ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14908879

Enhancing Conversational AI for Low-Resource Languages: A Case Study on Somali

Mohamud Osman Hamud[1]; Serpil Aydın (Supervisor)[2]

[2]Assistant Professor
[1,2]Ondokuz Mayıs University, Faculty of Science, Department of Statistics, Samsun, Turkey

Publication Date: 2025/02/22

Abstract: Conversational AI has made huge strides in understanding and generating human language. However, these advances have mostly benefited high-resource languages such as English and Spanish. In contrast, languages like Somali—spoken by an estimated 20 million people—lack the abundance of annotated data needed to develop robust language models. This study focuses on practical strategies to boost Somali text and speech processing capabilities. We explore three core approaches: (1) transfer learning, (2) synthetic data augmentation, and (3) fine-tuning multilingual models. Our experiments, featuring XLM-R, mBERT, and OpenAI's Whisper API, show that well-adapted models significantly outperform their baseline counterparts in Somali text translation and speech-to-text tasks. Beyond the numbers, our findings underscore the societal value of creating accessible AI tools for underrepresented linguistic communities, providing a template for extending these methods to other low-resource languages.

How to Cite: Mohamud Osman Hamud; Serpil Aydın (Supervisor). (2025). Enhancing Conversational AI for Low-Resource Languages: A Case Study on Somali. *International Journal of Innovative Science and Research Technology*, 10(2), 290-293. https://doi.org/10.5281/zenodo.14908879.

I. INTRODUCTION

A. Background and Motivation

Conversational AI systems—commonly embodied by chatbots, virtual assistants, and interactive voice-response systems—have rapidly evolved to handle a range of tasks, from online customer service to language tutoring. Unfortunately, most of these systems are optimized for a few dominant languages that benefit from large, high-quality datasets. This leaves many low-resource languages without the same level of development and leads to a growing "linguistic digital divide" (Bender et al., 2021).

Somali, a Cushitic language with multiple dialects, illustrates this divide. Despite Somali being spoken by millions within the Horn of Africa and the global diaspora, its digital footprint remains relatively small, hindering the development of advanced NLP tools. Somali dialectal differences, limited text corpora, and insufficient annotated speech data further complicate the training process for AI models.

B. Research Objectives

➢ *This Study Aims to Address the Scarcity of Somali NLP Resources by Experimenting with Targeted Strategies:*

- Harnessing Transfer Learning: We fine-tune multilingual transformer models (XLM-R and mBERT) on a curated Somali dataset to see if these general-purpose architectures can achieve language-specific proficiency.
- Exploring Synthetic Data Augmentation: We generate additional Somali-like text using back-translation methods, hoping to enrich existing datasets without the labor-intensive task of manual annotation.
- Evaluating Speech Recognition: By incorporating OpenAI's Whisper API for Somali audio, we gauge how fine-tuned approaches affect tasks like speech-to-text.

Through these objectives, we seek both empirical improvements in NLP performance and practical insights into creating more inclusive AI systems.

II. LITERATURE REVIEW

A. State of Conversational AI

Over the last decade, the application of deep learning to language tasks has led to breakthroughs in machine translation, sentiment analysis, and question answering. Large-scale language models, such as GPT (Brown et al., 2020) and BERT (Devlin et al., 2019), introduced novel architectures that capture long-range dependencies and contextual relationships. Yet, most gains have centered on resource-rich languages due to the abundance of text data available for training.

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14908879

B. Challenges in Low-Resource Language Modeling

- **Data Scarcity** remains the core challenge. A robust conversational AI system typically demands massive annotated datasets that capture a wide array of linguistic nuances. Additionally, many low-resource languages exhibit:
- **Dialectal Variations**: A single language may have multiple dialects or distinct regional forms. Models trained on one dialect can struggle with another.
- **Limited Digitized Text**: Official documents and public materials might not be fully digitized, and what is online often lacks the volume required for data-hungry architectures (Conneau et al., 2020).
- **Sparse Annotated Speech**: Collecting audio data with corresponding transcriptions is both time-consuming and resource-intensive.

C. Previous Approaches

➢ *Several Approaches have Emerged to Counteract Limited Data Issues:*

- Transfer Learning: Pre-trained models—like mBERT or XLM-R—are adapted to a target language with fewer resources, potentially boosting performance significantly (Howard & Ruder, 2018).
- Synthetic Data Generation: Techniques like back-translation create additional training samples, effectively expanding the dataset without human intervention (Fadaee et al., 2017).
- Few-Shot and Zero-Shot Learning: Large-scale multilingual models leverage global language patterns to work on new tasks or languages with minimal or zero task-specific training (Brown et al., 2020).

