# Bilingual Word-Level Language Identification for Omotic Languages

Mesay Gemeda Yigezu [1], Girma Yohannis Bade [2], Atnafu Lambebo Tonja [3],
Olga Kolesnikova [4], Grigori Sidorov [5], Alexander Gelbukh [6],
{mgemedak2022[1], sidorov[5], gelbukh[6]}@cic.ipn.mx
{girme2005[2], atnafuatx[3], kolesolga[4]}@gmail.com

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

**Abstract.** Language identification is the task of determining the languages for a given text. In many real-world scenarios, text may contain more than one language, particularly in multilingual communities. Bilingual Language Identification (BLID) is the task of identifying and distinguishing between two languages in a given text. This paper presents BLID for languages spoken in the southern part of Ethiopia, namely Wolaita and Gofa. The presence of words' similarities and differences between the two languages makes the language identification task challenging. To overcome this challenge, we employed various experiments on various approaches. Then, the combination of the Bert-based pre-trained language model and LSTM approach performed better, with an F1-score of 0.72 on the test set. As a result, the work will be effective in tackling unwanted social media issues and providing a foundation for further research in this area.

**Keywords:** Keywords: LID · Wolayta and Gofa language · Omotic language · bilingual LID · language identification

## 1 Introduction

The process of automatically identifying the language(s) present in a document based on the content of the document is known as Language identification(LID) [1]. It was typically assumed that every text is written for one's known languages [2]. The task of identifying the most similar languages from the set of bilingual languages data sets is possible through LID approaches.

Over the past few decades, a fair amount of research has been done on the challenge of determining the language in which a certain text was written. In both academic and professional settings, automated methods for identifying written languages are frequently applied to spoken and written source materials. Despite this universal acceptance, from the point of [3], the research in the identification of written languages indicates a number of unanswered topics that are ready for additional study. One of the crucial steps in a natural language processing (NLP) challenge is LID. Because it might attempt to predict the natural language's texts

to identify one language from another. Prior to working on any NLP tasks such as sentiment analysis or translation, it is important to understand the language of the text. For instance, the Google Translator application has a menu called "Detect Language", because before Google translates a sentence, it must first determine the language [4].

Now a day many social media users leave their ideas with a small text, such as a tweet, headline, or comment on various issues but may not have enough information to identify the language accurately and may not be recognized or supported by LID tools. A single sentence can belong to more than one language, and this can cause difficulties in LID  [5].

From the best knowledge of researchers, there is no work conducted on LID for the Omotic language family of Ethiopian languages. Since specially, Ethiopian Omotic language groups use Latin scripts to write their languages and it's quite similar for those all  [6], identifying one's language from others needs giving high attention to catch up a little difference that might come from spoken accents. Therefore in this paper, we focus on bilingual LID for highly interrelated and geographically bordered Omotic languages(Wolayta and Gofa) which are spoken in the southern part of Ethiopia. This study has the following contributions and improvements:

- We have prepared a word-level open source dataset for two languages for this study.
- The proposed model will be applied in the other low-resource Ethiopian languages.
- Social medial issues like hate speech, spreading of rumours, conspiracy theories, racism, and prejudice speakers will be identified where is written from.
- Improved communication:- LID can make it easier for people to communicate with others who speak different languages, reducing language barriers. This means that when someone speaks or writes in a particular language, the system can quickly identify it without requiring the user to specify the language manually.
- Enhanced customer Service:- by identifying multiple languages, businesses can offer better customer service and support to a wider range of customers.
- Increased accessibility:- with the ability to recognize multiple languages, technology, and applications can be made more accessible to a wider audience.

## 2   Related Work

In recent years, there has been a shift towards using deep learning(DL) approaches for bilingual word-level LID. Sun et al.  [7] proposed using a convolutional neural network (CNN) to identify the language of a given word. They found that CNN performed well in identifying the language of words in the bilingual text and outperformed traditional approaches such as Hidden Markov Models.

