Lacuna PII Multilingual Data Paper

Andrew Katumba[a], Joyce Nakatumba-Nabende[b]
[a]Department of Electrical and Computer Engineering, Makerere University, Kampala, Uganda
[b]Department of Computer Science, Makerere University, Kampala, Uganda

[*]Corresponding authors: Peter Nabende, Ann Lisa Nabiryo, Bridget Cecile Nakakande, Eric Peter Wairagala, Jenifer Winfred Namuyanja

May 2024

Abstract

This dataset consists of text data containing personally identifiable information (PII) across four languages: Luganda, Lumasaaba, Hausa, and Kanuri. The dataset can be utilized for various natural language processing tasks, including but not limited to, named entity recognition, text classification, and privacy-preserving data analysis. The dataset aims to provide a valuable resource for linguistic and computational research in these languages.

Keywords: Luganda; Lumasaaba; Hausa; Kanuri; PII; Multilingual Data

Specifications table

Subject	Person identifiable information
Specific subject area	Person identifiable Information
Type of data	Text data
How data were acquired	Text data: a multilingual corpus of Luganda, Lumasaaba, Hausa, and Kanuri PII sentences.
Data format	Text data: CSV files.
Parameters for data collection	Text data: • Luganda: 4000 PII sentences with representation in entities of Person, Date, National identity, tribal identity, location, and organization. • Lumasaaba: 5000 PII sentences with representation in entities of Person, Date, National identity, tribal identity, location, and organization. • Hausa: 3000 PII sentences with representation in entities of Person, Date, National identity, tribal identity, location, and organization. • Kanuri: 3000 PII sentences with representation in entities of Person, Date, National identity, tribal identity, location, and organization.

Description of data collection	Text data: • Luganda: PII text data curated by Makerere AI Lab researchers. The text data was collected from websites and media sources, tweaked, and validated by Luganda language experts. • Lumasaaba: Text data collected and curated by Makerere AI Lab researchers from various sources, including local media and community documentation, validated by Lumasaaba language experts. • Hausa: Text data collected through the CLEAR Global organization from open-source datasets on Hugging Face, including sources like VOA News, and validated by Hausa language experts. • Kanuri: Text data collected through the CLEAR Global organization from open-source datasets on Hugging Face and a corpus created by CLEAR Global, validated by Kanuri language experts.
Data source location	Text data: • Luganda: Various sources, including MasakhaNER project, Dembe FM, Galaxy FM, Ssegwanga newspaper, Gambuuze newspaper. Collected by: Makerere AI Lab researchers. Location: Uganda. • Lumasaaba: Various sources including local media and community records. Collected by: Makerere AI Lab researchers. Location: Uganda. • Hausa: VOA News, open-source datasets on Hugging Face. Collected by: CLEAR Global. Location: Nigeria, Niger. • Kanuri: Corpus created by CLEAR Global, open-source datasets on Hugging Face. Collected by: CLEAR Global. Location: Nigeria, Niger, Cameroon.
Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?	Text data; This data is available and can be downloaded at https://doi.org/10.7910/DVN/CGHWZE
Related research articles	N/A

Value of the Data

The dataset was created to provide quality PII datasets for Luganda, Lumasaaba, Hausa, and Kanuri languages. The text data collected includes sentences with PII entities, forming a multilingual corpus that is useful for various natural language processing tasks. This data can be used for different tasks such as named entity recognition, text classification, and privacy-preserving data analysis, contributing to both linguistic research and practical applications in these languages.

1. Named Entity Recognition (NER) The dataset can be used to develop NER models that can accurately identify and classify PII entities in Luganda, Lumasaaba, Hausa, and Kanuri text. This is essential for applications that require the extraction of specific information from text, such as automated form filling and information retrieval systems.

2. Text Classification The PII data can be utilized to train text classification models that can automatically categorize documents based on the presence of PII. This is particularly useful for organizations that need to manage and protect sensitive information in large volumes of text data.

3. Privacy-Preserving Data Analysis In the context of GDPR and other privacy regulations, the dataset provides a valuable resource for developing techniques for anonymizing and pseudonymizing PII in text data. This can help organizations comply with legal requirements while still being able to analyze and use their data effectively.

4. Language Modeling The dataset can be used to create domain-specific language models for the included languages. These models can be fine-tuned for specific tasks such as machine translation, sentiment analysis, and conversational AI, improving the performance of NLP applications in these languages.

5. Machine Translation By incorporating PII data from different languages, the dataset can be used to improve machine translation systems for low-resource languages. This can enhance the quality of translations for texts containing PII, ensuring accurate and contextually appropriate translations.

6. Linguistic Research The dataset provides a rich resource for linguistic analysis and research, offering insights into the use of PII in different languages and cultural contexts. Researchers can study the structure, usage patterns, and frequency of PII entities in the text, contributing to the understanding of language use in digital communication.

