ISSN 1512-1232

MACHINE LEARNING ALGORITHMS FOR MULTILINGUAL TEXT CLASSIFICATION OF NIGERIAN LOCAL LANGUAGES

Abiodun Adeyinka.O.[1], Awoniran Olalekan[2], Ogundiran Daniel[3], Ozichi N. Emuoyibofarhe.[2]

Corresponding Author: aabiodun@noun.edu.ng
[1] Africa Centre of Excellence on Technology Enhanced Learning, National Open University of Nigeria
[2] Department of Computer Science, Bowen University Iwo, Osun State, Nigeria

ABSTRACT

Many Natural Language Processing tasks have seen significant advances thanks to multilingual text classification models in a range of languages. Due to the rising diversity in internet access from non-English speakers, the idea of multilingualism comes into play. However, developing a multilingual text classification model can become a very difficult task when it comes to the phase of choosing the machine learning algorithm to use.

A multilingual text classification models was created. A total of 3 models were trained in three languages, English, Yoruba and Hausa model containing 200853,1967, 2917 articles headlines using Python with Tensor Flow. The highest accuracy for English model using BERT Transformer is 89%, The Yoruba version had the Long Short Term Memory (LSTM) 74% and the Hausa scored 88%

Bert Transformer outperformed LSTM and Naïve Bayes (NB). Although the BERT Yoruba model was overfitting. it has been proven that BERT Transformer has the best accuracy for multilingual text classification when compared to LSTM and NB. Surprisingly, it is also demonstrated that LSTM and NB can be viable options when selecting a text classification algorithm.

KEYWORDS: *LSTM, Naïve Bayes, BERT, Transformer, NLP.*

1.0 INTRODUCTION

Artificial intelligence has made strides massively without requiring to alter the fundamental hardware infrastructure. Clients can run an Artificial intelligence program in an ancient computer system. On the other hand, the recipient impact of machine learning is boundless. Natural Language Processing is one of the branches of AI that gives the machines the capacity to examine, get it, and provide meaning. NLP has been exceptionally effective in healthcare, media, finance, and human resource.

Machine learning is a branch of artificial intelligence that allows computer systems to learn directly from examples, data, and experience. It has many algorithms and unfortunately, unable to select the right algorithm for the right problem (Amir Vansh, 2019).

Combined with machine learning algorithms, NLP makes systems that learn to perform assignments on their own and get way better through experience through text classification which is one of the important fields in natural language processing and the task of automatically sorting a set of documents into categories from a predefined set. (Khatun et al.,2020). It has numerous applications within the commercial world like automated ordering of scientific articles concurring to a predefined dataset of technical terms amongst others.

2.0 REVIEW OF LITERATURE

The multilingual transformer models have helped to push the state-of-the-art results on cross-lingual machine learning tasks. (Devlin et al., 2019; Conneau et al., 2020). Most multilingual models have a performance trade-off between low and high-resource languages. (Khandelwal et al., 2020).

In 2018, researchers at Google AI Language proposed a pre-train deep representations of words based on a large unlabeled corpus. The BERT model proved to be effective in extracting word features and contextual information from plain text, and therefore many studies have tried to incorporate the pre-trained BERT embed-dings as features into WSD systems. (Huang et al.,2019; Du et al., 2019). Also, (Tazroute, 2020; Gao et al., 2021; Michael et al., 2020) examined the application of text classification using BERT approach. An ALL-IN-1 simple model was built by (Plank, 2017) for multilingual text classification that does not require any parallel data. The text classification was based on customer feedback analysis in which data from four languages was available (English, French, Japanese and Spanish). (Al-Tamimi et al., 2021) shares the results of utilizing the active learning method to make strides in the learning capacity of AI agents to perform text classification on Arabic news articles.

(Jungo Kasai, 2021) converts a pre-trained transformer into its proficient recurrent counterpart, progressing the effectiveness whereas holding the accuracy. In comparison to the traditional transformer and other recurrent variations, (Jungo Kasai, 2021) approach uses a learnt feature map to give a better tradeoff between efficiency and accuracy. We also show that fine-tuning requires less training time than training these recurrent variants from the ground up.

