Development of a Language Translator Model for Yoruba Language using Bidirectional Long Short-Term Memory Encoder

Jumoke Falilat AJAO[1] Abdulazeez Olorundare Ajao[2] Rafiu Mope ISIAKA[3] Shakirat Ronke YUSUFF[4] Abdullahi YAHAYA[5]

1,3,4,5 Department of Computer Science, Kwara State University, Malete, Nigeria

2 Department of Computer Engineering Technology, Federal Polytechnic Offa, Kwara State, Nigeria.

KWASU Journal of Information, Communication and Technology (KJICT.) 2024, 3(1)
© 2024 Faculty of Information and Communication Technology, Kwara State University, Malete, Nigeria.
URL: http://kjict.ng/

Abstract

Nigeria is a nation of vast ethnic diversity, encompassing a multitude of cultures, languages, and values. These differences often present significant communication barriers among its various ethnic groups. To bring the language barrier, Google has integrated several Nigerian languages into its translation model. This study focuses on the development of an automatic speech recognition (ASR) system for Yoruba, one of Nigeria's indigenous languages. Speech dataset was collected from native Yoruba speakers and subjected to extensive preprocessing to eliminate background noise introduced during the recording process. The preprocessed data was then transformed into a speech spectrogram, serving as input to a bidirectional long short-term memory (BiLSTM) model. The hyperparameters of the BiLSTM model were optimised for improved performance. The translation model was evaluated using metrics such as BLEU, METEOR, and ROUGUE. Results demonstrated significant improvements in the ASR model's ability to accurately transcribe and translate Yoruba language speech. This advancement in translation accuracy highlights the potential for better cross-linguistic communication among Nigeria's ethnic groups, fostering greater inclusivity and understanding. This study contributes to ongoing efforts to incorporate underrepresented languages in global translation models, addressing the challenge of language diversity in multilingual societies.

Keywords: Bidirectional Long Short-Term Memory, Yoruba Language, Language Translator, English Language, Speech Translator

1. Introduction

Communication is imperative in our day-to-day activities to enhance easy interaction, transaction and peaceful coexistence among ourselves. The proliferation of digital technologies has revolutionised communication and information exchange (Imamov & Semenikhina, 2021). However, many languages, particularly those spoken by minority groups, remain underrepresented in the digital realm, limiting their accessibility and preservation. Yoruba is a widely spoken language in West Africa (Forde, 2017; Ajao, Yusuff & Ajao, 2022). The lack of robust language translation models for Yoruba poses significant challenges in areas such as education, business, and cultural exchange.

The development of a language translator model for Yoruba is crucial for several reasons. First, it would enable more effective communication between Yoruba speakers and those who speak other languages, fostering cultural exchange and understanding. Second, it would facilitate access to information and resources in Yoruba, empowering speakers and promoting education and development. Third, it would contribute to the preservation of Yoruba language and culture, as it would be more widely used and accessible in the digital age. Figure 1 illustrates the framework for English-Yoruba translation model. Deep learning-based approaches have significantly enhanced speech recognition accuracy for low-resource languages by leveraging techniques such as pre-trained self-supervised models, which have been shown to improve perfor-

mance even with limited labeled data (Romero et al., 2024). Additionally, the integration of deep learning architectures, such as Bi-LSTM networks combined with convolutional layers, has yielded high accuracy rates (Jang et al., 2020).

These advancements indicate a promising trajectory for improving ASR systems in low-resource settings, addressing the challenges posed by the scarcity of training data and enhancing overall recognition performance (Drexler, 2020). Though, overtime, tremendous efforts have been made by Google and other researchers, but efforts still need to be made to improve the translation accuracy of English to Yoruba.

