DATASHEET:
Lelapa-X-NER (isiZulu)

Lelapa AI
info@lelapa.ai

This document is based on *Datasheets for Datasets* by Gebru *et al.* [1]. Please see the most updated version here.

MOTIVATION

For what purpose was the dataset created?
The dataset was created to enable research on named entity recognition (NER) for isiZulu.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
The dataset was created by Masakhane.

What support was needed to make this dataset?
Lacuna Fund.

Any other comments?
None.

COMPOSITION

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?
The instances are texts collected from newspaper and their associated entity tags.

How many instances are there in total (of each type, if appropriate)?
128,658 (tokens) **

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?
The dataset was collected from Isolezwe Newspaper, so we can not confirm that it contains all possible instances/domains.

What data does each instance consist of?
Each instance consists of text and associated entity types. The data was collected, annotated, and curated with natural language processing best practices in mind.

Is there a label or target associated with each instance?
Yes, each word is annotated with one type. These are Personal name (PER), Location (LOC), Organization (ORG), date & time (DATE). and Other (O).

Is any information missing from individual instances?
Everything is included. No data is missing.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?
No.

Are there recommended data splits (e.g., training, development/validation, testing)?
The dataset has Train, Dev, and Test split.

Are there any errors, sources of noise, or redundancies in the dataset?
No.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? The dataset is entirely self-contained.

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 relate to people?
No.

Does the dataset identify any subpopulations (e.g., by age, gender)?
N/A.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?
No.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial 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)?
No.

Any other comments?
None.

COLLECTION

How was the data associated with each instance acquired?
The data was acquired from open-source resources/websites.

Over what timeframe was the data collected?
Unknown to the authors of the datasheet.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?
The dataset is crawled using a simple script.

What was the resource cost of collecting the data?
Unknown to the authors of the datasheet.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?
N/A

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?
The dataset is collected by scraping Isolezwe Newspaper websites

Were any ethical review processes conducted (e.g., by an institutional review board)?
N/A.

Does the dataset relate to people? If not, you may skip the remainder of the questions in this section.
No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?
The dataset is crawled from Isolezwe Newspaper websites

Were the individuals in question notified about the data collection?
N/A

Did the individuals in question consent to the collection and use of their data?
N/A

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?
N/A

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis)been conducted?
N/A.

Any other comments?
None.

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)?
None.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?
N/A.

Is the software used to preprocess/clean/label the instances available?
N/A.

Any other comments?
None.

USES

Has the dataset been used for any tasks already?
Yes, for research on Named Entity Recognition (NER) for isiZulu.

Is there a repository that links to any or all papers or systems that use the dataset?
`https://arxiv.org/pdf/2210.12391.pdf,https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00416/107614`

What (other) tasks could the dataset be used for?
N/A.

Is there anything about the composition of the dataset or the way it was collected and

preprocessed/cleaned/labeled that might impact future uses?
No.

Are there tasks for which the dataset should not be used?
No.

Any other comments?
None.

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, it is open-source.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?
The dataset is open-sourced and can be accessed from Masakhane Hugging Face[1] or GitHub[2] sites.

When will the dataset be distributed?
The dataset is open-source.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?
The dataset is distributed with Academic Free License (AFL) version 3.0

Have any third parties imposed IP-based or other restrictions on the data associated with the instances?
No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?
No.

Any other comments?
None.

MAINTENANCE

Who is supporting/hosting/maintaining the dataset?
Masakhane

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
Email: info@lelapa.ai

Is there an erratum?
No.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?
Yes. As required for company research and commercial purposes.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?
N/A.

Will older versions of the dataset continue to be supported/hosted/maintained?
Older versions will be kept around for consistency.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?
No.

Any other comments?
None.

[1] https://huggingface.co/masakhane
[2] https://github.com/masakhane-io/masakhane-ner/

REFERENCES

[1] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. Datasheets for Datasets. *arXiv:1803.09010 [cs]*, January 2020.