Towards End-to-End Training of Automatic Speech Recognition for Nigerian Pidgin 🗣️

Amina Mardiyyah Rufai[1*], Afolabi Abeeb[2*], Esther Oduntan[1], Tayo Arulogun[2], Oluwabukola Adegboro[1†], Daniel Ajisafe[1†]

[1]African Institute for Mathematical Sciences, [2]Ladoke Akintola University of Technology, Nigeria, [*]Equal Contribution, [†]Co-supervisory role

Introduction

Motivation

- Africa has more than **2000** languages 🌍 [1], while;
  - **Automatic speech recognition** (ASR) systems 🗣️ are increasing recently
  - However, African languages lack sufficient linguistic resources to support ASR systems
- This **study** focuses on developing an end-to-end ASR system for "**Nigerian Pidgin English**" – the most prevalent form in West Africa (🇳🇬)

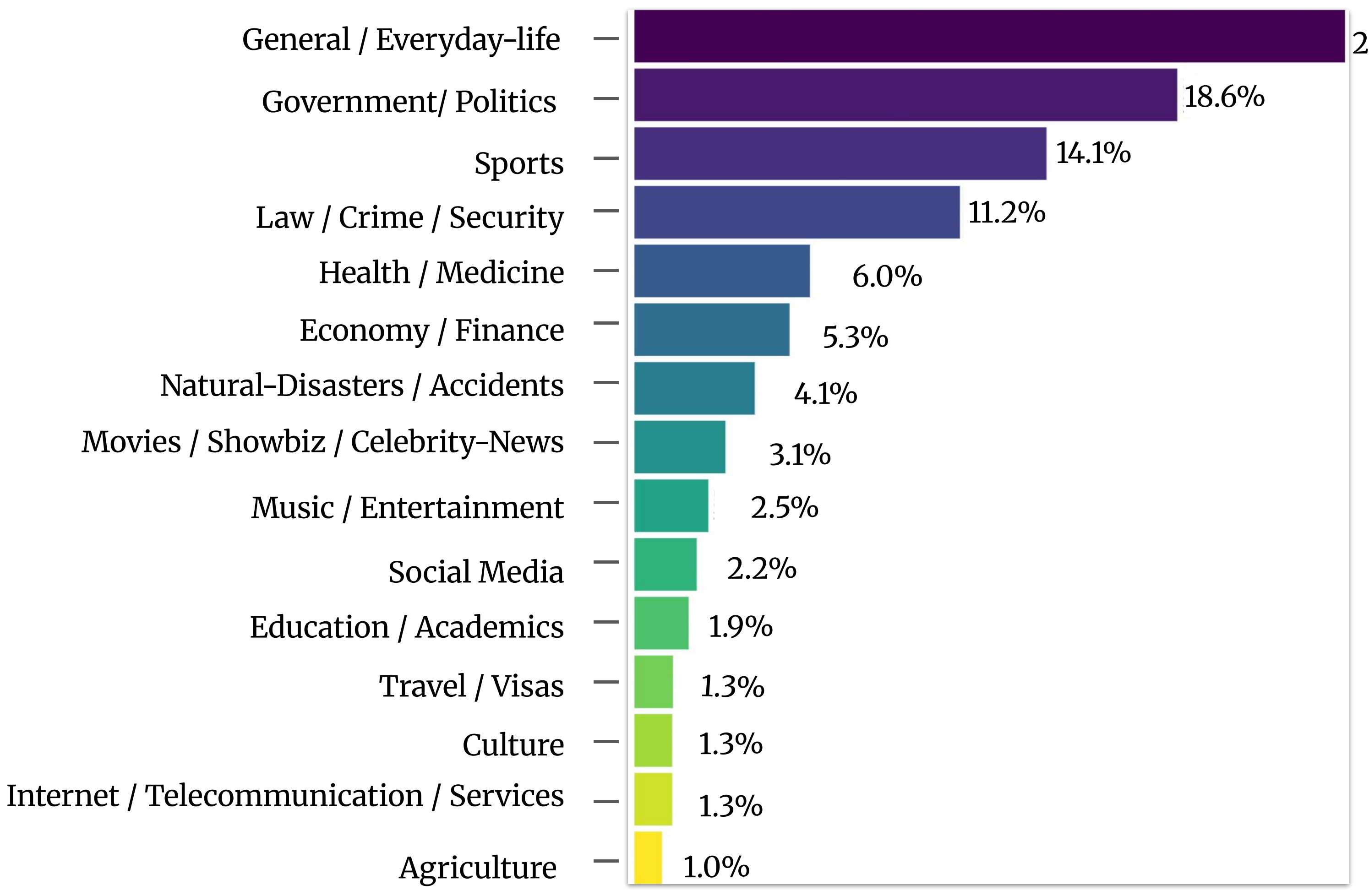
Key Contributions

- ❖ We demonstrate that a pretrained state-of-the-art model do not work well out-of-the-box, and reduce error rate by **59.84%** 🚀
- ❖ We release our unique **parallel** dataset (speech-to-text) on Nigerian Pidgin, as well as the model weights on Hugging Face 🤗
- ❖ We introduce a publicly accessible end-to-end ASR system for **community engagement** 👨 👩