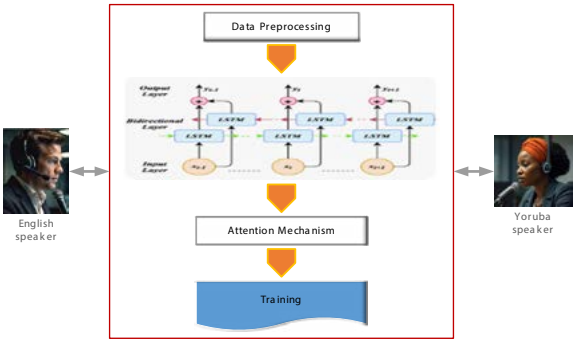


Figure 1. Architecture of English-Yoruba (adapted from Huang, Xu, & Yu, (2015))

This research aims to address the aforementioned challenges by developing a language translator model for Yoruba, using a bidirectional long short-term memory (LSTM) encoder. The specific objectives of this study are to (a) create a neural network-based language translator model capable of translating text from Yoruba to English and vice versa (b) evaluate the performance of the model using standard metrics such as BLEU score, ROUGE score, and METEOR score (c) analyse the impact of hyperparameter tuning on model performance, and (d) explore the potential applications of the model in the medical domain, such as doctor-patient dialogue in healthcare system.

2. Literature Review

Long Short-Term Memory Networks (LSTMs) have revolutionised machine learning and neurocomputing. Their ability to handle long-term dependencies has led to significant advancements in various applications (Van Houdt et al., 2020), including speech recognition (Oruh et al., 2022), machine translation (Rahit et al., 2020) and natural language processing (Setyanto et al., 2022). For instance, Google has leveraged LSTM-based models to improve the accuracy of its speech recognition and translation systems (Shewalkar, Nyavanandi, & Ludwig, 2019). Facebook, too, has extensively utilised LSTMs, processing billions of translations daily (Vathsala & Holi, 2020).

While traditional recurrent neural networks struggled with the vanishing/exploding gradient problem, LSTMs have proven to be more effective. This is due to their unique cell structure, which allows them to maintain long-term memory.

Bi et al. (2021) proposed an intelligent question-answering system that leverages a semantic mapping mechanism to normalise semantic representations. The system further employs a Bi-directional LSTM model with semantic role labeling and positional context to improve question-answering performance.

Direct Speech-to-Speech Translation (S2ST) has gained significant attention in recent years. Lee et al. (2021) introduced a self-supervised discrete speech encoder and S2UT model to improve S2ST performance. Huang et al. (2022) addressed challenges in S2ST by proposing TranSpeech, a model with bilateral perturbation. Popuri et al. (2022) tackled data scarcity in S2ST through self-supervised pre-training and data augmentation techniques.

Neural Machine Translation (NMT) has surpassed traditional Statistical Machine Translation (SMT) models. NMT models, often based on sequence-to-sequence architectures with LSTM or Transformer encoders and decoders, have demonstrated superior performance. Bidirectional LSTMs (BiLSTMs) are particularly effective in capturing long-range dependencies in both directions of a sequence.

Developing language technology for low-resource languages presents unique challenges. Ìyàndá et al. (2015) focused on Yoruba language, collecting and digitising text and speech data to develop a phonemic transcription system. Liu et al. (2021) proposed a unified system for multilingual ASR and LID, addressing the challenges of multilingual speech processing.

The scarcity of parallel corpora and other linguistic resources is a major hurdle for low-resource languages. Complex morphological systems further complicate the development of accurate language translation models.

3. Material and Method

The method employed in this study involved data acquisition, speech preprocessing, BiLSTM model development and the performance evaluation of the translation model. Figure 2 shows the model architecture of the translation model.

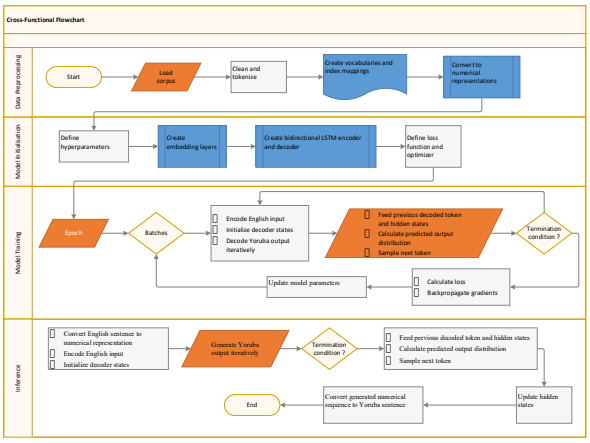


Figure 2: The model architecture of the translation system

Data Acquisition and Preprocessing

The data collection process focuses on the medical domain data; that is, the conversational speech that exists between patients and doctors in the hospital management system. There are various ailments that exist in the medical field. This study focuses on the conversational questions and answers that occur in respiratory tract infection. The conversation was established and certified by a medical Doctor. The consent form was also given to those whose voices were captured. This was done outside the hospital domain in order not to bridge the ethical issue of patient/doctor's conversations. Fifty people were engaged outside the hospital and one person acted like a doctor and other people acted like patients. A Sample of the message in Yoruba and English versions and the recorded speech is given as Table 1.

