forecasting traffic in the next year using this year's traffic and other macroeconomic variables like GDP and COVID cases. The following model was found:

Predictor	Estimate	Std. Error	t-value	p-value
Intercept	-1.011	0.378	-2.673	0.00756**
log_traffic	0.759	0.010	79.732	<2e-16***
log_cases	0.063	0.014	4.443	9.12e-06***
log_gdp	0.040	0.023	1.779	0.07538.
log_population	0.088	0.022	3.915	9.18e-05***

Table 2. Linear Model Results

Higher COVID cases may have pushed more workers into digital work. A more populous country has more workers for traffic to eventually grow from, while a higher GDP may be related to better internet infrastructure facilitating such work in the first place.

Potential Country Targets

By taking a look at traffic peaks and then declines across countries for the same platform, we could make educated guesses on which countries' workers are being intentionally targeted by these platforms with short-term incentives. The following chart shows how those countries compare to non-targeted countries:

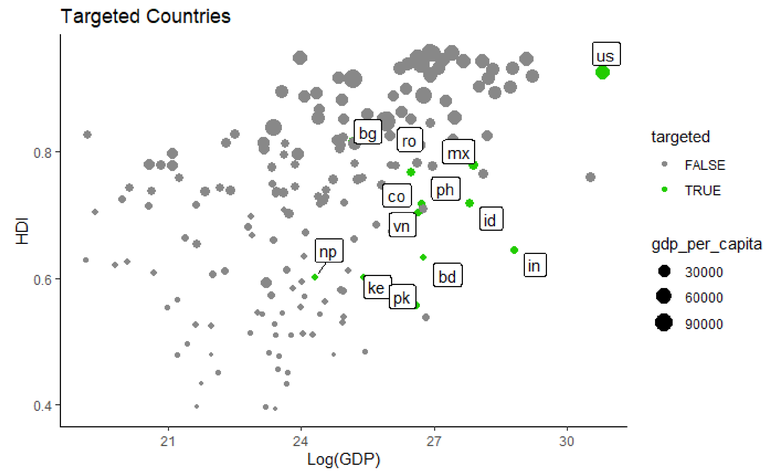

Figure 3. Targeted Countries

Excluding the United States, these countries share the following characteristics:

- Human Development Index between 0.55 and 0.83
- 2021 GDP between $1.45B and $3.2T
- GDP Per Capita between $2900 and $24500

Only 43% of the countries in the world fit this criteria, giving us insight into which other countries could be targeted next by these platforms.

Conclusion

Our analysis of the microtask gig economy sheds light on the geographical distribution of both workers and clients in this largely invisible population. Workers targeted by microtask platforms often come from countries with medium HDI, moderately high GDP, and low-to-medium GDP per capita. Our linear model revealed that current traffic, COVID cases, GDP, and population are significant predictors of future microtask labor traffic.

Future work could investigate factors behind the annual traffic fluctuations and incorporate inflation and unemployment rates into the analysis. Enhancements to the current study could involve accounting for VPN usage and finding ways to separate worker and client data for platforms without specific subdomains for a more granular understanding of the microtask labor landscape.