Transformer-based Model for Word Level Language Identification in Code-mixed Kannada-English Texts

Atnafu Lambebo Tonja[1], **Mesay Gemeda Yigezu**[2], **Olga Kolesnikova**[3], **Moein Shahiki Tash**[4], **Grigori Sidorov**[5], **Alexander Gelbukh**[6]

Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC), Mexico City, Mexico

{ [1]alambedot2022, [2]mgemedak2022, [5]sidorov, [6]gelbukh}@cic.ipn.mx

{ [3]kolesolga, [4] moein.tash}@gmail.com

Abstract

Using code-mixed data in natural language processing (NLP) research currently gets a lot of attention. Language identification of social media code-mixed text has been an interesting problem of study in recent years due to the advancement and influences of social media in communication. This paper presents the Instituto Politécnico Nacional, Centro de Investigación en Computación (CIC) team's system description paper for the CoLI-Kanglish shared task at ICON2022. In this paper, we propose the use of a Transformer based model for word-level language identification in code-mixed Kannada English texts. The proposed model on the CoLI-Kenglish dataset achieves a weighted F1-score of 0.84 and a macro F1-score of 0.61.

1 Introduction

In recent years, language identification of social media text has been a fascinating research topic (Ansari et al., 2021). Social media platforms have become more integrated in this digital era and have impacted various people's perceptions of networking and socializing (Tonja et al., 2022c). This influence allowed different users to communicate via various social media platforms using a mix of texts. NLP technology has advanced rapidly in many applications, including machine translation (Tonja et al., 2022b; Yigezu et al., 2021; Tonja et al., 2021), abusive comment detection (Balouchzahi et al., 2022b), fake news detection(Arif et al., 2022; Truică et al., 2022), aggressive incident detection (Tonja et al., 2022a), hope speech detection (Gowda et al., 2022), and others. However, numerous tools have not yet been created for languages with limited resources or languages with code-mixed data.

Code-mixing is the use of linguistic units—words, phrases, and clauses—at the sentence or word level from various languages. In casual communication, such as social media, it is typically seen. We have access to an enormous amount of code-mixed data because of the various social media platforms that allow individuals to communicate (Sutrisno and Ariesta, 2019). As a result, automatic language recognition at the word level has become an essential part of analyzing noisy content in social media. It would help with the automated analysis of content generated on social media. Currently, in the area of NLP, different researchers are developing different NLP applications in code-mixed datasets. Some of the applications are code-mixed sentiments analysis (Balouchzahi et al., 2021b), code-mixed offensive language identification (Balouchzahi et al., 2021a), etc. We took part in the **Kanglish shared task** (Balouchzahi et al., 2022a), which aims to identify language at the word level from code-mixed data for Kannada-English texts. For word-level code-mixed language identification tasks, we used Transformer -based (Vaswani et al., 2017) pre-trained language models (PLMs). Our transformer-based model consists of BERT (Devlin et al., 2018) and its three variants. We used PLMs and LSTM models for this word-level language identification task.

This paper discusses a Transformer-based model for word-level language identification in code-mixed Kannada-English texts for the Kanglish shared task. The paper is organized as follows: Section 2 describes past work related to this study, section 3 gives an overview of the dataset and its statistics, section 4 explains the methodology adopted in this study including the algorithms, section 5 emphasizes on the experimental results and descriptions. Finally, Section 6 concludes the paper.

2 Related Work

Currently, solving NLP problems in code-mixed data is getting attention from many researchers. For word-level language identification in code-

Proceedings of the 19th International Conference on Natural Language Processing (ICON), pages 18 - 24

mixed text, different researchers have suggested various models. Chittaranjan et al. (2014) proposed a Conditional Random Fields (CRF)- based system for word level language identification of code-mixed text for four language pairs, namely, English-Spanish (En-Es), English-Nepali (En-Ne), English-Mandarin (En-Cn), and Standard Arabic-Arabic (Ar-Ar) dialects. The authors explored various token levels and contextual features to build an optimal CRF using the provided training data. The proposed system performed more or less consistently, with accuracy ranging from 80% to 95% across four language pairs.