Table 1: Conversation between Patient and Doctor: The case of Lung Cancer

Conversation	Yoruba	English
Text Conversation	**Dókítà:** Kín ni àìlera tó ń ṣe yín gan-an ? **Aláìsàn:** Ikọ̀ àti àyà dídùn ló ti ń dààmú mi láti bí oṣù mélòó kan sẹ́yìn. Ó ti wá ń kọjá àfaradà nítorí pé ó ń dààmú mi gan-an.	Doctor: What would you say is the main problem? Patient: I've been experiencing a persistent cough and chest pain for a few months now. It's getting worse, and I'm worried.
Speech Conversation recordings	Youba speech.opus	English Speech.opus

The translation model consists of three modules:

1. Speech Input (ASR Module): The audio input is recorded in Yoruba and English as parallel corpora. Yoruba conversations are converted to spectrograms using the Mel-Spectrogram technique for feature extraction. Finally, a Wav2Vec2 pre-trained model is used to convert speech to text, feeding into the translation pipeline.
2. Translation Model: The extracted Yoruba text is fed into a BiLSTM-based translation model, which translates it into English. BiLSTM allows the model to capture the context from both forward and backward directions, providing high-quality translation results.
3. Text-to-Speech (TTS) Module: The translated English text is synthesised into audio, using Google TTS (gTTS) to generate the final speech output.

Speech Preprocessing: Silent Gap Removal

The audio preprocessing stage addresses the issue of silent gaps present in the conversational audio recordings, which could negatively impact translation performance, if not handled properly. A sliding win-

dow approach is applied to remove these gaps. The process is detailed below:

i. Sliding Window Setup

a. A sliding window is applied to the absolute intensities of the audio waveform.

b. The window starts from both the beginning and end of the waveform and moves towards the centre.

ii. Local Mean Calculation:

a. For each window segment, the local mean intensity is compared against the global mean intensity of the entire audio clip.

iii. Silent Segment Detection:

a. If the local mean is below the global mean, the segment is classified as insignificant and removed.

b. This ensures that only meaningful portions of the audio are retained for translation.

Figure 3 shows example of an audio waveform before and after silent gap removal, demonstrating how this technique improves the signal-to-noise ratio. Figure 4 illustrates the loss function during training and testing.

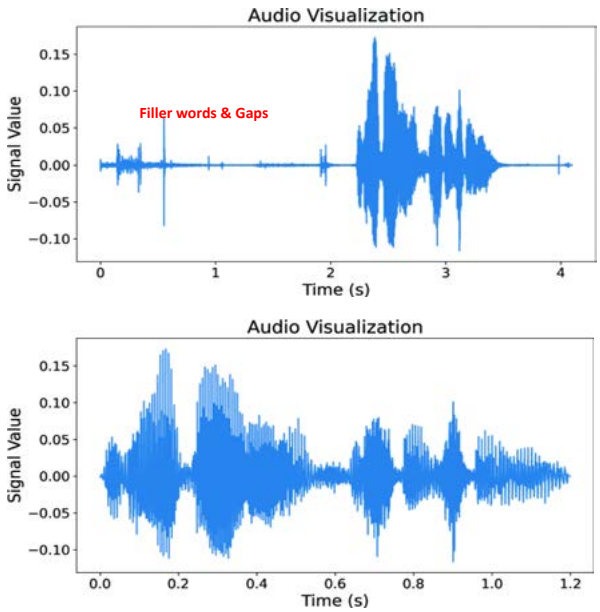


Figure 3: Audio waveform of patient-doctor conversation before and after clipping gaps

Bidirectional Long Short-Term Memory (LSTM) (BiLSTM) Model Components

The BiLSTM model forms the core of the translation system, enhanced with an attention mechanism to align input and output sequences effectively.

i. Input Layer: Input consists of 128 features per frame, extracted from the Mel-spectrogram of the audio signal, representing relevant time-frequency information.

ii. Hidden Layer: 2-layer bidirectional LSTM, each with 512 hidden units

iii. Optimiser and Loss Function: Adam with a learning rate of 0.001 and Cross-Entropy Loss

iv. Output Layer: 128 output size (corresponding to the predicted English text spectrogram)

v. Training Process: 100 epochs with early stopping are used to prevent overfitting by monitoring the validation loss and a Batch size of 32.