1 Data Description

The text data is a multilingual corpus, consisting of text sentences in Luganda, Lumasaaba, Hausa, and Kanuri. The full dataset contains:

- Luganda: 4000 PII sentences
- Lumasaaba: 5000 PII sentences
- Hausa: 3002 PII sentences
- Kanuri: 3005 PII sentences

2 Experimental Design, Materials and Methods

2.1 Data Collection

The sentences were obtained from various sources. These included:

- MasakhaNER project
- Dembe FM
- Galaxy FM
- Ssegwanga newspaper
- Gambuuze newspaper
- VOA News (Hausa)
- Open-source datasets available on Hugging Face (Hausa and Kanuri)
- Corpus created by CLEAR Global (Kanuri)

Sentences from all these various sources were written in Luganda, Lumasaaba, Hausa, and Kanuri. The text data was curated, validated, and preprocessed to ensure it contained accurate PII entities and was grammatically correct.

2.2 Data Preprocessing

The raw data was collected and stored on backup storage at the Makerere Artificial Intelligence Lab. The text data was validated by language experts to ensure accuracy and consistency. Preprocessing steps included text normalization, removing invalid sentences, and formatting the data for use in machine learning tasks.

3 Ethics Statement

The work presented in this paper is original and has not been published or submitted for consideration anywhere else. All the required consent to collect and publish the data was gathered from the responsible parties.

4 Credit Author Statement

Andrew Katumba: Conceptualization, review, and editing
Joyce Nakatumba-Nabende: Conceptualization, review, and editing
Ann Lisa Nabiryo: Writing, reviewing, and editing, data curation
Peter Nabende:Writing, reviewing, and editing
Bridget Cecile: Writing, reviewing, and editing
Eric Peter Wairagala:Writing, reviewing, and editing
Jenifer Winfred Namuyanja:Writing, reviewing, and editing

Creation dates Start date : 2023-06-06 and End date: 2024-05-30.

Dataset creators Marconi Lab-Uganda, Makerere Artificial Intelligence Lab-Uganda, Clear Global Organization-Nigeria

Language Languages used in the dataset include Luganda, Lumasaaba, Hausa and Kanuri

License The data has been uploaded under an open license CC0.

Repository name The name of the repository to which the data is uploaded is Harvard dataverse.

Publication date The dataset was published on 2024-06-03.

5 Reuse Potential

The dataset provided in this paper offers significant reuse potential across a variety of research domains and applications. Researchers in the field of natural language processing (NLP) can leverage this dataset for the development and improvement of named entity recognition (NER) models, particularly for low-resource languages such as Luganda, Lumasaaba, Hausa, and Kanuri. The dataset's inclusion of diverse PII entities makes it an invaluable resource for training and evaluating text classification models, enhancing their ability to categorize and manage sensitive information.

In addition to NLP applications, the dataset can be utilized for privacy-preserving data analysis, aiding in the development of techniques for anonymizing and pseudonymizing PII. This is crucial for organizations seeking to comply with privacy regulations while analyzing textual data. Furthermore, the dataset can be employed in linguistic research to study the structure and usage patterns of PII entities across different languages and cultural contexts, offering insights that can inform both academic research and practical applications.

Machine translation systems can also benefit from this dataset, as it provides a rich resource for improving translations involving PII, ensuring both accuracy and context sensitivity. The dataset's multilingual nature supports the creation of domain-specific language models, which can be fine-tuned for various NLP tasks such as sentiment analysis and conversational AI, thereby enhancing the performance of applications in these languages.

Overall, the dataset's comprehensive collection of PII sentences across multiple languages makes it a valuable asset for researchers and practitioners in NLP, privacy-preserving technologies, linguistic studies, and machine translation, among others..

Acknowledgements

We would like to thank Lacuna for funding the data collection activities in this research project. We would also like to acknowledge Clear global organization and stakeholders for their contributions to the text data collection of Kanuri and Hausa languages.

Funding Statement

Lacuna; Grant number : 2021-NLP-131.

Declaration of Competing Interest

The authors have no competing interests to declare.

References

Li, Z., Hu, C., Guo, X., Chen, J., Qin, W., & Zhang, R. (2022). An Unsupervised Multiple-Task and Multiple-Teacher Model for Cross-lingual Named Entity Recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 170-179. doi: 10.18653/v1/2022.acl-long.14.

Mager, M., et al. (2021). Findings of the AmericasNLP 2021 Shared Task on Open Machine Translation. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*. doi: 10.18653/v1/2021.americasnlp-1.5.

Ogueji, K., & Okagbue, H. I. (2020). MasakhaNER: Named Entity Recognition for African Languages using Transfer Learning. *arXiv preprint arXiv:2010.15300*.

Hedderich, M. A., Lange, L., Adel, H., Strötgen, J., & Klakow, D. (2021). A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. In *Proceedings of the 28th International Conference on Computational Linguistics*. doi: 10.18653/v1/2021.coling-1.

References