Gundapu and Mamidi (2020) also proposed a CRF based model for word-level language identification in English-Telugu code-mixed data. The authors used feature extraction as the main task for the proposed model. They used POS-tags, length of the word, prefix and suffix of focus word, numeric digit, special symbol, capital letter, and character N-grams (Uni-, Bi-, Trigrams of words) as features. The proposed CRF-based model had an F1-score of 0.91.

A Support Vector Machines (SVM)-based model for word level language identification of Tamil-English code-mixed text in social media is proposed by Shanmugalingam et al. (2018). The authors used dictionaries, double consonants, and term frequency to identify features. The proposed SVM model with a linear kernel gave 89.46% accuracy for the language identification system for Tamil-English code-mixed text at the word level.

Ansari et al. (2021) proposes transfer learning and fine-tuning BERT models for language identification of Hindi-English code-mixed tweets. The authors used data from Hindi-English-Urdu code-mixed text for language pre-training and Hindi-English code-mixed for subsequent word-level language classification. The authors first pre-trained Hindi-English-Urdu code-mixed text using BERT and fine-tuned the trained model in downstream Hindi-English code-mixed word-level language classification. Their proposed model for Hindi-English code-mixed language identification, both pre-training and fine-tuning with code-mixed text, gives the best F1-score of 0.84 as compared to their monolingual counterparts.

3 Data

During the experimental phase, we used the CoLI-Kenglish dataset (Hosahalli Lakshmaiah et al., 2022) which consists of English and Kannada words in Roman script and are grouped into six major categories, namely, Kannada (kn), English (en), Mixed-language (en-kn), Name, Location and Other. Table ?? shows some samples from the dataset used for training.

	word	**tag**
0	anusthu	kn
1	woww	en
2	staying	en
3	near	en
4	hostel	en
5	confirmed	en
6	faith	en
7	linked	en
8	gtila	kn
9	germany	en

Table 1: Training samples

3.1 Dataset Statistics

Figures 1 and 2 depict the training and test data distribution statistics with their assigned tags. The training dataset is slightly imbalanced: 43.9% of the words were labeled as *kn*, 30% were labeled as *en*, 9.28% were labeled as *en-kn*, 4.76% were labeled as *name*, 0.68% were labeled as *location* and 11.2% were labeled as other. This shows that approximately 73% of the training dataset was labeled as *kn* and *en*. Similarly, in the test dataset, words tagged as *en* and *kn* take a higher number than the rest of the dataset.

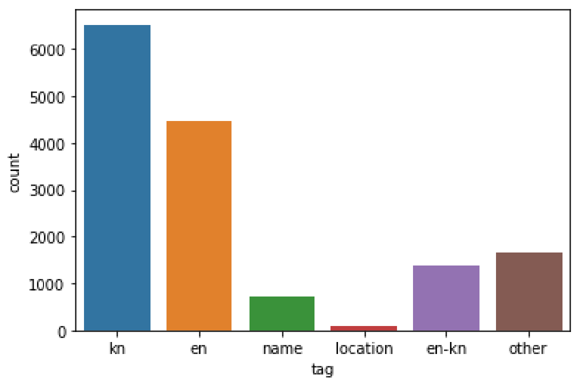

Figure 1: Training data distribution with tags

4 Methodology

This section presents a description of the data preprocessing, methodology, and models used in this

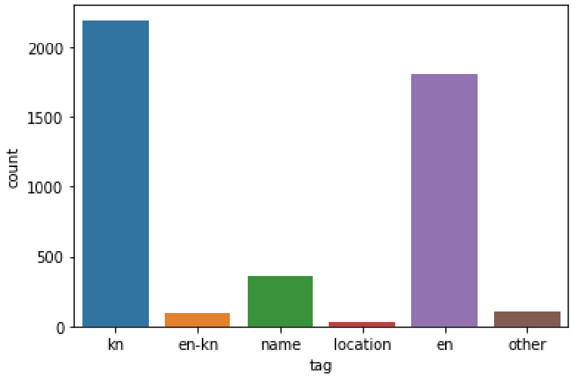

Figure 2: Test data distribution with tags

work. We used Transformer based pre-trained language models (PLMs) with the combination of the LSTM model for word level language identification in Kannada-English code-mixed text. We used PLMs in the embedding layer of the LSTM model layer.