The Wav2Vec2 ASR module and Google TTS API were integrated to handle the conversion between audio and text at both ends of the system.

Integration with Wav2Vec2 ASR and Google TTS API

The Wav2Vec2 ASR module and Google Text-to-Speech (TTS) API are integrated to handle the conversion between audio and text at both ends of the system.

i. Wav2Vec2 ASR: Converts the input Yoruba speech into Yoruba text, serving as the initial step in the translation process. The transcriptions are passed to the BiLSTM model for translation into English text.

ii. Google TTS API: The English text generated by the BiLSTM model is converted into synthesised English speech, completing the speech-to-speech translation cycle.

Figure 4 shows the training and testing loses for the BiLStM model

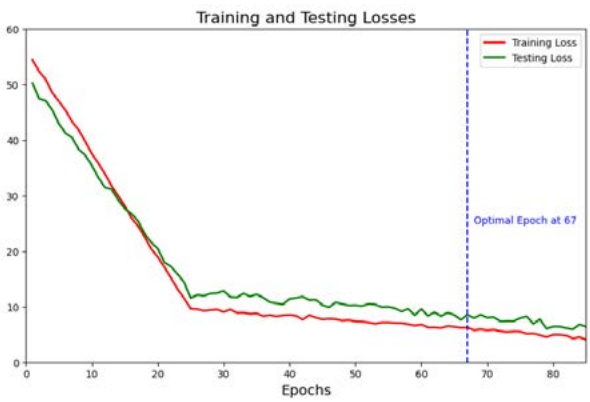


Figure 4: Training and Testing loss of the BiLSTM attention-based model

Evaluation Metrics

The performance evaluation used for the evaluation of speech-to-speech language translation Model for Yoruba to English is given in equation 1, 2 and 3.

BLEU (Bilingual Evaluation Understudy) Score

$$BP = \begin{cases} 1 & if\ c > r \\ e^{(1-\frac{r}{c})} & if\ c \leq r \end{cases}$$

Then,

$$BLEU = BP.exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

(ii) ROUGE (Recall-Oriented Understudy for Gisting Evaluation) Score

ROUGE-1 F1-score = 2 * (precision * recall) / (precision + recall)

(iii) METEOR (Metric for Evaluation of Translation with Explicit Ordering)

$$Fmean = \frac{10PR}{R + 9P}$$

Where Penalty is computed as

$$Penalty = 0.5 * \left(\frac{Number\ of\ Chunks}{Number\ of\ Unigrams + matched}\right)^3$$

4. Result and Discussion

The model's performance on the test set, after hyperparameter tuning, is summarised in Table 2.

Table 2: Model performance

Evaluation Metric	Score
BLEU Score	0.487
ROUGE-1 (Recall)	0.563
METEOR	0.438

These results indicate that the model achieved strong performance on the test set. The BLEU score of 48.7 suggests that the model accurately translated most phrases, with a good match between predicted and reference translations. The use of the attention mechanism in the BiLSTM model likely contributed to this success, enabling the model to produce coherent translations. The ROUGE score of 56.3% highlights the model's ability to capture word-level matches and preserve sentence structure, particularly in terms of longest common subsequences. The METEOR score of 43.8% reflects the model's capacity to capture word order, synonyms, and morphological variations, which is particularly important for the Yoruba language due to its complex tonal and morphological structure

Impact of Hyperparameter Tuning

Our hyperparameter tuning efforts led to significant improvements in the model's performance. Specifically:

i. Learning Rate Adjustment: Reducing the learning rate from 0.01 to 0.001 enhanced convergence and reduced the validation set loss by 12.5%, indicating that the model learned more effectively from the training data.

ii. BiLSTM Layer Depth: Increasing the BiLSTM encoder depth from 1 to 2 layers improved the model's ability to understand context, resulting in a 5-point increase in BLEU score. This suggests that the additional layer allowed the model to better capture complex relationships between input elements.

iii. Attention Mechanism: Incorporating Bahdanau Attention significantly enhanced translation fluency by enabling the model to focus on relevant parts of the input sequence, particularly for complex sentence structures. This improvement is likely due to the attention mechanism's ability to selectively weight input elements and emphasise key aspects of the input sequence.

