Data statement for Lacuna PII Multilingual Text data

The Dataset is a collection of 4000 Luganda , 5000 Lumasaaba, 3002 Hausa and 3005 Kanuri PII sentences. The dataset can be downloaded from https://doi.org/10.7910/DVN/CGHWZE

1 CURATION RATIONALE

Luganda text data that contains person identifiable information curated by Makerere AI Lab researchers. Text data was collected from various sources including; Masakhaner project, Dembe FM, Galaxy FM, Ssegwanga newspaper, Gambuuze newspaper. The Luganda text was pre-checked and then validated, such that the sentences have person identifiable information, had grammatically correct Luganda. The Hausa and Kanuri text data was collected through Clear global organization from some open source available datasets on Hugging Face. Hausa text data was obtained through sources like VOA news and Kanuri was obtained corpus created by CLEAR Global.

2 LANGUAGE VARIETY

The dataset includes text in four languages:

Luganda (LG): A Bantu language widely spoken across Uganda, particularly in the central region. Sentences were collected from various documentation and websites, ensuring they contain personally identifiable information and are grammatically correct.

Lumasaaba (LU): A Bantu language spoken primarily by the Luhya people in Eastern Uganda and Kenya. Lumasaaba sentences in the dataset contain PII and were collected from various trusted experts and sources, ensuring linguistic accuracy and relevance.

Hausa (HA): A Chadic language widely spoken in West Africa, particularly in Nigeria, Niger, and Ghana. The Hausa text data was obtained from sources such as VOA News, ensuring the inclusion of personally identifiable information and maintaining linguistic integrity.

Kanuri (KR): A Nilo-Saharan language spoken primarily in northeastern Nigeria, southeastern Niger, and northern Cameroon. The Kanuri text data was sourced from a corpus created by CLEAR Global, ensuring it includes personally identifiable information and is representative of the language.

The collected sentences for all languages have been curated to include personally identifiable information and are validated for grammatical correctness and authenticity.

3 SPEAKER DEMOGRAPHIC

N/A

4 SPEECH SITUATION

N/A

5 TEXT CHARACTERISTICS

The sentences that were used for voice contribution were obtained from various sources. These included:

- LUGANDA
  - MasakhaNER project

  - Dembe FM
  - Galaxy FM
  - Ssegwanga newspaper
  - Gambuuze newspaper
- LUMASAABA
  - (sources To be confirmed)
- HAUSA
  - VOA News
  - Open-source datasets available on Hugging Face (collected through CLEAR Global)
- KANURI
  - Corpus created by CLEAR Global
  - Open-source datasets available on Hugging Face (collected through CLEAR Global)

Sentences from all these various sources were written using the same language Orthography per language but in various styles, with PII entities as our focus. Other major aspects emphasized during the sentence collection process were checking for accuracy and sentence completeness and checking for words or phrases in the sentences that can be combined and used to disclose the identity of a person. the PII entities we looked for were; PERSON, NORP, ORG, LOCATION, USERID and DATE

6 RECORDING QUALITY

N/A

References

[1] Bender, Emily M., and Batya Friedman. "Data statements for natural language processing: Toward mitigating system bias and enabling better science." Transactions of the Association for Computational Linguistics 6 (2018): 587-604.

[2] Gebru, Timnit, et al. "Datasheets for datasets." Communications of the ACM 64.12 (2021): 86-92.