Yigezu et al. [8] proposed Bidirectional Long Short Term Memory (BiL-STM) with attention to identifying code-mixed English-Kannada language. The authors used a CoLI-Kanglish data set which contains 14,847 and 4,585 words for training and testing respectively. In addition to BiLSTM, the authors also employed different ML techniques for the same data sets. The performance evaluation result of the model developed with the deep learning approach gained a macro F1-score of 0.61 and which is better than that of a model developed with ML techniques. As a strength, the work outlined the annotation of data sets well, and it employed different text classification approaches like ML, DL, and combined ones. When it was compared with [9] text classification work, deep learning-based systems less perform than transformer-based approaches.

Barman et el. [10] used three approaches to identify trilingual languages such as Bengali, English, and Hindi. The proposed approaches are unsupervised dictionary-based approach, supervised word-level classification, and sequential labeling using random fields. Finally, the Unsupervised dictionary-based has performed better results on the given data sets. Similarly, Amine et al. [11] state that Unsupervised multilingual text recognition can analyze vast amounts of data fast and does not require any prior knowledge (classes, initial partition). It has the ability to only visualize the findings.

Tonja et al. [12] proposed transformer-based word LID for code-mixed English-Kannada language. The authors combined a transformer-based pre-trained language model with a deep learning approach. The authors used BERT [1] as the embedding layer and the LSTM algorithm for the classification of the language to assigned labels for LID. Their proposed model better outperformed other pre-trained language models.

Overall, there have been many different approaches proposed for bilingual word-level LID with varying levels of success. However, recent deep learning approaches have shown promising results and may be a promising direction for future text classification [13].

## 3    Languages

### 3.1    Ethiopian Language Overview

Ethiopia is home to a diverse range of languages, with over 83 distinct languages and up to 200 spoken dialects [14] [15]. These languages are grouped into four main families: Semitic, Cushitic, Omotic, and Nilo-Saharan. The Omotic language family is one of six within the Afro-Asiatic phylum Figure 1 and is primarily spoken in the southern rift valley and around the Omo River [16], [17]. Within the Omotic branch, the Ometo languages form a large group that shares similar phonology, grammar, and vocabulary, making them distinct genetic units. Ometic languages are divided into four categories: North, South, East, and West Omotic. Wolayta and Gofa languages are grouped under Northern Omotic [18].
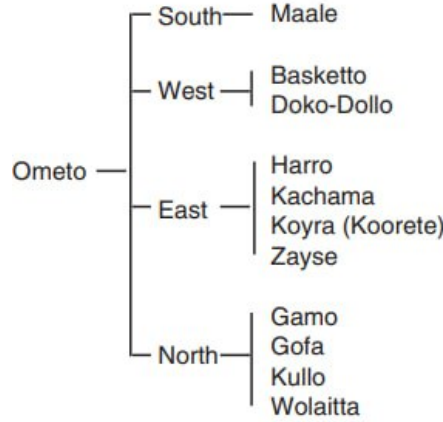
---

[1] https://huggingface.co/bert-base-uncased

Fig. 1: Overview for classification of Omotic languages

### 3.2   Wolayta Language

Wolayta is a north Omotic language spoken by the Wolayta people in southern Ethiopia [15], [19]. It is a member of the Omotic branch of the Afro-Asiatic language family. Wolayta is spoken by approximately 2.48 million people, primarily in the Wolayta Zone and some parts of Ethiopia's Southern Nations, Nationalities, and Peoples' Region (SNNPR). It is also spoken by a smaller number of people in neighbouring Zones and Regions such as Dawuro, Gofa, Sidama, and Gamo. Wolayta is one of the most widely spoken Omotic languages and is an official language in the Wolayta Zone. Figure 2 depicts the geographical location of the Wolayta and Gofa zone(yellow-shaded) [20].