(Haiming Wu, 2019) developed a shared-private LSTM (SP-LSTM), which allows domain-specific parameters to be updated on a three-dimensional recurrent neural network, was explored for tackling the challenge of shared-private models.

3.0 METHODOLOGY

This research proposed a system that used classification and regression algorithms to perform multilingual text classification on Nigerian languages namely English, Yoruba, and Hausa and reported the f1 score, recall, precision, and accuracy of each algorithm used. Naïve Bayes was used for classification or regression algorithms, recurrent long short-term memory for deep neural networks and Bert transformer for transformers.

3.1 Data Collection

1. The English Dataset comprises of 200853 article headlines gotten from series of BBC News Article Headlines. The dataset had four major categories namely 'ENTERTAINMENT', 'POLITICS', 'BUSINESS', 'SCIENCE& TECHNOLOGY.
2. The Yoruba Dataset comprises of 1967 article headlines gotten from series of random news outlets. The dataset had five major categories namely ['WORLD', 'POLITICS', 'ENTERTAINMENT', 'HEALTH', 'SPORT'].
3. The Hausa Dataset comprises 2917 article headlines gotten from series of random news outlets.

The dataset had 3 categories, namely ['HEALTH', 'POLITICS', 'WORLD'].

The machine learning algorithm is fed with training information that comprises pairs of feature sets (vectors for each content illustration) and labels (e.g. news, legislative issues) to create a classification model. Once it's trained with sufficient training samples, the machine learning show can start to create accurate predictions. The same feature extractor is utilized to convert unseen text to feature sets, which can be encouraged into the classification model to induce forecasts on labels (e.g., sports, legislative issues).

