

DHS Final Report Senegal 1986*

Nicholas Bosco

30 April 2022

Abstract

An analysis of the data was collected by the authors of the Senegal DHS final report in 1986. This analysis of a specific datasheet I had created to test the data extracted from the report. The datasheet looks at the percentage of women in a polygamous relationships.

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 enable insight of the population trends of Polygamy in Senegal. Being in the western part of the world, I would hear about Polygamy being practiced in some parts of West Africa but never really knew how prominent polygamy was. Though this data was collected in the 1980s so it may be a bit outdated but I think it is all the more reason to look, to possible give insight to the trends of polygamy in the past.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
 - The data itself was created by “Ndiaye, Salif, Ibrahima Sarr, and Mohamed Ayad. 1988. 1986 Senegal Demographic and Health Survey. Dakar, Senegal: Ministry of Economy and Finance Direction de la Statistique Division des Enquêtes et de la Démographie/Senegal and the Institute for Resource Development/Westinghouse”(Sen1986?). This was from the recommended citation from the website SENDHS1986
 - But I, Nicholas Bosco who is a part of the “World Becomes Data Research School” had been given this task to create the dataset.
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.*
 - No one had funded me to generate the dataset.
4. *Any other comments?*
 - NA

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.*

*Code and data are available at: <https://github.com/bosconic/SoLongPolygamy>

- Within the data itself it just deals with people and their willingness to share an intimate part of their lives in order to be enumerated. So the enumerator who was collecting said information had a lot of responsibility. Also for the specific data sheet here it deals with percentages rather than number of people.
2. *How many instances are there in total (of each type, if appropriate)?*
 - TBD
 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 data attempts to capture the whole country of Senegal, however within the data itself it is a sample size as it captures only a large percentage of the country. Varying from different demographics woman identify with.
 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.*
 - NA
 5. *Is there a label or target associated with each instance? If so, please provide a description.*
 - NA
 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.*
 - Yes, Within the data there is there was a comment made about some women who identified within a certain age range who posses one of the desired socio-demographics might not have been significant enough to be mentioned as a percentage, or that there were very few people who fit into a specific category matching the age and their socio-demographic so it was hard to generate a percentage that is a good representation of the country’s woman population.
 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.*
 - TBD
 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.*
 - TBD
 9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*
 - There are a few errors noted when looking at the data, firstly there can be a possiblity of double counting women if you try to specify the parameters of the question. For example the socio-demographics are divided according to Milieu, Region, Ethnic Group, Instruction and Literacy. But what if i wanted to know about the percentage of women who are in a polygamous relationship from who identify as Wolof, who can read and had a secondary education. The data does not account for that kind of accuracy.
 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.*

- In order to access the translated data, you would have to have access to a translator software. If you prefer the original french data and you would like to electronically use it at your will, you would need a tool that is able to run OCR on the document in order to import the document.
 - The links I had used for the data are as follows: SENDHS1986,R programming language, and Microsoft Azure
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.*
 - TBD
 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.*
 - TBD
 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.*
 - TBD
 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.*
 - TBD
 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.*
 - TBD
 16. *Any other comments?*
 - TBD

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

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

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

- TBD
4. *Any other comments?*
- TBD

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
 - TBD
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.*
 - TBD
3. *What (other) tasks could the dataset be used for?*
 - TBD
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?*
 - TBD
5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
 - TBD
6. *Any other comments?*
 - TBD

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

- TBD
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.*
 - TBD
 7. *Any other comments?*
 - TBD

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
 - TBD
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
 - TBD
3. *Is there an erratum? If so, please provide a link or other access point.*
 - TBD
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)?*
 - TBD
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.*
 - TBD
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.*
 - TBD
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.*
 - TBD
8. *Any other comments?*
 - TBD

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.