### 3.3   Gofa Language

The term "Gofa" encompasses the language, culture, and region of southern Ethiopia, spoken by the inhabitants of the Gofa Zone and neighbouring communities. With over 362,000 speakers, it is used as the primary language of instruction in primary schools and is also taught as a subject in secondary and high schools. Additionally, it serves as the main language of communication in government offices within the Gofa Zone [21].

### 3.4   Challenges of languages

The study [10] says "automatic LID on social media data becomes a necessary and challenging task." Particularly, the challenges of Omotic languages are the irregularity of the words, stop word identification issues, compounding, and languages' limited access to digital data, according to [6]. Wolayta and Gofa languages are members of the Omotic languages that are spoken around river

Fig. 2: Southern region of Ethiopia map  [20]

Omo, the southern part of Ethiopia. LID is still a difficult subject to solve, especially when trying to distinguish between languages that are quite similar  [22]. A number of factors influenced the performance of BLID, including the similarity of the languages being identified and the length of the text. For example, the proposed languages that are more similar to each other are typically more difficult to distinguish than those that are less similar. Similarly, the longer text is typically more reliable for BLID than the shorter one, as it provides more information about the language being used. Further research is needed to continue to improve the performance of BLID methods and to explore new applications for these methods.

Table 1: sample words/phrases in English, Wolayta, Gofa, Wolayta-Gofa and its remark as similar/common or different in written text.

| No | In English | In Wolayta | In Gofa | Wolayta-Gofa (common) | Remark |
|---|---|---|---|---|---|
| 1 | While following | Kaallidi | Kaallidi | Kaallidi | Use common words for both |
|   | Country | Biittaa | Biittaa | Biittaa | |
|   | Bad | Iita | Iita | Iita | |
|   | Many | Daro | Daro | Daro | |
| 2 | From his/her/its talk | Haasayaappe | Hayssafe | - | Use different but closest words |
|   | While I writing | Xaafayida dayiis | Xaafettidayssi | - | |
|   | Are you going? | Bayii? | Biyey? | - | |
|   | His/her born | Yeletay | Yeletethay | - | |

Table 1 shows us the similarity and differences between both languages. The first, No column shows the similarity between Wolayta and Gofa words, and the second No column, shows their difference. The differences came from the accents of the users and the similarities came from due to the two language speakers' geographical location in the same area and their languages belonging to the same group, thus the structure and morphology of both languages are quite similar. Thus these circumstances made it difficult to identify the two languages one form other.

## 4    Methodology

We followed different methods such as dataset collection, pre-processing, annotation, approach selection, parameter tuning, conducting experiments and so on.

### 4.1    Dataset development

For this study, we collected a novel dataset from religious, educational, and social media sources. For the religious domain, we used the data from  [23], for the educational domain, we collected dataset from Wolayta Sodo University, Wolayta language department[2] and for social media domain, we collected from Facebook user comments.

### 4.2    Data pre-processing

To maintain the model quality, the data pre-processing task must be carried out before the experiment. Pre-processing is the process of cleaning and making data suitable for execution  [24] - [26]. In order to generate a high-quality training dataset, we pre-processed the data that was acquired using a variety of different methods. During the process, we carried out a variety of tasks, such as removing any URLs, non-alphabetic characters, numbers, marks, and HTML tags and converting text into lowercase. Then we compiled the pre-processed dataset into a single file and shuffled it. We used Excel Kutools [3] to remove the matching common words from both Wolayta and Gofa to annotate the data.

### 4.3    Data annotation

Before starting the annotation of the data set, deciding what specific annotating tasks should be performed is necessary. This could be labeling data and tagging entities. The annotation indicates whether the selected word is spoken in either Wolayta, Gofa, or both for each language. The compiled dataset is manually normalized and annotated with LID. Annotation tasks are performed as mentioned

---

[2] ftp://ftp.wsu.edu.et/
[3] https://es.extendoffice.com/product/kutools-for-excel.html

in the LID. The shuffled data is given to annotators of each language. These examples are presented in Table 2 which shows how the annotators annotated the training datasets.