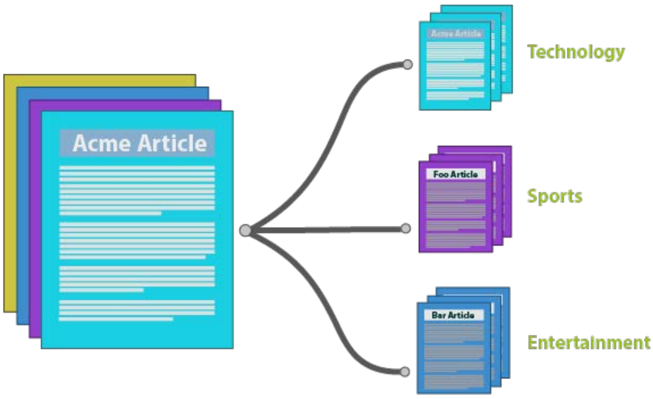

Figure 3.1: Text classification diagram

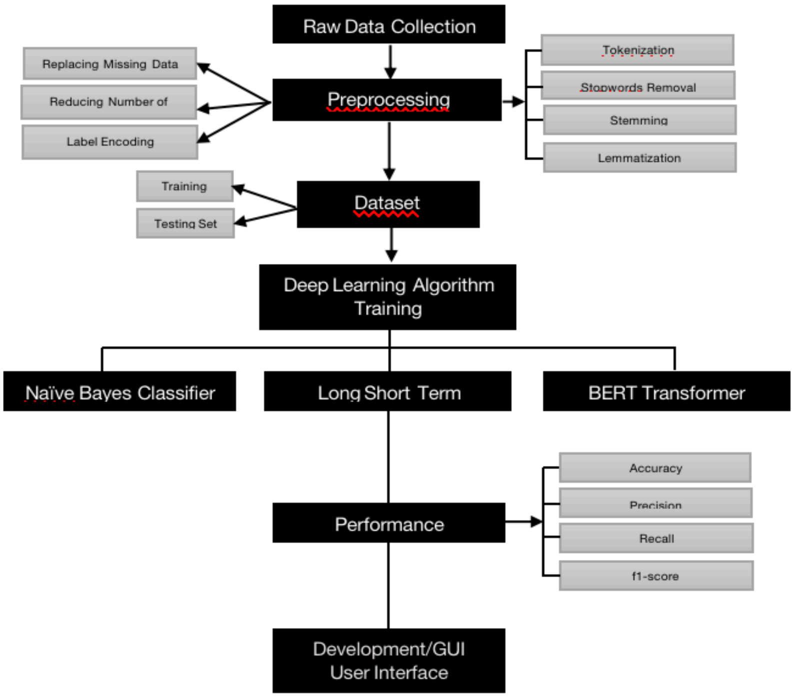

Figure 3.2: Architectural design of a proposed system

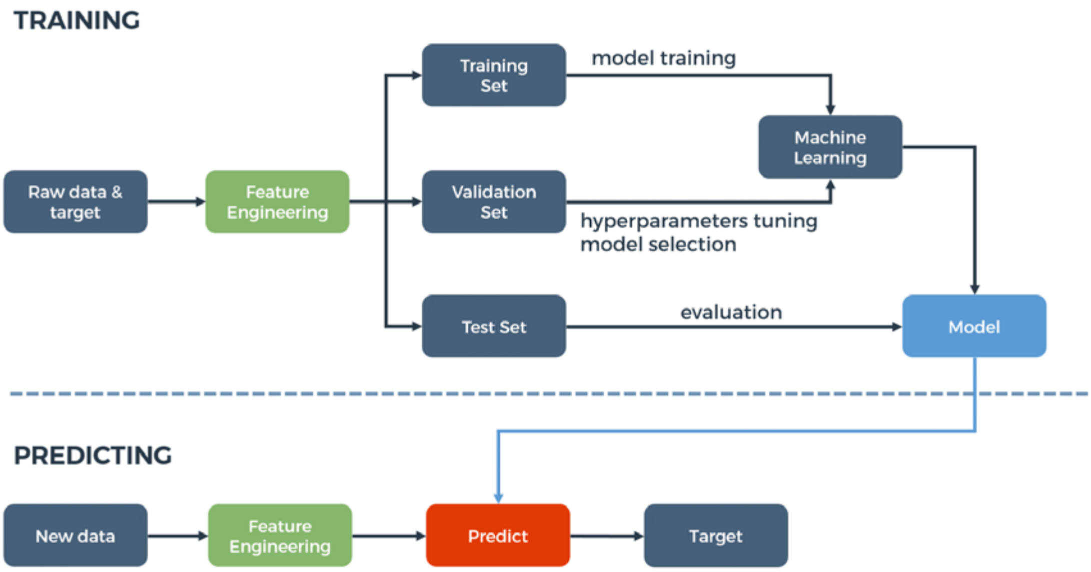

Figure 3.3: Classification Process

4.0 RESULTS AND DISCUSSION

4.1 English Model

All text classification model for the three machine learning algorithms were trained.

The three (3) machine learning models were trained in Jupyter Notebook 2019 with TensorFlow during the model development phase.

The performance metrics employed for checking the accuracy of the models includes the f1_score, precision, accuracy and recall is below.

Table 4.1: Performance metrics of the English model

Model	Class	F1_score	Precision	Recall	Accuracy
BERT	0: 1: 2: 3:	71% 90% 93% 74%	74% 90% 92% 77%	68% 90% 94% 72%	89%
LSTM	0: 1: 2: 3:	65% 89% 91% 70%	69% 88% 90% 71%	61% 90% 92% 69%	86%
Naïve Bayes (NB)	0: 1: 2: 3:	53% 87% 89% 59%	80% 88% 83% 89%	40% 86% 96% 45%	84%

Where 0 = Business, 1 = Entertainment, 2 = Politics, 3 = Science and Technology

Yoruba model gave a 5-Class prediction where 0 is Entertainment, 1 is Health, 2 is Politics, 3 is Sport, 4 is World.

