Supported By
tetfund
TERTIARY EDUCATION TRUST FUND

# TRANSFER LEARNING ON CROSS LINGUAL NEWS CLASSIFICATION OF LOW RESOURCE YORUBA LANGUAGE USING BI-LSTM ON A SIAMESE NETWORK

**[1]Abdullah, Khadijha-Kuburat Adebisi, [2] Sodimu Segun Michael ,**
**[3] Efuwape Biodun Tajudeen , [4] Olasupo Ahmed Olalekan.**
Olabisi Onabanjo University
Department of Mathematical Sciences
Ago Iwoye, Ogun State, Nigeria.
**Corresponding Email: abdullah.adebisi@oouagoiwoye.edu.ng, +2348060046592**

**Received: December 13, 2021 Accepted: March 20, 2022**

ABSTRACT Most existing language models cannot handle low-resource textual data due to diversity in language representation and non-availability of text corpora. Transfer learning from high-resource help in such language, thus, disregard vocabulary overlap. Hence, cross-lingual news classification in Siamese network learn and build better model to encode sentences with few samples into shared embedding features from monolingual pretrained model using Bidirectional Long Short Term Memory (BiLSTM). The BiLSTM sequence model takes sentences of news articles for each language as input sequences, independently learns monolingual embeddings from parallel corpora using Skip-gram embeddings with negative sampling. Employed a lexicon to enhance the language model of low resource language and encodes into respective features representations. These embeddings are jointly aligned into a common cross-lingual features to capture semantic structure of the languages. The model is minimised with *L2*-regularization softmax cross-entropy loss $(L_{RCE})$ and enhanced with Adam optimizer. At end of 100 epochs, the result shows an accuracy of 0.84 with loss of 0.24 while precision, recall and F1-score are 0.88, 0.92 and 0.89 respectively. The model confusion matrices increase as the epoch increases with decrease in loss function. The experiments show an aligned sentences task in two languages; English and Yoruba, also, embedding trained with pretrained sequence BiLSTM is improved with monolingual data.

Keywords**:** BiLSTM, Cross-lingual, Low-resource, *L2*-regularization, Siamese network

## INTRODUCTION

Internet has gained widespread in the world and the linguistic diversity representations has grown but most existing work in Natural Language Processing (NLP) focuses on English or other languages such as German, French that have text corpora for processing. Meanwhile, there are thousands of languages spoken globally such that the openness of resources is imbalanced (Nestle, 1998). But with the help of NLP tool; machine translation has grown and diversify to the extent that speakers of low-resource continue to use the language with low samples in the distribution, though, this only supported by about 100 languages (Johnson, et al., 2017). Using only machine translator on the corpus of the high-resource language (source) has many advantages/ disadvantage over building new corpus for low-resource language (target). Machine translation overlap with different language pairs and domain mismatches (Guzman et al., 2019), however, this may not fix for some languages. Yoruba language is one of the low-resource also one of the indigenous Nigeria languages alongside Hausa and Igbo. It is spoken by over 30 million people in Nigeria and some other neighbouring countries such as Republic of Benin, Ghana, Sierra Leone and Togo (all in Africa) as well as some communities in Cuba and Brazil (Adeoye *et al.*, 2014) The task of learning from one language to another involved cross-lingual learning. Cross-lingual transfer involves transfer learning used data and weights of the neural models available from high-resource language for sample of which the resources are available (e.g., English) to solve tasks in low-resource language with fewer samples. Most existing work on learning representation have focused on transferring knowledge across tasks for a single language (English). In NLP, initialising word embedding with pre-trained word representations obtained from Word2Vec (Mikolov et al., 2013) or GloVe (Pennington et al., 2014) has become a common way of transferring learning from large labelled data to low language tasks. This represents each word as low dimensional vectors that capture syntactic and semantic information. Recently, learning cross-lingual sentence representation enhanced high-resource language by mapping semantic words in the two languages and transform the functions over corresponding word embeddings features. The data distributions are highly similar in their sentence representations and allow corresponding hypothesis in the two languages. Apparently, the need for large amounts of data in low-resource languages is an issue of the current methods for learning embedding. There is need to build cross-lingual that encode sentence into a shared embedding features from high-resource to improve NLP models. There have been several bilingual representations using sentence learning which have achieved good results for classification *(*Hermann & Blunsom, 2014; Gouws & Søgaard, 2015). Cross-lingual have different languages but the domain is similar, therefore, transfer learning of pretrained models are done *across languages. The pretrained models trained the news articles in different languages with pretrained weights to handle textual information.*