4.1 Pre-processing

Pre-processing is one of the preliminary steps in training NLP tasks, with the aim of making the training data suitable during the training phase. The dataset provided by the organizers for this task has passed the basic pre-processing steps, and we carried out one pre-processing step to prepare the training data during the experimental phase. We applied label encoding to tags, to convert the tags into a numeric form. As discussed in section 3, the dataset contains six tags (*kn, en, en-kn, name, location* and *other*). We converted these tags into numeric values using one-hot encoding.

4.2 Proposed Experimental Architecture

Figure 3 shows the experimental architecture of our Transformer-based model for word level language identification in code-mixed Kannada-English texts. As shown in Figure 3, our experimental architecture consists of five steps:

- **Step 1** - preparing labelled data for training, the data set contains **words** and their **tags** as discussed in section 3.
- **Step 2** - we converted the tags into a numeric machine-readable form.
- **Step 3** - after label encoding the representation for each token is fed to transformer layers to obtain contextualized tokens using PLMs.

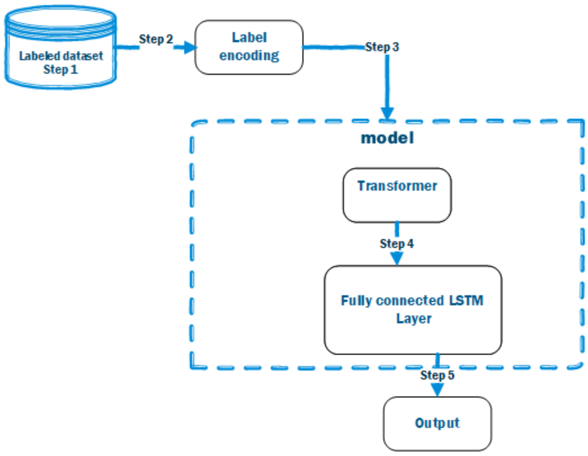

Figure 3: Experimental architecture for Transformer-based model for word level language identification in code-mixed Kannada-English texts

- **Step 4** - the embeddings obtained in step-3 are fed into LSTM model to obtain their corresponding language tag.

We used the following pre-trained language models (PLMs) in the embedding layer of the LSTM model for our experiment.

- BERT (Devlin et al., 2018) - stands for Bidirectional Encoder Representations from Transformers. As the name suggests, it is a way of learning representations of a language that uses a transformer, specifically, the encoder part of the transformer.
- mBERT (Devlin et al., 2018) - is a Multilingual BERT, it provides sentence representations for 104 languages, which are useful for many multi-lingual tasks. Previous work probed the cross-linguality of mBERT using zero-shot transfer learning on morphological and syntactic tasks.
- XLM-R (Conneau et al., 2019) - uses self-supervised training techniques to achieve state-of-the-art performance in cross-lingual understanding, a task in which a model is trained in one language and then used with other languages without additional training data.
- RoBERTa (Liu et al., 2019) is a self-supervised transformers model that was trained on a large corpus of English data. This means it was pre-trained on raw texts only, with no human labeling in any way (which is

why it can use lots of publicly available data) and an automatic process to generate inputs and labels from those texts.

Table 2 shows models used in our experiments and their parameters.

5 Experiments and Results

This section presents the description of the experimental setups, training parameters, results, and analysis. We conducted four experiments by replacing embedding layers with different pre-trained language models, the results are presented in section 5.2.