Error Analysis

Despite the promising results, our model encountered several common translation challenges in the Yoruba language:

i. Tone and Homonym Resolution: The model struggled with the tonal nature of Yoruba, where the meaning of words can change based on tonal inflections. For example, the model misinterpreted the tone of words, resulting in the incorrect contexts, such as translating "àìsàn" as "suffering" instead of "illness".

ii. Long Sentence Translation: The model occasionally struggled with long and complex sentences, leading to incomplete translations or missing elements. For instance, the translation of "Ó máa ń jí mi síle̩ lóru" (It keeps me up at night) was sometimes translated as "I stay awake", losing the nuance of the statement.

iii. Named Entity Handling: The model had dif-

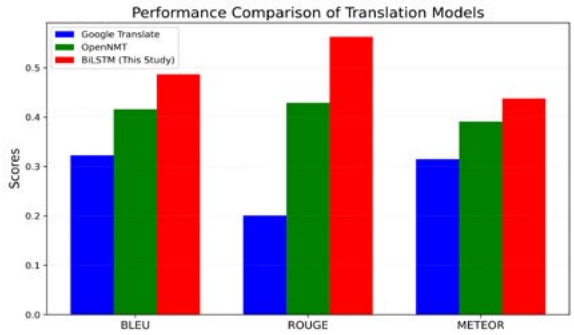
ficulty with medical terms or proper nouns, such as "àyà dídùn" (chest pain), which were sometimes incorrectly translated.

Comparison with Existing Yoruba Language Models

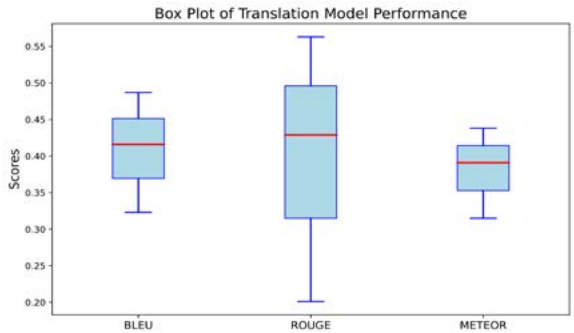
We evaluated the performance of our BiLSTM model against two existing Yoruba language models:

Table 2: Performance Comparison

Model	BLEU	ROUGE	METEOR
Google Translate (Yoruba-English)	0.323	0.201	0.315
OpenNMT Pre-trained (Yoruba)	0.416	0.429	0.391
BiLSTM (This Study)	0.487	0.563	0.438



(a)



(b)

Figure 5: Result performance of trained model

Our proposed BiLSTM model outperformed Google Translate in terms of BLEU and METEOR scores, achieving a 50.5% increase in BLEU score (from 0.323 to 0.487) and a 39.0% increase in METEOR score (from 0.315 to 0.438). This improvement can be attributed to the use of domain-specific data and the attention mechanism employed in our model. Compared to OpenNMT's pre-trained Yoruba model, our BiLSTM model excelled across all metrics, particularly for longer sentences and medical-related phrases, with a 16.7% increase in BLEU score (from 0.416 to 0.487), a 31.5% increase in ROUGE score (from 0.429 to 0.563), and a 12.0% increase in METEOR score (from 0.391 to 0.438). The bar plot in Figure 5(a) illustrates the performance comparison of the translation models based on BLEU, ROUGE, and METEOR scores. Each colour corresponds to a specific model, as indicated in the legend. Notably, our BiLSTM model achieves the highest scores across all metrics. Figure 5(b) provides a more detailed view of the performance comparison, with the red line representing the median score for each metric and the whiskers indicating the range of scores across the three models.

5. Conclusion and Recommendation

This study successfully developed a speech-to-speech translation model for the Yoruba language, leveraging a BiLSTM encoder-decoder architecture with an attention mechanism. The model achieved notable improvements in translation performance, surpassing existing translation systems for Yoruba with BLEU, ROUGE, and METEOR scores of 48.7, 56.3, and 43.8, respectively. The incorporation of medical conversational data proved instrumental in enabling more accurate translations, demonstrating the value of domain-specific training and the potential for tailored translation systems in specialised fields. The success of this model highlights the importance of addressing the unique challenges of low-resource languages, such as Yoruba, and the need for innovative approaches to bridge the gap in translation capabilities. Future work will focus on addressing the remaining challenges, including handling tonal variations more effectively, which is critical for accurately capturing the nuances of the Yoruba language.