Though, the goal of cross-lingual approaches is to capture linguistic regularities in words or sentences that share same semantic and syntactic features across languages. It is understandable that the concept across languages is to enable the learning between different languages. Some existing cross-lingual used supervised bilingual aligned from multilingual corpora at sentence level *(*Hermann & Blunsom, 2014; Luong *et al*., 2015) while Ruder *et al*, (2018) required bilingual supervision with seed translation dictionaries with aligned pretrained monolingual embeddings. Recent works on cross-lingual involved parallel corpus either to learn a bilingual document or sentence representation (Zhou *et al*., 2016) or used machine translation to address the issue of cross lingual (Zou *et al*., 2013). According to Ruder *et al*., (2019), models from machine translation that involve word or sentence alignment are motivated with cross-lingual representation which can be fine-tuned on labelled data.in different languages. Nagoudi et al., (2017) presented work on cross-lingual language semantic similarity while Vulic & Moen, (2015) presented cross language information retrieval. Although, embedding of cross lingual is a difficult task due to transfer of knowledge between different languages. Faruqui & Dyer (2014) shown that training on parallel data additionally enriches monolingual representation quality. This effectively share semantics of the text sequences across the two or more independent embedding to solve the issues of both word aligned polysemy. Consequently, Gouws & (2015) presented cross lingual objective model using monolingual and sentence with aligned parallel corpora. Apparently, cross lingual training of language models has been successful in learning sentence representations from high-resource tasks to improve low-resource tasks (Conneau et al., 2017) as well as improves text classification (Xiao, & Guo, 2014; Adams et al., 2017). Lample & Conneau (2019) presented translation language modeling (TLM) to strengthen parallel data and obtain better results on cross-lingual language processing.

This study proposes cross-lingual BiLSTM transfer learning model using parallel learning representation in both monolingual languages for Yoruba-English news articles classification on a Siamese network. The approach involves an independent monolingual embedding from parallel corpora with online translation from source to target language to obtain joint cross-lingual embedding. The similarity level of the two embedding is computed by determining the absolute difference of the two embedding. This captures a common structure feature representation of the two languages. Subsequently, Bi-LSTM is adopted to learned and trained the embedding by summing up the feature in forward and backward directions of the embedding then, merge features of each representation. Thus, the performance of low-resource Bi-LSTM classification is improved by using Skip-gram embedding with negative sampling (SGNS) for cross-lingual objective function. Gouws et al. [7] minimised skip-gram with negative sampling model with *L2*-loss function using bag-of word vectors on parallel sentences. This is similar to our work but in the study, *L2* regularization softmax cross-entropy loss $(L_{RCE})$ **is** employed with Adam optimizer by learning parameters on top of the cross-lingual representations. The main contributions of this work are:

(i) Independent parallel corpora is crawl and build for bilingual languages to learn the embedding in order to bridge the language barrier and exploit semantic of the two languages with pre-trained embedding of each language.
(ii) The language labels used in the embeddings is enhanced by Bi-LSTM cross-lingual objective to capture a common language structure for the classification
(iii) Overfitting is prevented by *L2* regularization softmax cross-entropy loss $(L_{RCE})$ with Adam optimizer.

The rest of this work is organised as follows: section 2 describes the collection of dataset and the approach methodology including the skip-gram word embedding with cross lingual embeddings model with Bi-LSTM network. Section 3 presents the experimental results with discussion. Conclusion of the work with our findings and possible directions for future work is discussed in 4.

## Materials and Methods

### Search Strategy

References for this study were identified through searches of ACM, arXiV, Google Scholar, IEEE Xplore, digital library from 2005 to 2020.

### Data collection and Preprocessing

In this study, 3532 documents dataset are randomly crawled from different news sources to form news corpus then translated into corresponding source (Yoruba) language using online translator. The datasets are set into training, test set and validation of 72:20:8 respectively. Normalization is performed by lowercasing all the sentences in the document sets, tokenization and stopword are performed on both the English and Yoruba as in Eq. 1a & 1b for translation of English to Yoruba (low resource)

Let $S=\{s_1,s_2\cdots s_{|s|}\}$= vocabulary of $l_1$ with $|s|$ tokenize words Eq. (1a)

$T=\{t_1,t_2\cdots t_{|t|}\}$ = vocabulary of $l_2$ with $|t|$ tokenize words. Eq. (1b)

English (S) = “Firm targets middle income earners in new housing scheme maureen ihuamaduenyi alpha development company has revealed plans to meet the housing needs of middle income earners in its new housing scheme”.