5.1 Experiments

We used Google colab [1] for GPU support with the Python programming language. Sci-kit-learn [2] and Keras [3] (with TensorFlow backend) were utilized for the LSTM model, for PLMs we used Hugging Face [4] transformer libraries. We used PLMs for embedding and the LSTM model as the classifier, To optimize the model, we used an Adam optimizer with a batch size of 64 and a learning rate of 0.0001. We used the maximum number of epochs of 30, with early stopping based on the performance of the validation set. We also used a dropout of 0.2 to regularize the model.

We added a batch normalization layer to speed up training, and make learning easier, and a fully-connected output layer with a SoftMax function so that a probabilistic output of all tags for language identification would be produced. For further information, all the parameters and their summaries are depicted in Figure 4. Figure 4 shows our proposed model summary for word-level language identification in code-mixed Kannada-English texts.

5.2 Results

Table 3 depicts the overall results (official) of four experiments conducted in this work. From four experiments, using *bert-base-uncased* in the embedding layer with the LSTM model out-performs other pre-trained languages models used in the embedding layer with the LSTM model with a weighted score of 0.85 precision, 0.84 recall, 0.84 F1-scores and a micro score of 0.62 precision, 0.62 recall, 0.61 F1-scores.

[1]https://colab.research.google.com/

[2]https://scikit-learn.org/stable/

[3]https://keras.io/

[4]https://huggingface.co/

The official rank of the top three teams participating in the shared task of word-level language identification in code-mixed Kannada-English texts is shown in Table 4. As shown in Table 4 our model ranked second in overall results among all participant teams.

Figures 5 and 6 display the training and, validation losses, training, and validation accuracy of the BERT-based approach for code-mixed language identification tasks. It is seen that the BERT-based model's training loss decreases and stabilizes at a specific point, but the validation loss is not as stable as the training loss. This shows that the more specialized the model becomes with training data, the worse it is able to generalize to new data, resulting in an increase in generalization error.

The above result demonstrates that transformer-based models can give promising results when applied to NLP tasks like word-level language identification in code-mixed texts without considering any linguistic features.

6 Conclusion

In this paper, we explored the application of BERT-based pre-trained language models to identify languages at the word level in code-mixed data for Kannada-English texts. Pre-trained models with a combination of the LSTM model and a BERT-based model outperformed the others and have shown promising results in identifying languages in code-mixed Kannada-English texts. Our team achieved the second place in CoLI-Kanglish: word-level language identification in the code-mixed Kannada-English texts competition.

Acknowledgements

The work was done with partial support from the Mexican Government through the grant A1S-47854 of CONACYT, Mexico, grants 20220852, 20220859, and 20221627 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

Model	Transformer blocks	Hidden layer size	Self-attention heads	#Parameters
bert-base-uncased	12	768	12	110M
bert-base-multilingual-uncased	12	768	12	110M
xlm-roberta-large	24	1024	16	355M
roberta-base	12	768	12	110M

Table 2: Transformers used in this paper and their parameters

```
Model: "model"
__________________________________________________________________________________________________
 Layer (type)                   Output Shape         Param #     Connected to
==================================================================================================
 input_ids (InputLayer)         [(None, 10)]         0           []

 attention_mask (InputLayer)    [(None, 10)]         0           []

 tf_bert_model (TFBertModel)    TFBaseModelOutputWi  109482240   ['input_ids[0][0]',
                                thPoolingAndCrossAt               'attention_mask[0][0]']
                                tentions(last_hidde
                                n_state=(None, 10,
                                768),
                                 pooler_output=(Non
                                e, 768),
                                 past_key_values=No
                                ne, hidden_states=N
                                one, attentions=Non
                                e, cross_attentions
                                =None)

 lstm (LSTM)                    (None, 128)          459264      ['tf_bert_model[0][0]']

 batch_normalization (BatchNorm  (None, 128)         512         ['lstm[0][0]']
 alization)

 dense (Dense)                  (None, 768)          99072       ['batch_normalization[0][0]']

 activation (Activation)        (None, 768)          0           ['dense[0][0]']

 dense_1 (Dense)                (None, 768)          590592      ['activation[0][0]']

 dropout_37 (Dropout)           (None, 768)          0           ['dense_1[0][0]']

 outputs (Dense)                (None, 6)            4614        ['dropout_37[0][0]']

==================================================================================================
Total params: 110,636,294
Trainable params: 1,153,798
Non-trainable params: 109,482,496
__________________________________________________________________________________________________
```