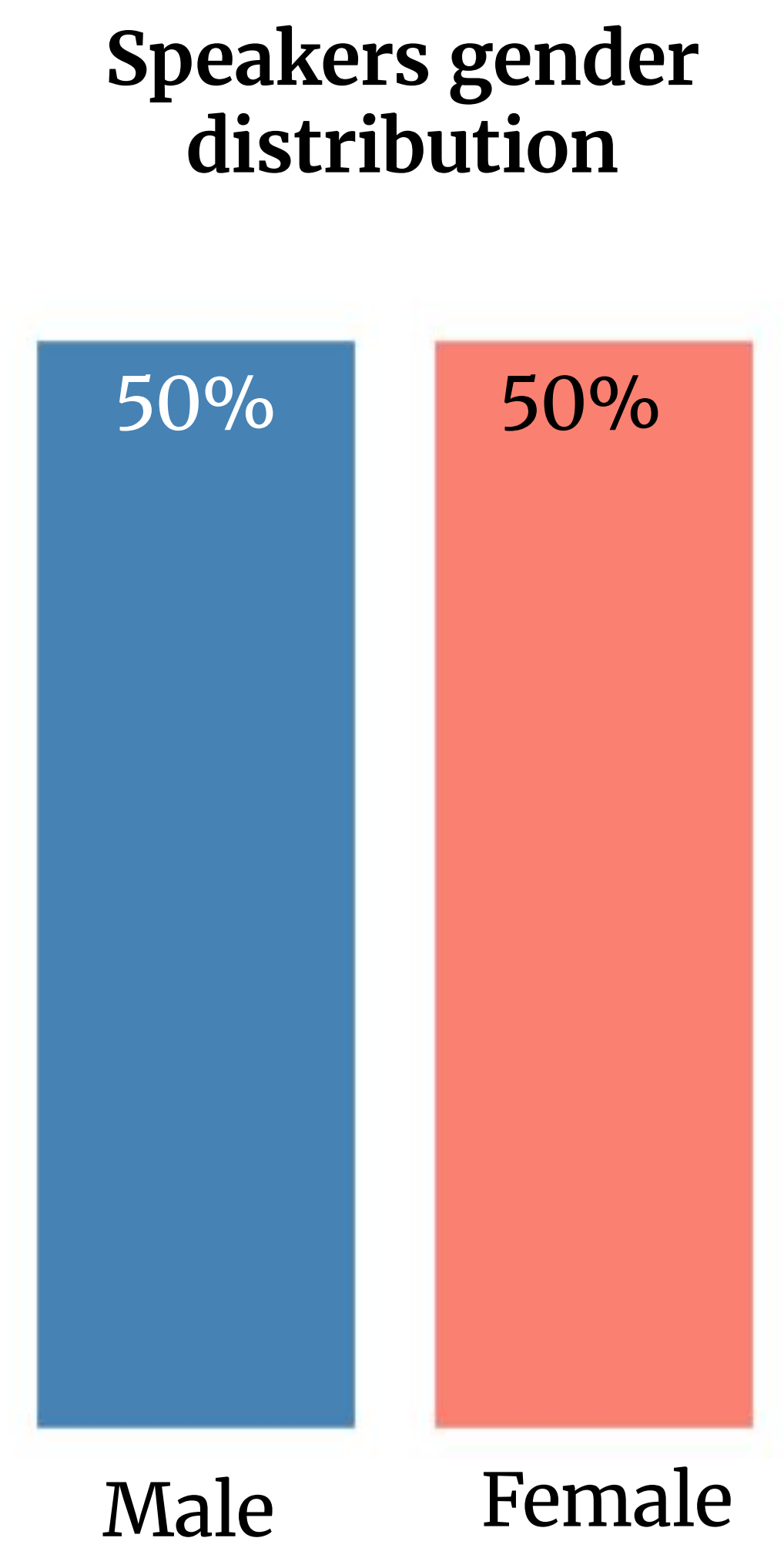
Methodology

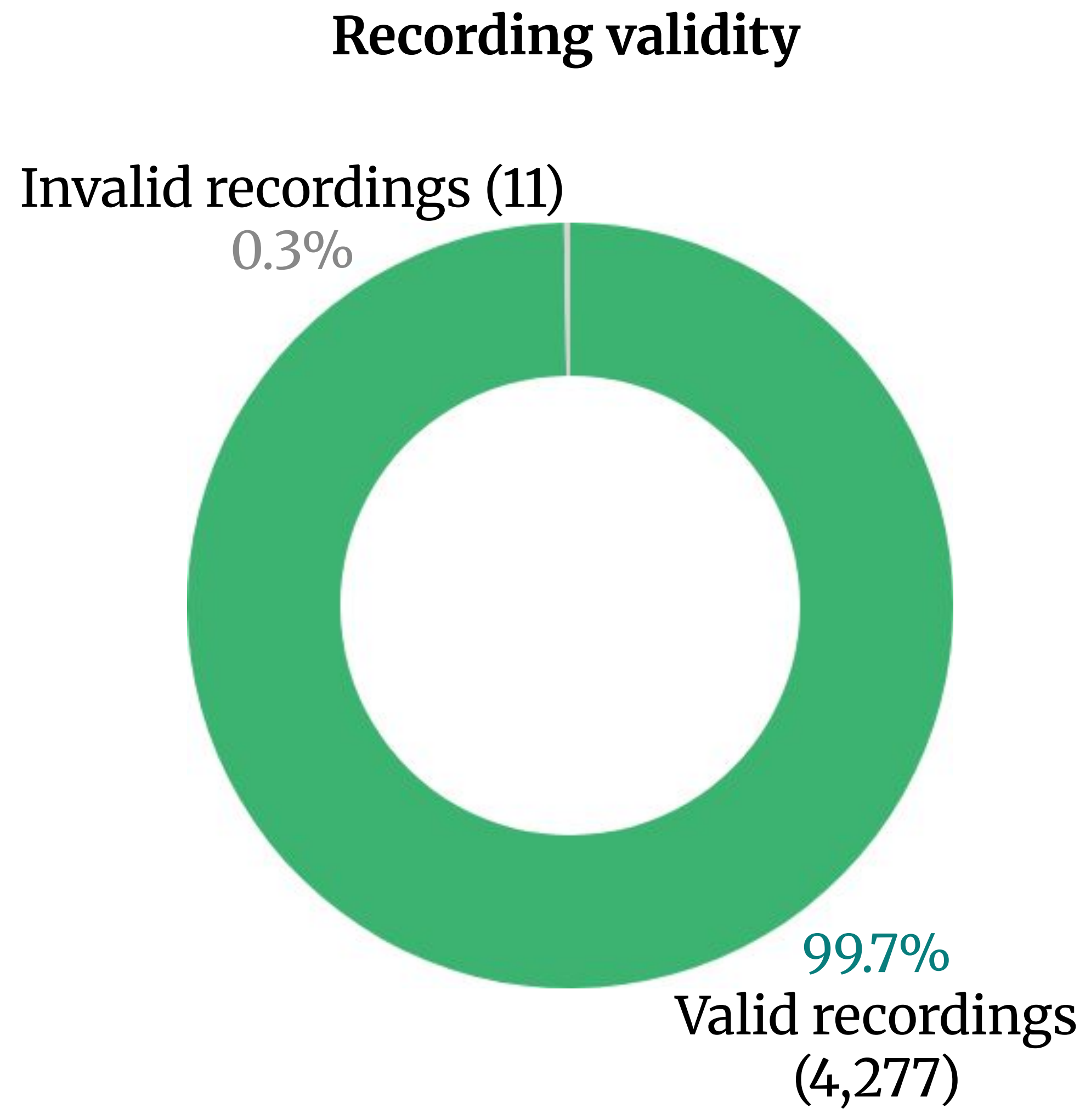
Topic Distribution

Using BERTopic, we revealed 15 themes in the Nigerian Pidgin text dataset, with "Everyday Conversation" and "Politics" emerging as the most **prominent** across the collected texts