Yoruba (T) = “ile-iṣẹ fojusi awọn ti n gba owo oya arin ni ero ile tuntun maureen ihuamaduenyi ile-iṣẹ idagbasoke alpha ti ṣe afihan awọn ero lati pade awọn aini ile ti awọn ti n wọle owo-arin ni eto ile tuntun rẹ.”

## Approach Methodology with Bidirectional Long Short Term Memory (Bi-LSTM)

The approach methodology requires news articles to form corpus *C*, there exist sentences in source language $S=\{s_1,s_2,\cdots s_n\}$ and the corresponding translations of the sentence in a target language $T=\{t_1,t_2,\cdots t_n\}$. The source and the target corpora are then converted into sequences $\overline{S},\overline{T}$ of varying lengths in Eq. 2 such that

$\overline{S}=\{s_1',s_2',\cdots s_{n_1}'\}$ and $\overline{T}=\{t_1',t_2',\cdots t_{n_2}'\}$ Eq. (2)

where $s_i'\in\overline{S}$ , $t_i'\in\overline{T}$ are sequences which are padded to fixed lengths of size 250. Our method of embedding combine by taking advantage of each method by independently and concurrently learn monolingual sentence embeddings from parallel corpora $C_1,C_2$ and joint into common feature space. The model takes the advantage of the fact that translated sentences from source *S* to target *T* are similar in its sentence representations. The training takes as inputs the sequences which are fed to the parallel skip-gram with negative sampling of window size 5, learn from the network each node $n$ and encode to generate corresponding feature representation of matrices $A=(A_T)=a_{T1},\cdots a_{Tk}$, embedding $\hat{S}_i,\hat{T}_i\in\Re^d$ of fixed output length *d*-dimensional (300) semantic representations. The embeddings are jointly fed into Bi-LSTM cross lingual embedding classification.

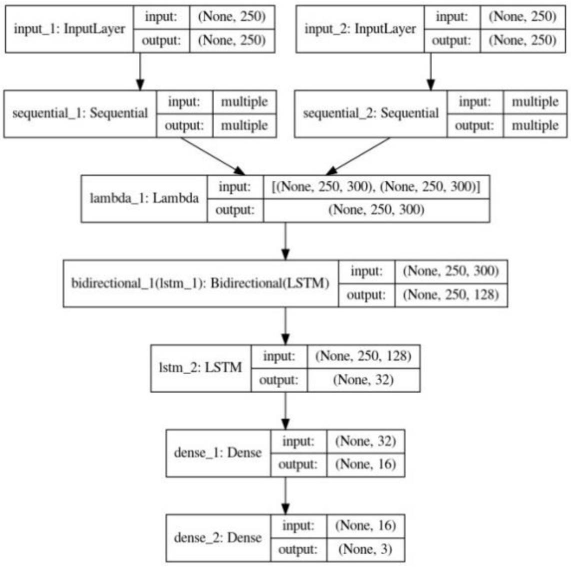

Figure 1: The Architecture of the proposed Model

The BiLSTM learns representation of the cross lingual embedding to predict the news categories. Given the input of the two embeddings $\hat{S}_i,\in\Re^{d_1}$, $\hat{T}_i\in\Re^{d_2}$ that are directly fed into the forward LSTM layer $(m_1\cdots m_n)$ without alteration and the reverse are fed into the backward $(m_n \cdots m_1)$ to learn the embeddings. At each time step $t$, BiLSTM of the hidden state $(hs_t)$ summed the forward $\vec{hs}_t$ and backward $\overleftarrow{hs}_t$ for cross lingual feature vectors of the hidden states and projected linearly through a dense layer as represented in Eq. (3)

$$wt_t = \max\{0, (\vec{wt}_t, \overleftarrow{wt}_t)wt + b\} \in \Re^d$$

Eq. (3)

The parameter weighted matrix $(wt_t)$ is learned on the dense layer with a fixed length of dimension $d$ such that $wt_t \in \Re^d$ in the LSTM, the output layer consist of three (3) neurons (128), such that categorical cross entropy loss function is used to formulate the training objective as depicted in Eq. 4

$$\delta_i(x) = \frac{e^{x_i}}{\sum_{i=1}^{N} e^{x_i}}$$

Eq. (4)

The proposed model is trained by maximizing such that solution space is limit for the training samples, *L2* regularization softmax cross-entropy loss $(L_{RCE})$ is adopted for forward and backward directions for each language.