Our work integrates and expands upon these methods, delivering a hands-on exploration of how to optimize them for Somali.

III. METHODOLOGY

A. Data Collection

➢ *Textual Resources:*

- Somali Wikipedia: Although smaller than English Wikipedia, it still offers articles on culture, geography, and history.
- Online News Outlets: News websites such as Hiiraan Online and BBC Somali provided more contemporary and journalistic language styles.
- Open-Source Somali Corpora: We included any publicly available Somali-language datasets, although these remain quite limited.

➢ *Speech Resources:*

- Manually Transcribed Audio: Volunteers contributed short recordings in various Somali dialects, which we hand-transcribed to ensure accuracy.
- OpenAI's Whisper API: For additional Somali audio, we applied the Whisper ASR model to yield preliminary transcripts, which were then lightly reviewed by fluent speakers.

(Figure 1, if inserted, could depict the data collection pipeline, from web scraping to speech transcription.)

B. Data Preprocessing

➢ *After Data Collection, we Performed Essential Preprocessing Steps:*

- Tokenization: Using language-specific tokenizers to split text into words or subword units.
- Normalization: Lowercasing and removing extraneous symbols, especially from scraped online sources.
- Stemming: Reducing words to their linguistic roots, which helps unify variations in inflected forms.
- Noise Reduction: Filtering out incomplete sentences, HTML tags, or any non-Somali text.

C. Model Architecture

➢ *We Focused on Two Transformer-Based Architectures known for their Multilingual Capabilities:*

- XLM-R (Conneau et al., 2020): A robust cross-lingual language model pre-trained on 100+ languages.
- mBERT (Devlin et al., 2019): Multilingual BERT, known for decent cross-lingual transfer, but it sometimes underperforms on heavily underrepresented languages compared to more recent models like XLM-R.

D. Training and Fine-Tuning

➢ *Three Experimental Setups were Conducted:*

- Baseline Model: We tested XLM-R and mBERT in their vanilla (pre-trained) forms without any Somali-specific fine-tuning. This served as our control group.
- Fine-Tuned Model: We trained XLM-R on Somali text and transcripts. This step allowed the model's internal representations to better adapt to Somali lexicon and syntax.
- Hybrid Model (mBERT + Augmentation): We combined mBERT with synthetic Somali data. By back-translating English texts into Somali using an existing Somali-English NMT model, we augmented the training set, then further fine-tuned mBERT on this enriched corpus.

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14908879

E. Ethical Considerations

- Consent and Privacy: Participants who contributed their speech data were briefed on how and where the audio would be used, with options to withdraw.
- Dialectal Fairness: We recognized that some dialects might be underrepresented; attempts were made to include at least a moderate variety in the dataset.
- Data Bias: Overrepresentation of certain topics (e.g., news about specific regions) could bias the models. We worked to balance news, literary text, and encyclopedic content.

IV. EXPERIMENTAL SETUP

A. Performance Metrics

➢ *We used Multiple Metrics to Capture Both Quantitative Accuracy and the More Subjective Aspects of Language Fluency:*

- BLEU: Standard measure for evaluating machine-translated text against a human reference.
- Word Error Rate (WER): To assess the accuracy of speech-to-text outputs.
- Perplexity: Reflects how well the model predicts words in text sequences.
- Human Evaluation: We had a panel of bilingual Somali-English speakers rate the clarity, fluency, and correctness of the AI-generated responses on a 1–5 Likert scale.

B. Implementation Details

Our codebase was primarily in Python, using Hugging Face Transformers for model training. All experiments ran on a GPU-enabled environment (NVIDIA Tesla V100), ensuring efficient fine-tuning even with limited Somali data. For speech recognition experiments, we utilized OpenAI's Whisper API to generate preliminary transcripts, which were then refined by the model in the fine-tuning phase.

(Figure 2 could illustrate the system architecture, including how textual and audio streams feed into the models.)

V. RESULTS

A. Quantitative Analysis

We observed marked improvements when the models were adapted to Somali data (Table 1).