Figure 4: Proposed model summary

Model	Weighted Score			Macro score		
	P	R	F1-score	P	R	F1-score
bert-base-multilingual-uncased	0.83	0.82	0.82	0.62	0.57	0.57
xlm-roberta-large	0.84	0.85	0.84	0.64	0.59	0.61
roberta-base	0.83	0.8	0.81	0.63	0.55	0.52
bert-base-uncased	**0.85**	**0.84**	**0.84**	**0.62**	**0.62**	**0.61**

Table 3: Performance of our models on the test set (official results)

Rank	Team name	Weighted Score			Macro score		
		P	R	F1-score	P	R	F1-score
1	tiya1012	0.87	0.85	0.86	0.67	0.61	0.62
2	**Our team**	**0.85**	**0.84**	**0.84**	**0.62**	**0.62**	**0.61**
2	Habesha	0.85	0.83	0.84	0.66	0.6	0.61
3	lidoma	0.83	0.83	0.83	0.64	0.56	0.58

Table 4: Official rank of top 3 teams

References

Mohd Zeeshan Ansari, MM Sufyan Beg, Tanvir Ahmad, Mohd Jazib Khan, and Ghazali Wasim. 2021. Language identification of hindi-english tweets using code-mixed bert. In *2021 IEEE 20th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*, pages 248–252. IEEE.

Muhammad Arif, Atnafu Lambebo Tonja, Iqra Ameer, Olga Kolesnikova, Alexander Gelbukh, Grigori

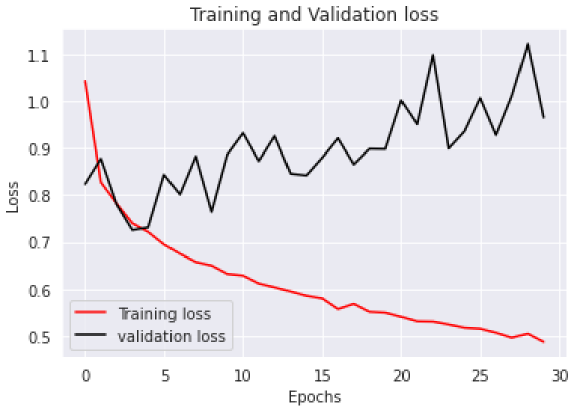

Figure 5: Training and validation loss of BERT based approach

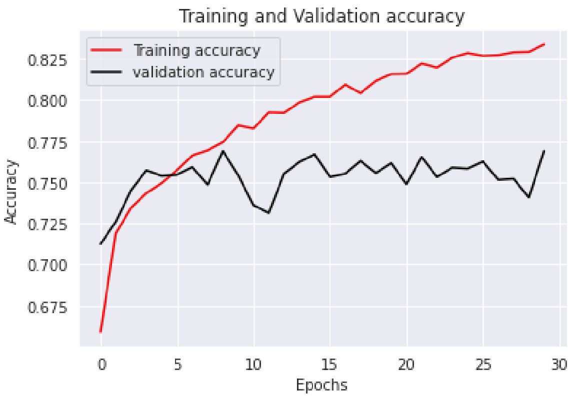

Figure 6: Training and validation accuracy of BERT based approach

Sidorov, and AG Meque. 2022. Cic at checkthat! 2022: multi-class and cross-lingual fake news detection. *Working Notes of CLEF*.

F Balouchzahi, S Bashang, G Sidorov, and HL Shashirekha. 2021a. Comata oli-code-mixed malayalam and tamil offensive language identification. In *Working Notes of FIRE 2021-Forum for Information Retrieval Evaluation (Online). CEUR*.