LID refers to the process of automatically determining the language in which a given piece of text is written. The task of LID is an important component in (NLP). Its primary goal is to classify the input data into specific language categories from a predefined set of languages. The LID classify into three categories: wal (Wolayta), gof (Gofa) and wal-gof (Wolayta-Gofa).

**wal:** pertains to the Wolayta language used by the Wolayta people.

**gof:** denotes the Gofa language spoken by the Gamo and Gofa people.

**wal-gof:** is a LID term utilized when both the Wolayta and Gofa people communicate, and it indicates the presence of shared words or linguistic elements during their interactions.

Table 2: Sample words with their tags and their English meanings.

| Word | Tag | Translated in English |
|---|---|---|
| Hayassafe | gof | don't speak |
| Xaafettidayssi | gof | while I'm writing |
| Hintte | gof | You all |
| Asa | wal | People |
| Giddiis | wal | Enough |
| kaalletanawu | wal | To lead |
| Kaallidi | wal-gof | While following |
| Hara | wal-gof | Other |
| Eridi | wal-gof | knowingly |
| Doonan | wal-gof | His/her language |

### 4.4   Annotation procedure

In this task, we recruited three native speaker annotators for each language and the criteria to be selected as an annotator is a person who speaks the chosen language and knows about its structure and morphology. We held multiple one-on-one sessions with the annotators to thoroughly go over the annotation guidelines. The dataset was split into two batches, and once each batch was annotated, the authors of the paper reviewed the labeled samples. Throughout the annotation process, annotators were instructed to carefully read each text and initially determine the appropriate label for the LID in the text. Subsequently, they were asked to assign the most fitting LID based on the predefined categories. Finally, we collected the annotated data and calculated the number of language IDs in each word, and we decided on a single label based on a majority vote. when three annotators are involved in the language identification task, the minimum requirement for a language ID to be selected is that at least two out of the three annotators must agree and assign the same language label to a given

text. If there is a consensus among two annotators for a specific language ID, then that ID is considered valid and eligible for further consideration or analysis. This ensures a level of agreement and consistency among the annotators in the language identification process.

### 4.5   Dataset statistics

After being data annotated we collected a total of 144K words from the above three domains and tagged them as Wolayta, Gofa, and wal-gof. Figure 3 shows the distribution of the collected data with their respective language tags.
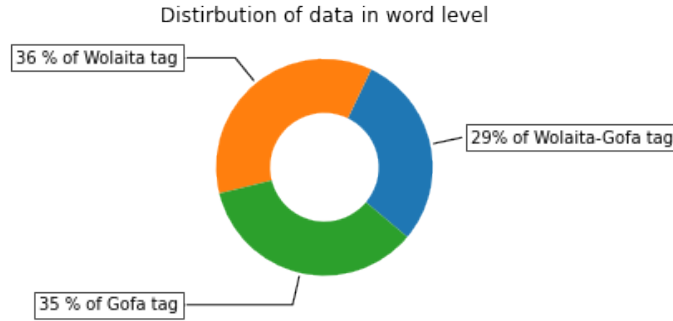


Fig. 3: Distribution of tags among whole data sets(in %)

## 5   Design of proposed Model

### 5.1   Proposed Experimental Architecture

The proposed architecture is designed to demonstrate how the model identifies bilingual languages. The main phases of the architecture include:

**Phase-1** - creating labeled data for training, the data set contains words together with their respective tags.

**Phase-2** - preprocessing such as tokenization, removing irrelevant data.

**Phase-3** - the tags are converted into machine-readable form .

**Phase-4** - after the label encoding process is complete, the representation for each token is transferred to ML and DL then evaluates the trained model. Also, each token is transferred to transformer layers which allow a pre-trained language model to produce contextualized tokens.

**Phase-5** - the embeddings that were obtained in step 4, were then fed into an LSTM model in order to generate the language tag that corresponds to them and finally evaluate the trained model.

