Model Card

Lelapa-X-ASR (isiZulu and seSotho)

Model Details **1**

- Intended use 2
  - Primary intended uses 2
  - Primary intended users 2
  - Out-of-scope use cases 2
- Factors 2
  - Relevant factors 2
  - Evaluation factors 3
- Metrics 3
  - Model performance measures 3
  - Decision thresholds 4
  - Approaches to Uncertainty and Variability 4
- Evaluation data 4
  - Datasets 4
  - Motivation 4
  - Preprocessing 5
- Training data 5
- Quantitative analyses 5
  - Unitary results 5
  - Human evaluation 6
  - Intersectional result 7
- Ethical considerations 7
- Caveats and recommendations 7

Model Details

Basic information about the model: Review section 4.1 of the model cards paper.

Organization	Lelapa AI
Product	Vulavula
Model date	7 November 2023
Feature	ASR
Lang	isiZulu and seSotho
Domain	Call Center
Model Name	Lelapa-X-ASR (isiZulu & seSotho)
Model version	1.0.0
Model Type	Fine-Tuned Proprietary Model

Information about training algorithms, parameters, fairness constraints or other applied approaches, and features: Proprietary Fine-tuning of a Base Model on Transcription Data

License: Proprietary

Contact: info@lelapa.ai

Intended use

Use cases that were envisioned during development: Review section 4.2 of the model cards paper.

Primary intended uses

Intended use is governed by the language and domain of the model. The model is intended to be used in the call center domain for transcription of calls that are conducted in isiZulu and seSotho. The model is not suitable for the general conversation domain and should be used with extreme caution in high-risk environments.

Primary intended users

Transcription to enable analysis for downstream tasks in the call center domain for isiZulu and seSotho:

- Compliance monitoring for Customer Interactions
- Quality assurance
- Enabling search and filter of conversations

Out-of-scope use cases

All domains and languages outside of the call center analytics space for isiZulu and seSotho.

Factors

Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3: Review section 4.3 of the model cards paper.

Relevant factors

Groups:

- Users who recorded utterances used to train the model are diverse across several factors such as age, location (primarily South Africa but from several regions/parts of the country depending on the language), and gender (both males and females are equally distributed across speakers). There is no record of the social class of speakers, as well as their health conditions, names, and any other sort of privacy details. Further details of groups and their constituents can be found in the datasheet
- Performance across groups is underway

Environmental conditions, Instrumentation and technical attributes:

- Audio utterances are recorded in environments such as rooms, and call centers with a noiseless background.
- Audio segments' length varies from 3 seconds to 30-40 minutes.

Evaluation factors

- In our development setting (training and evaluation) we used the factors described above with additional synthetic arrangements to improve the robustness of the model to real-world factors

Metrics

The appropriate metrics to feature in a model card depend on the model being tested. For example, classification systems in which the primary output is a class label differ significantly from systems whose primary output is a score. In all cases, the reported metrics should be determined based on the model's structure and intended use: Review section 4.4 of the model cards paper.

Model performance measures

The model is evaluated using WER as well as human evaluation: The models' performances are measured by both automatic metrics and human evaluation. As an automatic metric, we use the Word Error Rate (WER) which is based on the edit distance also called Levenshtein distance. WER is not a symmetric distance metric, since it measures the number of `operations`: substitution, deletion, insertion, number of correct words needed to leave a reference sentence `A` to a predicted sentence `B`. Read more. As far as human evaluation is concerned, this stage is performed by paid linguists, native speakers of the languages under study. Evaluation is also done after post-processing techniques are performed on the outputs.

WER: Testing on general isiZulu and seSotho data (incl call center)

WER after post-processing: Post-processing of output predictions.

Decision thresholds

No decision thresholds have been specified

Approaches to Uncertainty and Variability

For `fairness`, `robustness`, and `generalization` with respect to languages and datasets, we have leveraged standard downsampling and normalization techniques which have proven to be useful.

Evaluation data

All referenced datasets would ideally point to any set of documents that provide visibility into the source and composition of the dataset. Evaluation datasets should include datasets that are publicly available for third-party use. These could be existing datasets or new ones provided alongside the model card analyses to enable further benchmarking.

Review section 4.5 of the model cards paper.

Datasets

- Publicly available datasets in the general domain
- Proprietary call center dataset

Motivation

These datasets have been selected because they are open-source, high-quality, and cover the targeted languages - and utterances are recorded by a variety of speakers living in required regions. These help to capture interesting cultural and linguistic aspects that would be crucial in the development process for better performance.

Preprocessing

Data utterances are filtered initially by audio length, sampled, and normalized transcripts. We also make sure to select actual recordings i.e. recordings that are not just noise or blank.

Training data

Review section 4.6 of the model cards paper.

Refer to the datasheet provided

Quantitative analyses

Quantitative analyses should be disaggregated, that is, broken down by the chosen factors. Quantitative analyses should provide the results of evaluating the model according to the chosen metrics, providing confidence interval values when possible.

Review section 4.7 of the model cards paper.

Unitary results

WER	
WER isiZulu (General and Call Center)	0.4570
WER isiZulu (After Post-Processing)	0.3692
WER seSotho (General and Call Center)	0.3697
WER seSotho (After Post-Processing)	0.2943

Human evaluation

This is a breakdown of the types of errors we are seeing based on a sample of the evaluation dataset.

*Note: some samples suffered from more than 1 type of error

isiZulu	**Yes**	**No**	**Total**
Transcription Correct	84	16	100
Prediction Correct	42	58	100

Ambiguous Audio Input		1	
Context-Breaking		12	
Flawed Audio Input		7	
Homophone		10	
Name, Anglicism, Loan Word		5	
Non-Context-Breaking		10	
Flawed Ground Truth Transcript		6	
Negligible		7	

seSotho	**Yes**	**No**	**Total**
Transcription Correct	70	30	100
Prediction Correct	58	42	100
Ambiguous Audio Input		4	
Context-Breaking		9	
Flawed Audio Input		8	
Homophone		0	
Name, Anglicism, Loan Word		5	
Non-context-Breaking		16	
Flawed Ground Truth Transcript		0	
None		0	

Intersectional result

In progress

Ethical considerations

This section is intended to demonstrate the ethical considerations that went into model development, surfacing ethical challenges and solutions to stakeholders. The ethical analysis does not always lead to precise solutions, but the process of ethical contemplation is worthwhile to inform on responsible practices and next steps in future work: Review section 4.8 of the model cards paper.

All call center data is synthetic and so the model does not contain any personal information. More details in the datasheet.

Caveats and recommendations

This section should list additional concerns that were not covered in the previous sections.

Review section 4.9 of the model cards paper.

Additional caveats are outlined extensively in our Terms and Conditions.