Table 1: Somali Data

Model	BLEU Score	Word Error Rate (WER)	Human Rating (Out of 5)
Baseline (XLM-R)	28.6%	31%	3.1
Fine-Tuned (XLM-R)	47.2%	18%	4.3
Hybrid (mBERT + Augmentation)	42.5%	20%	4.1

- **BLEU Score**: Fine-tuned XLM-R excelled, surpassing the baseline by nearly 20 points. This underscores the impact of directing model capacity toward language-specific data.
- **WER**: We noticed a substantial drop from 31% to 18% in the best-performing model. This improvement suggests a better alignment with Somali phonetic and lexical patterns.
- **Human Ratings**: Participants praised the fine-tuned XLM-R for generating more natural, context-specific Somali text.

B. Qualitative Observations

In pilot conversations, the fine-tuned model displayed improved understanding of colloquial phrases. It also handled some regional expressions better than the baseline, likely reflecting the diversity of our curated dataset. However, occasional errors surfaced around highly specialized domain terms, indicating the model's limited exposure to certain technical or academic jargon in Somali.

VI. DISCUSSION

A. Addressing Dialectal Variations

Although improvements were significant, not all Somali dialects were equally represented. Users from certain regions noted the model did not always grasp nuanced idiomatic expressions. Future versions could incorporate region-specific corpora or adopt domain adaptation strategies to handle these variations more gracefully.

B. Potential for Real-World Applications

- **Educational Tools**: Enhanced Somali conversational AI could facilitate literacy and language-learning applications, particularly in remote areas.
- **Customer Support**: Companies offering digital services in Somalia could integrate such systems to handle queries in Somali.
- **Accessibility for Diaspora**: Many Somali speakers residing abroad can benefit from AI-driven translation and interpretation services for bureaucratic or healthcare-related tasks.

C. Limitations

- **Dataset Size**: Despite our efforts, the overall Somali corpus remains comparatively small.
- **Quality of Synthetic Data**: The back-translated text can occasionally include awkward phrasing or grammatical errors.

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14908879

- **Computational Resources**: Fine-tuning large transformers is expensive, which could be a barrier for many researchers or organizations in lower-income settings.

VII. CONCLUSION AND FUTURE WORK

This study highlights the transformative effect of adapting multilingual models for low-resource languages like Somali. Our experiments reveal that transfer learning, synthetic data augmentation, and careful fine-tuning can collectively reduce performance gaps, enabling far better text and speech processing than baseline models.

A. Key Findings

- Fine-tuned XLM-R consistently outperformed baseline versions, indicating the value of Somali-specific training data.
- Synthetic data augmentation offered a moderate yet meaningful performance boost for mBERT.
- Human evaluators confirmed that improved lexical coverage and contextual understanding directly translated into more coherent conversations.

B. Future Directions

- **Extended Dialect Coverage**: Collect additional transcripts from multiple Somali-speaking communities to create a more balanced dataset.
- **Advanced Data Augmentation Techniques**: Explore generative models (e.g., GPT-like approaches) for more fluent synthetic Somali text.
- **Open-Source Collaboration**: Encourage community-driven efforts to build and maintain a publicly accessible Somali corpus, fostering wider replication and innovation.
- **Cross-Lingual Adaptation**: Investigate whether these methods can be swiftly repurposed for other Cushitic or Afro-Asiatic languages facing similar data scarcity.

Ultimately, bridging the digital divide demands inclusive language technologies that cater to the full tapestry of human languages. By focusing on Somali, we hope to inspire further research that elevates other marginalized languages, ensuring no community is left behind in the AI revolution.

ACKNOWLEDGMENTS

The authors would like to thank the volunteers who provided audio recordings and assisted in manual transcription. Their invaluable contributions helped to diversify the dataset and ensure greater linguistic coverage.

REFERENCES

[1]. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*.

[2]. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). “Language Models are Few-Shot Learners.” *arXiv preprint arXiv:2005.14165*.

[3]. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., … & Stoyanov, V. (2020). “Unsupervised Cross-lingual Representation Learning at Scale.” *arXiv preprint arXiv:1911.02116*.

[4]. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*.

[5]. Fadaee, M., Bisazza, A., & Monz, C. (2017). “Data Augmentation for Low-Resource Neural Machine Translation.” *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*.

[6]. Howard, J. & Ruder, S. (2018). “Universal Language Model Fine-tuning for Text Classification.” *arXiv preprint arXiv:1801.06146*.