**Phase-6** - the trained model was evaluated by accuracy metrics.

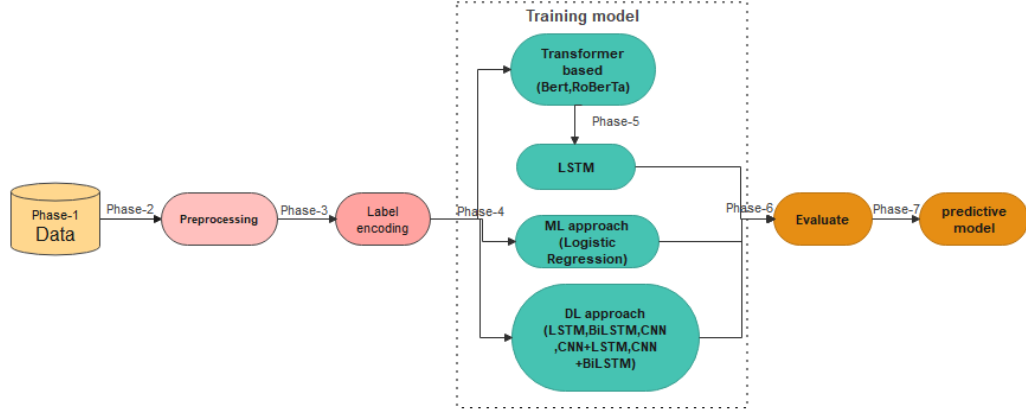**Phase-7** - after being evaluated, the output which is the predictive model is developed.



Fig. 4: Architecture of proposed model

## 5.2 Experiment and parameter setting

**Experiment** In this study, we present an experimental evaluation of several approaches to BLID. We developed the proposed model using the following approaches:

**Machine learning** is a branch of artificial intelligence that focuses on developing algorithms and statistical models that enable computers to automatically improve their performance on a given task through experience [27].

**Deep learning** is a subfield of machine learning that uses algorithms inspired by the structure and function of the brain, known as artificial neural networks, to process and analyze large amounts of data [28] - [30].

**Transformer based** is a deep learning model based on neural networks, specifically designed for processing sequences of data, such as natural language text [31] - [32].

In order to train the algorithms and develop the model, we have used various platforms. Among those, the basics are mentioned below. We have used a cloud-based platform known as Google Colab. [4] which is a popular open-source

---

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

framework for deep learning with rich libraries such as TensorFlow [5], PyTorch [6], and Keras [7].

**parameter setting** It's important that the optimal parameter settings for a BLID model depend on the specific iterative problem and may require some trial and error to find the best settings. When we train a model using a deep learning approach for BLID, there are several parameters that must be set in order to optimize the performance of the model.

For the purpose of defining the number of times that the learning algorithm will run through the full training data set, with a range of 10 to 30 epochs. We used the Leaky ReLU function, which allows returning a small positive value to "leak" through for negative input values while using standard relu 0 for all negative input values. In addition to that, the softmax function is used in the output layer of a classifier, such as a neural network, to convert the raw output values into probabilities. This is useful in multi-class classification problems where there are multiple possible outcomes.

Furthermore, we applied batch normalization to the input of each layer, which helped stabilize the training process and improve the performance of the network. It is typically applied to the input of each layer before the activation function, and it is typically used in convolutional neural networks and fully connected networks. We adjusted the batch size to 128 in the model configuration to impact the speed and accuracy of the training process.

We utilized the learning rate amount of 0.001, which helps to control the step size at which the algorithm updates the weights and biases of the neural network during the training process. This is done by iteratively adjusting the parameters of the model and evaluating the error until the minimum error is found. We applied the Adam optimizer in deep learning to find the optimal parameters for a model.

Generally, in this experiment, we experimented with four categories.

The experiment was trained on given datasets with three tags, allowing them to learn the distinctive features of each language and identify them accurately. In the first category, we applied the machine-learning technique of logistic regression.

In the second category, to evaluate the performance of the LID model, we used a deep learning approach such as LSTM, BiLSTM, and CNN. After pre-processing our textual data, The text was tokenized and trained with the mentioned approach with the help of a trained data set.

