Datasheet for 'global findex'

Yizhuo Li

November 30, 2024

Extract of the questions from Gebru et al. (2021).

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*
    - The dataset was created to analyze global financial inclusion across 139 countries, with a focus on understanding access to financial services, usage patterns, and barriers. It fills the gap of comprehensive, structured data on financial behaviors, services, and inclusion globally.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - The dataset was created by the Development Research Group, Finance, and Private Sector Development Unit at the World Bank.
3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*
    - The dataset was funded by the World Bank, but no specific grant details are provided.
4. *Any other comments?*
    - The dataset is valuable for policymakers, financial institutions, and researchers, helping them track and improve global financial inclusivity.

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

- The instances represent individual survey responses from people across 139 countries. These responses include data on financial behaviors, access to banking services, and financial literacy.

2. *How many instances are there in total (of each type, if appropriate)?*

   - The dataset contains 143,887 instances, with 128 variables representing a broad sample from various countries.

3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*

   - The dataset is a sample of a larger set, focused on individuals from countries that participated in the survey. The sample is geographically representative across 139 countries.

4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*

   - The data consists of survey responses, which include raw data such as age, income, access to bank accounts, mobile phone ownership, and derived features like financial literacy.

5. *Is there a label or target associated with each instance? If so, please provide a description.*

   - There is no specific target variable. The dataset is used to explore financial inclusion outcomes, but no explicit label is attached.

6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*

   - Missing information may occur due to incomplete responses or countries not participating in certain aspects of the survey. These gaps may be due to regional differences in data availability.

7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

   - Relationships are made explicit through geographic and demographic linkages, allowing for analysis by country, income group, age group, and other demographic variables.

8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

    - Specific data splits are not provided, but researchers might commonly split data by countries or income groups for validation and testing purposes.

9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

    - Some noise may exist due to inconsistencies in data reporting across countries or demographic groups, as survey collection methods vary globally.

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

    - The dataset is self-contained. However, it may rely on national surveys or other data sources for specific countries. There is no guarantee of long-term availability of these external resources.

11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*

    - No, the dataset does not contain confidential data. It includes aggregated survey data with no personally identifiable information.

12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

    - No, the dataset does not contain any such content. It focuses on financial behaviors and access to services.

13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

    - Yes, the dataset identifies sub-populations based on demographic factors such as age, gender, income, and country of residence. The dataset is distributed across various regions and income levels.

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*

    - No, individuals cannot be identified directly or indirectly from the dataset as it is anonymized.

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

    - No, the dataset does not contain sensitive data such as race, health data, or criminal history, but it includes financial behaviors and access to financial services, which could be sensitive.

16. *Any other comments?*

    - The dataset is useful for large-scale analyses of financial inclusion and global financial behaviors but should be treated with care when drawing conclusions about specific individuals or groups

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

    - The data was acquired through direct survey responses from individuals across 139 countries. Data was self-reported and validated through the survey design and methodology.

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*

    - The data was collected via surveys conducted by national agencies and international survey teams, using standardized survey instruments across participating countries.

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

    - The dataset uses probabilistic sampling to ensure that it is representative of the population in each country.

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

    - Data was collected by professional survey teams and national agencies. Specific compensation details are not provided.

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*

    - The data was collected between 2021 and 2023, reflecting the most recent trends in financial inclusion.

6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

    - Ethical review processes were likely conducted by the World Bank, but specific review outcomes or documentation are not publicly provided.

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

    - The data was collected directly from survey participants across multiple countries.

8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

    - Individuals were notified through the survey process, but specific notifications are not detailed.

9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

    - Consent was likely obtained through the survey process, although the specific language used is not provided.

10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

    - There is no specific mechanism described for revoking consent, as the dataset focuses on anonymized survey data.

11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

    - A data protection impact analysis was likely conducted as part of the World Bank's data governance procedures, though specific details are not publicly provided.

12. *Any other comments?*

    - The data collection process adheres to international standards for survey data collection, ensuring broad participation and representative sampling.

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*

    - Yes, the data was cleaned and processed to standardize responses and ensure comparability across countries and regions.

2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*

    - It is unclear if raw data is available publicly; only the cleaned data is generally provided.

3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*

    - Specific software used for preprocessing is not mentioned.

4. *Any other comments?*

    - The dataset appears to be processed with the goal of making it suitable for broad, cross-country financial inclusion analysis.

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*

    - Yes, it has been used to analyze global financial inclusion trends and to develop policies for improving access to financial services.

2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*

- The dataset is hosted on the World Bank's microdata repository, and relevant papers can be found there.

3. *What (other) tasks could the dataset be used for?*

   - The dataset could be used for research on financial behavior, economic development, and access to banking services across different demographic groups.

4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*

   - Researchers should be cautious when generalizing about individual financial behaviors, as the data is aggregate and anonymized. Misuse could lead to biases in financial decision-making or policy formulation.

5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*

   - The dataset should not be used for individual-level predictions or to make conclusions about specific people without appropriate context.

6. *Any other comments?*

   - The dataset is a powerful tool for understanding global financial inclusion but should be used with a critical understanding of its limitations and the potential for biases.

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*

   - Yes, the dataset is publicly available and distributed to third parties, including researchers, policymakers, and financial institutions, to promote financial inclusion globally.

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*

   - The dataset is distributed through the World Bank's Microdata Library website, where users can download it directly. It has a Digital Object Identifier (DOI): https://doi.org/10.48529/jq97-aj70.

3. *When will the dataset be distributed?*

- The dataset was first released in October 2022 and has been updated periodically since then. The latest update was in May 2023.

4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

    - The dataset is distributed under the World Bank's terms of use, which allow for free use, reproduction, and distribution with proper attribution. The terms of use can be found here: https://microdata.worldbank.org/index.php/terms-of-use.

5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*

    - No, there are no additional IP-based or other restrictions imposed by third parties on the data.

6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*

    - No, there are no export controls or other regulatory restrictions applying to the dataset.

7. *Any other comments?*

    - The dataset is a valuable resource for understanding global financial inclusion trends and is freely accessible to support research and policy development.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*

    - The World Bank's Development Research Group, Finance and Private Sector Development Unit, is responsible for supporting, hosting, and maintaining the dataset.

2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*

    - The dataset can be accessed and inquiries can be directed to the World Bank's Microdata Library: https://microdata.worldbank.org/index.php/contact.

3. *Is there an erratum? If so, please provide a link or other access point.*

    - As of now, there are no known errata for this dataset.

4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*

    - Yes, the dataset is periodically updated to include new data and correct any errors. Updates are communicated through the World Bank's Microdata Library website and associated publications.

5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*

    - The dataset contains anonymized data and does not include personally identifiable information, so retention limits are not applicable.

6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*

    - Older versions of the dataset are still accessible for historical comparison. Users are informed of the latest version through the World Bank's Microdata Library website.

7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*

    - While the dataset itself is not open for direct contributions, users can contact the World Bank with suggestions or data that may enhance future editions. Contributions are reviewed and, if appropriate, incorporated into subsequent releases.

8. *Any other comments?*

    - The World Bank encourages the use of this dataset for research and policy development to promote financial inclusion worldwide.

References

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92.