Acknowledgements

The authors would like to express their sincere gratitude to the Tertiary Education Trust Fund (TETFund) for its financial support with NRF Number TETF/ES/DR&D-CE/NRF2023/SET1/ICT/00085/VOL1, which made this research possible. We also extend our thanks to Kwara State University, Malete for providing the necessary research facilities and academic environment.

References

Ajao, J. F., Yusuff, S., & Ajao, A. O. (2022). Yorùbá character recognition system using convolutional recurrent neural network. *Black Sea Journal of Engineering and Science*, 5(4), 151-157.

Bi, M., Zhang, Q., Zuo, M., Xu, Z., & Jin, Q. (2021). Bi-directional long short-term memory model with semantic positional attention for the question answering system. *Transactions on Asian and Low-Resource Language Information Processing*, 20(5), 1-13.

Drexler, J. F. (2020). Improving end-to-end neural network models for low-resource automatic speech recognition (Doctoral dissertation, Massachusetts Institute of Technology)

Forde, D. (2017). The Yoruba-Speaking Peoples of South-Western Nigeria: Western Africa Part IV. Routledge.

Huang, R., Liu, J., Liu, H., Ren, Y., Zhang, L., He, J., & Zhao, Z. (2022). Transpeech: Speech-to-speech translation with bilateral perturbation. arXiv preprint arXiv:2205.12523.

Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991.

Imamov, M., & Semenikhina, N. (2021). The impact of the digital revolution on the global economy. *Linguistics and Culture Review*, 968-987

Ìyàndá, A. R., Odéjobí, O. A., Soyoye, F. A., & Akinadé, O. O. (2014). Development of grapheme-to-phoneme conversion system for yorùbá text-to-speech synthesis. INFOCOMP Journal of Computer Science, 13(2), 44-53.

Jang, B., Kim, M., Harerimana, G., Kang, S. U., & Kim, J. W. (2020). Bi-LSTM model to increase accuracy in text classification: Combining Word2vec CNN and attention mechanism. *Applied Sciences*, 10(17), 5841.

Liu, D., Xu, J., Zhang, P., & Yan, Y. (2021). A unified system for multilingual speech recognition and language identification. *Speech Communication*, 127, 17-28.

Oruh, J., Viriri, S., & Adegun, A. (2022). Long short-term memory recurrent neural network for automatic speech recognition. *IEEE Access*, 10, 30069-30079.

Popuri, S., Chen, P. J., Wang, C., Pino, J., Adi, Y., Gu, J., ... & Lee, A. (2022). Enhanced direct speech-to-speech translation using self-supervised pre-training and data augmentation. arXiv preprint arXiv:2204.02967.

Rahit, K. T. H., Nabil, R. H., & Huq, M. H. (2020). Machine translation from natural language to code using long-short term memory. In Proceedings of the Future Technologies Conference (FTC) 2019: Volume 1 (pp. 56-63). Springer International Publishing.

Romero, M., Gómez-Canaval, S., & Torre, I. G. (2024). Automatic Speech Recognition Advancements for Indigenous Languages of the Americas. *Applied Sciences*, 14(15), 6497.

Setyanto, A., Laksito, A., Alarfaj, F., Alreshoodi, M., Oyong, I., Hayaty, M., ... & Kurniasari, L. (2022). Arabic language opinion mining based on long short-term memory (LSTM). *Applied Sciences*, 12(9), 4140.

Shewalkar, A., Nyavanandi, D., & Ludwig, S. A. (2019). Performance evaluation of deep neural networks applied to speech recognition: RNN, LSTM and GRU. *Journal of Artificial Intelligence and Soft Computing Research*, 9(4), 235-245.

Van Houdt, G., Mosquera, C., & Nápoles, G. (2020). A review on the long short-term memory model. *Artificial Intelligence Review*, 53(8),

5929-5955.

Vathsala, M. K., & Holi, G. (2020). RNN based machine translation and transliteration for Twitter data. *International Journal of Speech Technology*, 23(3), 499-504.