In the third experiment category, we evaluated our model by using a combination of CNN with LSTM and BiLSTM. To extract features for the given text, we used CNN and then used LSTM and BiLSTM as classifiers.

In the fourth experiment, We used transformer-based models such as BERT-

---

[5] https://tensorflow.org/

[6] https://pytorch.org/

[7] https://keras.io/

based uncased and RoBERTa-based to develop our model. Finally, We evaluated the performance of our models using several techniques such as accuracy, precision, recall, and F1-score. We discussed our results in Section 6.

## 6    Results and Discussion

The proposed models are trained on given words in both languages with three tags, allowing them to learn the distinctive features of each language and identify them accurately. Table 3 shows the experimental results

| Experiment | Approaches | Model | Precision | Recall | F1-score |
|---|---|---|---|---|---|
| Exp-1 | Machine Learning(ML) | Logistic Regression | 0.55 | 0.76 | 0.47 |
| Exp-2 | Deep learning(DL) | LSTM with attention | 0.78 | 0.43 | 0.55 |
| | | BiLSTM wiith attention | 0.7 | 0.66 | 0.62 |
| | | CNN | 0.77 | 0.55 | 0.64 |
| Exp-3 | Combination of DL | CNN + LSTM | 0.69 | 0.64 | 0.66 |
| | | CNN + BiLSTM | 0.71 | 0.67 | 0.69 |
| Exp-4 | Combination of PLM and DL | Roberta base + LSTM | 0.77 | 0.64 | 0.69 |
| | | Bert-base-uncased + LSTM | **0.76** | **0.69** | **0.72** |

Table 3: Results for the experiments

In the first experiment, we perform an ML technique which is Logistic Regression and achieves 0.47 with F1-score.
In the second experiment, we explored three DL techniques for performing LID at the word level in the bilingual language. We used character-level LSTM and BiLSTM with attention and CNN techniques to perform and showed different performance outcomes based on the same dataset being used. The results showed that the BiLSTM with attention model performed well in identifying the language of text, with an F1-score of 0.68.
Thirdly, the experiment was conducted with a combination of LSTM, BiLSTM, and CNN. In this experiment, we used CNN for feature extraction. By using convolutional and max-pooling layers, the model was able to capture the features of the text. In that experiment, the combination of BiLSTM and CNN performed better, with an F1-score of 0.69.
The fourth experiment indicates that the transformer-based model has been found to perform better than an entire experiment. We applied a pre-trained language model, which is Bert uncased in the base layer and Roberta in the embedding layer, with the LSTM model. The Bert base uncased model outperforms the Roberta base pre-trained language models that were used in the embedding layer with the LSTM model, with a macro score of 0.76 precision, 0.69 recall, and 0.72 F1-scores.
The model is based on the LSTM architecture, which is a type of RNN that can capture long-term dependencies in sequential data. The model computes the

embeddings of the input text using a pre-trained model. The embeddings are a dense representation of the input text that can capture semantic and syntactic information. The embeddings are fed to an LSTM layer with 128 units. The output of the LSTM layer is passed through a batch normalization layer, which normalizes the activations of the previous layer. The normalized activations are then fed to a dense layer with 768 units, followed by a ReLU activation function. Another dense layer with 768 units and a dropout layer with a dropout rate of 0.1 is applied next. Finally, the output of the dropout layer is passed through a dense layer with 3 units and a softmax activation function, which produces the output probabilities for each of the 3 classes.

Therefore, this model is able to effectively learn the language patterns in the bilingual text dataset. This allowed the model to effectively distinguish between different languages even when the text was short and had limited information.

## 7   Conclusion

The experiments conducted on bilingual LID at the word level using various machine learning, deep learning techniques, and transformer-based model, specifically the Bert base uncased model in the embedding layer of LSTM, performed the best with a 0.72 F1-score among the other models tested. The combination of BiLSTM and CNN also showed good performance but was slightly outperformed by the Bert base uncased model. These results suggest that using pre-trained language models such as Bert and Roberta can be an effective approach for identifying the language of bilingual text.