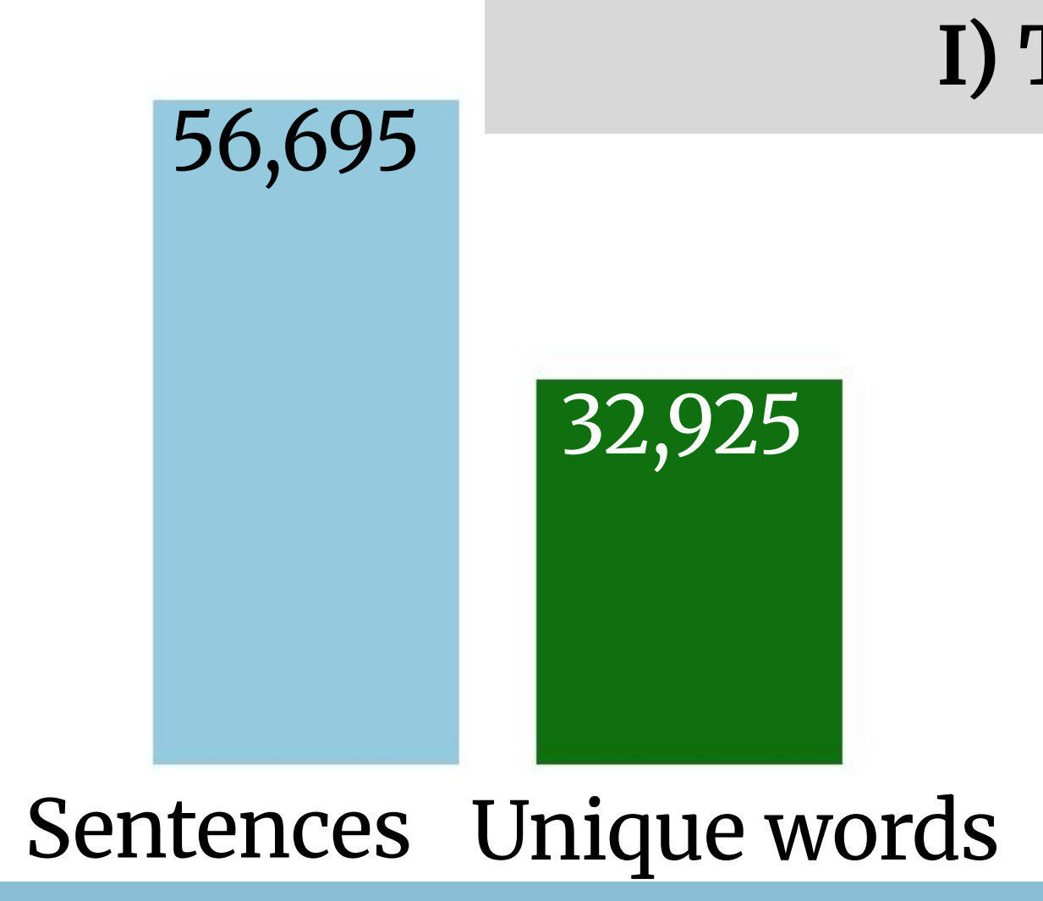


II) Speech Corpus





I) Textual Data



- ❏ Utterances selected: 4,288
- ❏ Avg. words/sentence: 8-17
- ❏ Avg. audio duration: ~17 seconds

Model Architectures

The study evaluated several ASR models, including:

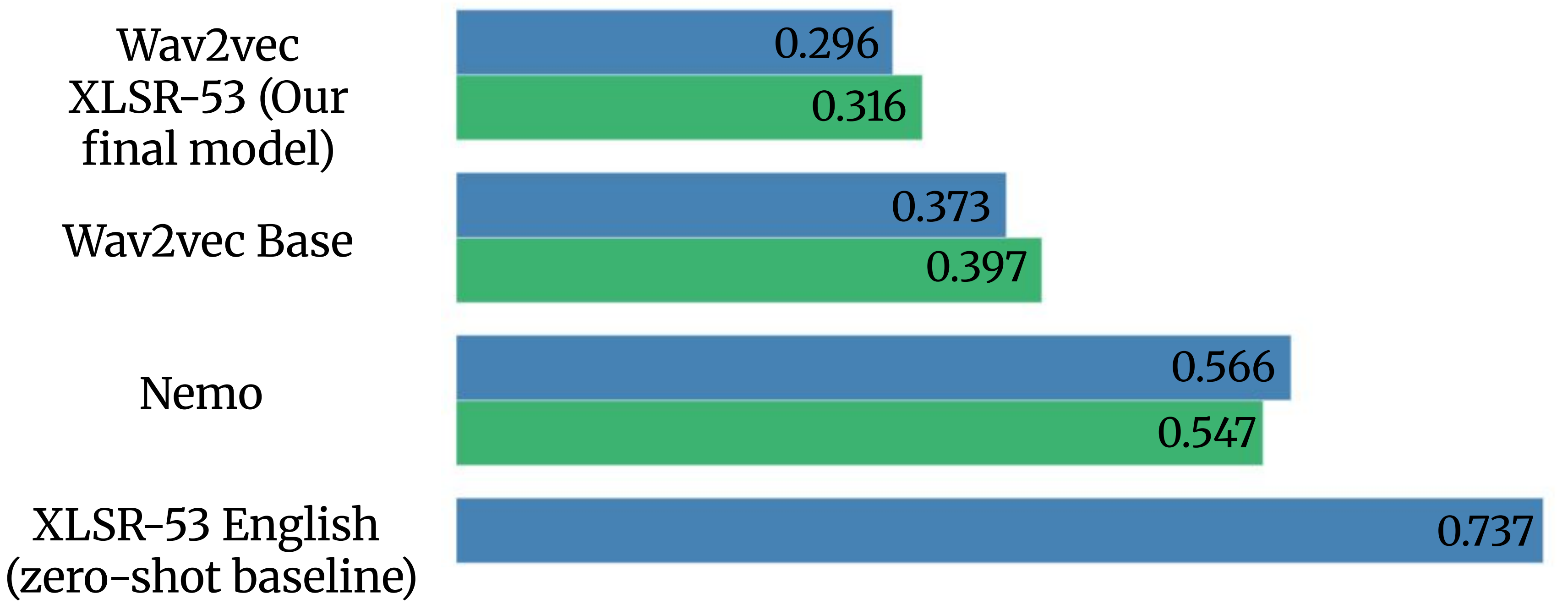
- XLSR-English [2] (**zero-shot baseline**)
- Nemo QuartzNet [3]
- Wav2Vec 2.0 Base-100h [4]
- Wav2Vec XLSR-53 [5] (Our final model)

Result and Discussion

Result

Word Error Rate (WER) was used to evaluate performance, with Wav2Vec XLSR-53 achieving the lowest WER and effectively capturing Nigerian Pidgin terms, though it struggled with accurate number recognition

Model Comparison: Validation and Test WER



*feature encoder weights for Wav2vec models were unfrozen

Qualitative Comparison of Predictions

Reference	Our final model	Zero-shot prediction
pipo and all di poor pipo wey govment gats take care of	pipo and all di poor pipo wey govrment gats take care of	people and ol the poor peopleway government gats take care of
so dat one con mean say no show for dem next year	so dat one con mean say no show for dem next year	so thats on't calm me in senushu for them next year

Insights / Discussion

- Superior performance courtesy of an effective **cross-lingual** architecture
- Effective fine-tuning on Nigerian Pidgin data capturing language **nuances**
- Access to a high-quality, training-augmented **native** speech dataset

Ethics, Limitation and Conclusion

Ethics and Limitation

- Informed consent from speakers and privacy protection
- Limitations in data size, regional dialect coverage and numerical elements, constrains model generalisability and robustness → an **avenue** for future work

Conclusion

Fine-tuning our best model on Nigerian Pidgin reduced error-rate from 73.7% to 29.6%, highlighting the need for **domain-specific** data, **effective** approaches and continued **collaboration**

References

[1] Jade Abbott and Laura Martinus. Benchmarking neural machine translation for southern african languages. In Proceedings of the 2019 Workshop on Widening NLP, pages 98–101, 2019

[2] Grosman, J. (2021). Fine-tuned XLSR-53 large model for speech recognition in English. Hugging Face. Link: https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-english

[3] Samuel Kriman, Stanislav Beliaev, Boris Ginsburg, Jocelyn Huang, Oleksii Kuchaiev, Vitaly Lavrukhin, Ryan Leary, Jason Li, and Yang Zhang. Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions. In ICASSP 2020 - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6124–6128. IEEE, 2020

[4] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems, 33:12449–12460, 2020

[5] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised cross-lingual representation learning for speech recognition. arXiv preprint arXiv:2006.13979, 2020