**Table 1: The Precision, Recall and F1-score for Business News Categorization**

| Epoch | Precision | Recall | F1Score |
|---|---|---|---|
| 20 | 1 | 0.35 | 0.52 |
| 40 | 1 | 0.36 | 0.53 |
| 60 | 0.92 | 0.99 | 0.95 |
| 80 | 0.80 | .0.98 | 0.88 |
| 100 | 0.92 | 0.97 | 0.95 |

Given training samples as $M = \{(x_i, y_i) | i \in \{1,2,\cdots K\}\}$ and $Y_i$ as one-hot vector of classes $N$, the non-zero dimension of the class label of sample $x_i$, $wt_j$ represents the parameter weights and $b_j$, $j \in \{1.2,\cdots N\}$ represent bias of the $j^{th}$ class. The objective function is defined as follows in Eq. 5:

$$\ell_{RCE}(Y_i) = -\sum_{i=1}^{K} \log \frac{\exp(wt_{x_i}^T x_i + b_{x_i})}{\sum_{j=1}^{N} \exp(wt_j^T x_i + b_j)} + \delta \sum_{i \neq j} \left\| wt_i - wt_j \right\|^2$$

Eq. (5)

## RESULTS AND DISCUSSION

In the experiments, there are total number of 3532 documents for source and target language respectively with the training samples of 2542, 707 for test and validation of 283 for English and Yoruba languages each. The model is trained with hyperparameters of 20, 40,60,80,100 epochs, with learning rate of 0.0001, Adam optimizer with skip-gram consisting of contextual window of size 5 and negative sampling of size 5. The news categories are Business, Entertainment and Family with confusion matrices of precision, recall and F1score for each of the categories as in table1, table 2 and table3 respectively

In the business categories in table 1, it shows that as the epoch increases, the precision decreases but at 100 epochs it increases again. While recall increases as the epoch increases as well as F1 scores also increases as the epochs increase but 80 epochs it decreases but increases when the epochs get to 100 epochs, hence, increases the predicted values.

**Table 2: The Precision, Recall and F1-score for Entertainment News Categorization**

| Epoch | Precision | Recall | F1Score |
|---|---|---|---|
| 20 | 0.00 | 0.00 | 0.00 |
| 40 | 0.00 | 0.00 | 0.00 |
| 60 | 0.42 | 0.94 | 0.58 |
| 80 | 1.00 | .0.64 | 0.78 |
| 100 | 0.97 | 0.80 | 0.88 |

For the entertainment table 2, as the epoch increases, the precision increases but decreases at 100 epochs. While recall increases as the epoch increases but at 80 epochs, it decreases and increase as the epoch increase again. However, F1score increases as the epochs increase Therefore,
predicted values increases.

**Table 3: The Precision, Recall and F1-score for Family News Categorization**

| Epoch | Precision | Recall | F1Score |
|---|---|---|---|
| 20 | 0.00 | 0.00 | 0.00 |
| 40 | 0.00 | 0.00 | 0.00 |
| 60 | 0.97 | 0.53 | 0.69 |
| 80 | 0.56 | 1.00 | 0.72 |
| 100 | 0.75 | 0.98 | 0.85 |

For the family table 3, it shows that as the epoch increases, the precision, recall and F1 score did not have effect at all until the epochs increases to 60 epochs, then there is changes in the confusion matrices with increases in precision, recall and F1 scores.

**Table 4: Accuracy and Loss for the Training and Validation**

| Epoch | Training Accuracy | Training Loss | Validation Accuracy | Validation Loss |
|---|---|---|---|---|
| 20 | 0.3733 | NAN | 0.34982 | NAN |
| 40 | 0.3690 | NAN | 0.3568 | NAN |
| 60 | 0.7738 | 0.5541 | 0.7491 | 0.5078 |
| 80 | 0.8233 | 0.5191 | 0.7915 | 0.5943 |
| 100 | 0.9272 | 0.2383 | 0.9010 | 0.3166 |

From the training and validation in table 4, it shows that at epochs 20, 40 there is no training at all that is why precision, recall and F1 score did not have any value as shown in table 2 and 3 for entertainment and family. This is based on the number of training values. But as the epochs increases at 60, the model is trained and increases as the epoch increases while the loss decrease as shown in figure 2.