Fazlourrahman Balouchzahi, Sabur Butt, Asha Hagde, Noman Ashraf, Shashirekha Hosahalli Lakshmaiah, Grigori Sidorov, and Alexander Gelbukh. 2022a. Overview of CoLI-Kanglish: Word Level Language Identification in Code-mixed Kannada-English Texts at ICON 2022. In *19th International Conference on Natural Language Processing Proceedings*.

Fazlourrahman Balouchzahi, Anusha Gowda, Hosahalli Shashirekha, and Grigori Sidorov. 2022b. Mucic@ tamilnlp-acl2022: Abusive comment detection in tamil language using 1d conv-lstm. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 64–69.

Fazlourrahman Balouchzahi, Hosahalli Lakshmaiah Shashirekha, and Grigori Sidorov. 2021b. Cosad-code-mixed sentiments analysis for dravidian languages. In *CEUR Workshop Proceedings*, volume 3159, pages 887–898. CEUR-WS.

Gokul Chittaranjan, Yogarshi Vyas, Kalika Bali, and Monojit Choudhury. 2014. Word-level language identification using crf: Code-switching shared task report of msr india system. In *Proceedings of The First Workshop on Computational Approaches to Code Switching*, pages 73–79.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Anusha Gowda, Fazlourrahman Balouchzahi, Hosahalli Shashirekha, and Grigori Sidorov. 2022. Mucic@ lt-edi-acl2022: Hope speech detection using data re-sampling and 1d conv-lstm. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 161–166.

Sunil Gundapu and Radhika Mamidi. 2020. Word level language identification in english telugu code mixed data. *arXiv preprint arXiv:2010.04482*.

Shashirekha Hosahalli Lakshmaiah, Fazlourrahman Balouchzahi, Anusha Mudoor Devadas, and Grigori Sidorov. 2022. CoLI-Machine Learning Approaches for Code-mixed Language Identification at the Word Level in Kannada-English Texts. *acta polytechnica hungarica*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Kasthuri Shanmugalingam, Sagara Sumathipala, and Chinthaka Premachandra. 2018. Word level language identification of code mixing text in social media using nlp. In *2018 3rd International Conference on Information Technology Research (ICITR)*, pages 1–5. IEEE.

Bejo Sutrisno and Yessika Ariesta. 2019. Beyond the use of code mixing by social media influencers in instagram. *Advances in Language and Literary Studies*, 10(6):143–151.

Atnafu Lambebo Tonja, Muhammad Arif, Olga Kolesnikova, Alexander Gelbukh, and Grigori Sidorov. 2022a. Detection of aggressive and violent incidents from social media in spanish using pre-trained language model. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022), CEUR Workshop Proceedings. CEUR-WS. org*.

Atnafu Lambebo Tonja, Olga Kolesnikova, Muhammad Arif, Alexander Gelbukh, and Grigori Sidorov. 2022b. Improving neural machine translation for low resource languages using mixed training: The case of ethiopian languages. In *Mexican International Conference on Artificial Intelligence*, pages 30–40. Springer.

Atnafu Lambebo Tonja, Olumide Ebenezer Ojo, Mohammed Arif Khan, Abdul Gafar Manuel Meque, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2022c. Cic nlp at smm4h 2022: a bert-based approach for classification of social media forum posts. In *Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task*, pages 58–61.

Atnafu Lambebo Tonja, Michael Melese Woldeyohannis, and Mesay Gemeda Yigezu. 2021. A parallel corpora for bi-directional neural machine translation for low resourced ethiopian languages. In *2021 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, pages 71–76. IEEE.

Ciprian-Octavian Truică, Elena-Simona Apostol, and Adrian Paschke. 2022. Awakened at checkthat! 2022: fake news detection using bilstm and sentence transformer. *Working Notes of CLEF*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Mesay Gemeda Yigezu, Michael Melese Woldeyohannis, and Atnafu Lambebo Tonja. 2021. Multilingual neural machine translation for low resourced languages: Ometo-english. In *2021 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, pages 89–94. IEEE.