Lacuna PII Multilingual Text Datasheet

The Lacuna PII text data sheet was created by Makerere Artificial Intelligence Lab in collaboration with Marconi lab and Clear Global org created by a group of researchers from the Makerere Artificial Intelligence Lab, Marconi Lab in Makerere University in Uganda and Clear Global organization in Nigeria. We followed the datasheet for the dataset framework created by (Gebru et al. 2021) to create this datasheet.

Motivation	
For what purpose was the dataset created	The text we collected was used to create labels for Pesron Identifiable Information in Luganda and Lumasaba in Uganda and Hausa and Kanuri in Nigeria.
Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?	This dataset was created by a group of researchers from the Makerere Artificial Intelligence Lab, Marconi Lab and Clear Global Organization
Who funded the creation of the dataset?	This work was done with funding from the Lacuna fund
Data format	The text data is in CSV format.

Composition	
What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?	Documents, people and countries
How many instances are there in total (of each type, if appropriate)?	4000 Luganda sentences, 5000 Lumasaba sentences, 3000 Kanuri sentences, 3000 Hausa sentences
Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?	The dataset contains all possible instances of Person Identifiable Information including Person, Location, Date, Organisation, Political/national affiliation and User/Government ID
What data does each instance consist of?	The dataset contains sentences in each language that have specifically identifiable information entities including Person, Location, Date, Organisation, Political/national affiliation and User/Government ID
Is any information missing from individual instances?	None
Are relationships between individual instances made explicit?	There are no relationships between instances made explicit in the dataset.
Are there recommended data splits (e.g., training, development/validation, testing)?	No recommendations
Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.	No
Is the dataset self-contained, or does it link to or otherwise rely on external resources?	N/A
Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)?	No

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?	No
Does the dataset contain data that might be considered sensitive in any way (e.g., 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)?	Yes

Collection Process	
How was the data associated with each instance acquired?	Text data was collected from various sources, including; online news articles, and crowd-sourcing. Websites and online news articles from which data was collected; MasakhaNER data, Wikipedia and Coast data
What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)?	Text data -Luganda and Lumasaba text data was curated by Makerere AI researchers while the Hausa and Kanuri text data was curated by researchers from Clear Global Organization. All the text data was also validated by linguistic consultants such that the data has PII entities and is grammatically correct with regards to the language rules and lexicon .
Over what time frame was the data collected?	Two months in 2023 for collection, two months in 2023 for validation
Were any ethical review processes conducted (e.g., by an institutional review board)?	No.

Preprocessing/cleaning/labeling	
Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?	The data was annotated and labels of personal identifiable information entities such as including Person, Location, Date, Organisation, Political/national affiliation and User/Government ID.
Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?	The raw unprocessed data is stored locally on backup data storage owned by the Marconi Lab and Makerere Artificial Intelligence Lab
Is the software that was used to preprocess/clean/label the data available?	Yes

Uses	
Has the dataset been used for any tasks already?	We have used the dataset to identify and annotate PII entities
Is there a repository that links to any or all papers or systems that use the dataset?	No
What (other) tasks could the dataset be used for?	1.Named Entity Recognition (NER) for PII entities in multiple domains 2.Text Classification; 3.Privacy-Preserving Data Analysis and Research 4.Language Modeling 5.Machine Translation 6.Linguistic Research

Is there anything about the composition of the dataset or the way it was collected and pre-processed/cleaned/labeled that might impact future uses?	None
Are there tasks for which the dataset should not be used?	Not at the moment

Distribution	
Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created	Yes
How will the dataset be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?	The text dataset is available and can be accesssed at https://doi.org/10.7910/DVN/CGHWZE
When will the dataset be distributed?	TBD
Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?	TBD
Have any third parties imposed IP-based or other restrictions on the data associated with the instances?	No

Maintenance	
Who will be supporting/hosting/maintaining the dataset?	The Makerere Artificial Intelligence Lab will support, host, and maintain the dataset.
How can the owner/curator/manager of the dataset be contacted (e.g., email address)?	Email address: themarconilab@gmail.com
Is there an erratum?	No
Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?	All updates to the dataset will be documented and communicated by the Marconi Lab.
Will older versions of the dataset continue to be supported/hosted/maintained?	Yes
If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?	Contact can be made through the email address: themarconilab@gmail.com

References

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. "Datasheets for datasets." Communications of the ACM 64, no. 12 (2021): 86-92.