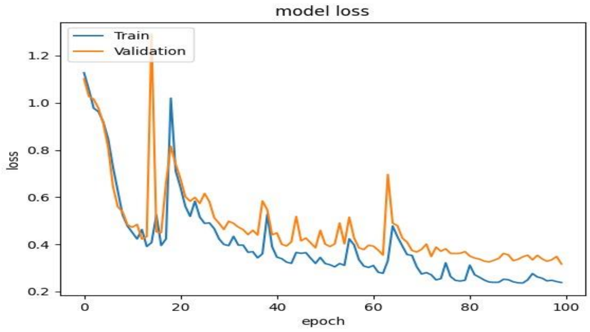

**Figure 2: The Model Validation and Loss**

## CONCLUSION

The BiLSTM model is jointly learned on cross-lingual embeddings with parallel monolingual data, enriched with lexicon to improve the part-of-speech in sentence translation. The pretrained language models of skip-gram with negative sampling are used to aligned the bilingual sentences representation of vocabularies of the languages to reduce the polysemy in the two languages; English and Yoruba. It has been demonstrated that embedding trained with parallel corpora are valuable for sematic representation that the embeddings are joined into a common feature space, an additional loss function is added to bilingual embedding and also improves the quality of monolingual word feature despite training on low-resource small datasets.

## ACKNOWLEDGEMENT

The information used in the experiment for the source code is found in Tensorflow, Keras, Python

## DECLARATION OF COMPETING INTEREST

The author(s) declares that there is no competing financial interests or personal relationships that influence the work in this paper.

## REFERENCES

Adams, O.; Makarucha, A.; Neubig, G.; Bird, S.; & Cohn, T. 2017. Cross-lingual word embeddings for low-resource language modeling. *In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Association for Computational Linguistics 1*: 937–947.

Adeoye, O.B., Adetunmbi, A.O., Fasiku A.I., & Olatunji, K. A. 2014. *International Journal of English and Literature, 5(3):71-78*

Conneau, A.; Kiela, D.; Schwenk, H.; Barrault, L.; and Bordes, A. 2017. Supervised learning of universal sentence representations from natural language inference data. In Empirical Methods Natural Language Processing, 670–680.

Faruqui, M. & Dyer, C. 2014. Improving vector space word representations using multilingual correlation. *In Proceeding of the 14th Conference of the European Association for Computational Linguistics*, 462-471

Gouws S. & Søgaard A. 2015. Simple task-specific bilingual word embeddings. *In North American Association for Computational Linguistic –HLT Institute.*, 1386–1390

Guzman, F., Chen, P-J. Ott, M. Pino, J., Lample, G. Koehn, P. Chaudhary, V. & Ranzat, M. 2019. The FLORES Evaluation Datasets for Low-Resource Machine Translation: Nepali–English and Sinhala–English, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China, November 3–7, 2019. c 2019 Association for Computational Linguistics, 6098–6111.

Hermann, K. M. & Blunsom, P. 2014. Multilingual models for compositional distributed semantics. *In Proceedings of 52rd Annual Meeting of the Association for Computational Linguistic*, 58–68.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Vi´egas, F., Wattenberg, M., & Corrado, G., 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Lample G. & Conneau, A. (2019). Crosslingual language model pretraining. Neural Information Processing Systems (NeurIPS).

Luong, T. Pham, H. & Manning, C. D. 2015. Bilingual word representations with monolingual quality in mind. *In Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, 151–159.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. 2013. Efficient estimation of word representations in vector space. *In Proceedings of International Conference on Learning Representations (ICLR)* Scottsdale, AZ. arXiv:1301.3781v3 2013; 746–751.

Nagoudi, M. B., Ferrero, J., Schwab, D. & Cherroun, H. 2017. Word embedding-based approaches for measuring semantic similarity of arabic-english sentences. *In International Conference on Arabic Language Processing, Springer* 19–33.

Nettle, D. Explaining global patterns of language diversity. 1998. *Journal of anthropological archaeology,*17(4):354–374.

Pennington, J., Socher R., & Manning C. D. 2014. Glove: Global vectors for word representation. *In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP),* 1532–1543.

Ruder, S. Søgaard, A. & Vulic., I. 2018. A survey of cross-lingual embedding models. *arXiv preprint arXiv:1706.04902*.

Ruder, S. Vulić, I. & Søgaard, A. 2019. A Survey of Cross-lingual Word Embedding Models, Journal of Artificial Intelligence Research 65 (2019) 569-631

Vulic, I. & Moen, M-F. 2015. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. *In Proceedings of the 38th international Association of Computing Machinery (ACM) SIGIR Conference on research and development in information retrieval*, 363–372.

Xiao, M., & Guo, Y. 2014. Distributed word representation learning for cross-lingual dependency parsing. *In Proceedings of the Eighteenth Conference on Computational Natural Language Learning, Association for Computational Linguistics (ACL)* 119–129.

Zhou, X., Wan, X. & Xiao, J. 2016. Attention-based LSTM Network for Cross-Lingual Sentiment Classification. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics,* 247–256.

Zou, W. Y. Socher, R. Cer, D. & Manning, C. D. 2013. Bilingual word embeddings for phrase-based machine translation. *In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*,1 393–1398.