This study can be used as a starting point for further research and development in this area. Additionally, this study also highlights the importance of creating and annotating datasets for BLID.

To improve the performance of the bilingual LID model, further research can focus on developing more advanced algorithms, increasing the size [33] and diversity of the training data, and experimenting with different feature representations. Additionally, an important application of the bilingual LID model lies in machine translation tasks. By accurately detecting the source language, the appropriate language model for translation can be selected. This way, the translation system can apply language-specific rules, context, and vocabulary, resulting in more precise and contextually relevant translations [34].

## 8   Acknowledgments

## References

1. Lui, M., Lau, J. H., and Baldwin, T. (2014). Automatic detection and LID of multilingual documents. Transactions of the Association for Computational Linguistics, 2, 27-40.
2. Lillian Lee, Mark Johnson, Kristina Toutanova, and Brian Roark. 2018. Transactions of the Association for Computational Linguistics, Volume 6. Transactions of the Association for Computational Linguistics, 6.
3. Hughes, B., Baldwin, T., Bird, S., Nicholson, J., and MacKinlay, A. (2006, May). Reconsidering LID for Written Language Resources. In LREC (pp. 485-488).
4. Indhuja, K., Indu, M., Sreejith, C., Sreekrishnapuram, P., and Raj, P. R. (2014). Text based LID system for indian languages following devanagiri script. International Journal of Engineering, 3(4).
5. Kroll, J. F., Dussias, P. E., Bogulski, C. A., and Kroff, J. R. V. (2012). Juggling two languages in one mind: What bilinguals tell us about language processing and its consequences for cognition. In Psychology of learning and motivation (Vol. 56, pp. 229-262). Academic press.
6. Bade, G. Y. (2021). Natural Language Processing and Its Challenges on Omotic Language Group of Ethiopia. Journal of Computer Science Research, 3(4), 26-30.
7. Shiliang Sun, Chen Luo, and Junyu Chen. A review of natural language processing techniques for opinion mining systems. Information fusion, 36:10–25, 2017
8. Yigezu, M. G., Tonja, A. L., Kolesnikova, O., Tash, M. S., Sidorov, G., and Gelbukh, A. (2022, December). Word Level LID in Code-mixed Kannada-English Texts using Deep Learning Approach. In Proceedings of the 19th International Conference on Natural Language Processing (ICON): Shared Task on Word Level LID in Code-mixed Kannada-English Texts (pp. 29-33).
9. Jauhiainen, T., Ranasinghe, T., and Zampieri, M. (2021). Comparing approaches to Dravidian LID. arXiv preprint arXiv:2103.05552.
10. Barman, Utsab, Amitava Das, Joachim Wagner, and Jennifer Foster. "Code mixing: A challenge for LID in the language of social media." In Proceedings of the first workshop on computational approaches to code switching, pp. 13-23. 2014.
11. Abdelmalek Amine, Zakaria Elberrichi, and Michel Simonet. Automatic LID: An alternative unsupervised approach using a new hybrid algorithm. Int. J. Comput. Sci. Appl., 7 (1):94–107, 2010
12. Atnafu Lambebo Tonja, Mesay Gemeda Yigezu, Olga Kolesnikova, Moein Shahiki Tash, Grigori Sidorov, and Alexander Gelbuk. Transformer-based model for word level LID in code-mixed Kannada-English texts. arXiv preprint arXiv:2211.14459, 2022
13. Mager, M., Çetinoğlu, Ö., and Kann, K. (2019). Subword-level LID for intra-word code-switching. arXiv preprint arXiv:1904.01989.
14. waldemariam, H. (2007). The challenges of mother-tongue education in Ethiopia: The case of North Omo area. Language Matters, 38(2), 210-235.
15. Wakasa, M. (2014). A sketch grammar of walaytta. Nilo-Ethiopian Studies, 2014(19), 31-44.
16. Azeb, A. (1996). Aspects of the verb in Ometo: Ethiopia. J. Lang. Lit., 6, 43-64.
17. Amha, A. (2017). The Omotic language family. Cambridge University Press.

18. Bedecho, A. T., and waldeyohannis, M. M. (2022, November). walaytta-English Cross-lingual Information Retrieval using Neural Machine Translation. In 2022 International Conference on Information and Communication Technology for Development for Africa (ICT4DA) (pp. 96-101). IEEE.

19. Michael, L., de Carvalho, F., Chacon, T., Rybka, K., Sabogal, A., Chousou-Polydouri, N., and Kaiping, G. (2022). Deriving calibrations for Arawakan using archaeological evidence. Interface Focus, 13(1), 20220049.

20. Melesse, A., and Negesse, T. (2011). Phenotypic and morphological characterization of indigenous chicken populations in southern region of Ethiopia. Animal Genetic Resources/Resources génétiques animales/Recursos genéticos animales, 49, 19-31.

21. Hirboro, S. C. (2015). Documentation and grammatical description of Gofa. Addis Ababa Univeristy.

22. Bjerva, J. (2016). Byte-based LID with deep convolutional networks. arXiv preprint arXiv:1609.09004.

23. Tonja, A. L., waldeyohannis, M. M., and Yigezu, M. G. (2021, November). 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) (pp. 71-76). IEEE.

24. Kumari, S. (2022). Text mining and pre-processing methods for social media data extraction and processing. In Handbook of Research on Opinion Mining and Text Analytics on Literary Works and Social Media (pp. 22-53). IGI Global.

25. Yigezu, M. G., Kolesnikova, O., Sidorov, G., Gelbukh, A. (2023). Transformer-Based Hate Speech Detection for Multi-Class and Multi-Label Classification.

26. Yigezu, M. G., Bade, G. Y., Kolesnikova, O., Sidorov, G., Gelbukh, A. (2023). Multilingual Hope Speech Detection using Machine Learning.

27. Mahesh, B. (2020). Machine learning algorithms review. International Journal of Science and Research (IJSR).[Internet], 9, 381-386.

28. Fourie, C. M. (2003). Deep learning? What deep learning?: research in higher education. South African Journal of Higher Education, 17(1), 123-131.

29. Yigezu, M. G., Kebede, T., Kolesnikova, O., Sidorov, G., Gelbukh, A. (2023, September). Habesha@ DravidianLangTech: Utilizing Deep and Transfer Learning Approaches for Sentiment Analysis. In Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages (pp. 239-243).

30. Yigezu, M. G., Kanta, S., Kolesnikova, O., Sidorov, G., Gelbukh, A. (2023, September). Habesha@ DravidianLangTech: Abusive Comment Detection using Deep Learning Approach. In Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages (pp. 244-249).

31. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... and Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

32. Yigezu, M. G., Mehamed, M. A., Kolesnikova, O., Guge, T. K., Gelbukh, A., Sidorov, G. (2023, October). Evaluating the Effectiveness of Hybrid Features in Fake News Detection on Social Media. In 2023 International Conference on Information and Communication Technology for Development for Africa (ICT4DA) (pp. 171-175). IEEE.

33. Tash, M. S., Ahani, Z., Tonja, A., Gemeda, M., Hussain, N., and Kolesnikova, O. (2022, December). Word Level LID in Code-mixed Kannada-English Texts using traditional machine learning algorithms. In Proceedings of the 19th International Conference on Natural Language Processing (ICON): Shared Task on Word Level LID in Code-mixed Kannada-English Texts (pp. 25-28).

34. Yigezu, M. G., waldeyohannis, M. M., and Tonja, A. L. (2021, November). Multilingual neural machine translation for low resourced languages: Ometo-english. In 2021 International Conference on Information and Communication Technology for Development for Africa (ICT4DA) (pp. 89-94). IEEE.