Table 4.2: Performance metrics of the Yoruba model

Model	**Class**	**F1_score**	**Precision**	**Recall**	**Accuracy**
BERT	0: 1: 2: 3: 4:	0% 0% 62% 0% 0%	0% 0% 45% 0% 0%	0% 0% 100% 0% 0%	45%
LSTM	0: 1: 2: 3: 4:	61% 29% 86% 57% 57%	77% 29% 87% 99% 45%	50% 29% 85% 40% 77%	74%
Naïve Bayes (NB)	0: 1: 2: 3: 4:	52% 8% 80% 27% 41%	83% 99% 68% 99% 80%	38% 4% 98% 15% 28%	70%

Where 0 = Entertainment, 1 = Health, 2= Politics, 3 = Sport, 4 = World.

Hausa model gave a 3-Class prediction where 0 is Health, 1 is Politics, 2 is World.

Table 4.3: Performance metrics of the Hausa model

Model	**Class**	**F1_score**	**Precision**	**Recall**	**Accuracy**
BERT	0: 1: 2:	83% 80% 82%	89% 81% 79%	78% 79% 86%	82%
LSTM	0: 1: 2:	93% 85% 87%	89% 81% 93%	97% 89% 82%	88%
Naïve Bayes (NB)	0: 1: 2:	87% 82% 84%	87% 84% 82%	86% 79% 86%	84%

Where 0 = Health, 1 = Politics, 2 = World

5.0 CONCLUSION

The developed models use machine learning to perform multilingual text classification, and it has been proven that BERT Transformer has the best accuracy for multilingual text classification when compared to LSTM and Naive Bayes. Surprisingly, it is also demonstrated that LSTM and Naive Bayes can be viable options when selecting a text classification algorithm.

REFERENCES

A. Conneau, K. K. (2020). Unsupervised cross-lingual representation learning at scale.

1. Abdel-Karim Al-Tamimi, E. B.-I.-A. (2021). *Active Learning for Arabic Text Classification.* DOI: 10.1109/ICCIKE51210.2021.9410758.
2. Amina Khatun, R. M.-A. (2020). *Comparative Study on Text Classification.* International Journal of Engineering Science Invention (IJESI).
3. Flynn, S. (2016). What do we mean by 'Development' in multilingual language acquisition: Where do we start, where do we end and how do we get there? *Plenary presentation at the 10th International Conference on Third Language Acquisition and Multilingualism.*
4. Haiming Wu, Y. Z. (2019). *Shared-Private LSTM for Multi-domainText Classification.* In book: Natural Language Processing and Chinese Computing, 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II.
5. Hekmat Moumivand, R. S. (2021). *A new model for automatic text classification.* DOI: 10.30564/ese.v3i1.3170.
6. Jacob Devlin, M.-W. C. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computation* (pp. 4171-4186). Minneapolis, Minnesota: Human Language Technologies.
7. Jiaju Du, F. Q. (2019). *Using bert for word sense disambiguation.*
8. Jonathan-Raphael Reichert, K. L. (2017). *A Supervised Machine Learning Study of Online Discussion Forums about Type-2 Diabetes.* Mobile Technology Laboratory, Westerdals Oslo School of Arts, Communication and Technology, Norway.
9. Jungo Kasai, H. P. (2021). *Finetuning Pretrained Transformers into RNNs.* CC BY 4.0.
10. K. Cho, B. v. (2014). *"Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation.* arXiv:1406.1078 [cs, stat].
11. Luyao Huang, C. S. (2019). Gloss BERT: BERT for word sense disambiguation with gloss knowledge. *n Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (pp. 3509-3514). Hong Kong: Association for Computational Linguistics.
12. Michael A. Hedderich, D. I. (2020). *Transfer Learning and Distant Supervision for Multilingual Transformer Models: A Study on African Languages.* arXiv:2010 .03179v1 [cs.CL] 7 Oct 2.
13. Plank, B. (2017). *ALL-IN-1: Short Text Classification with One Model for All Languages*. Retrieved from https://arxiv.org/pdf/1710.09589v1.pdf
14. Shang Gao, M. A. (2021). *Limitations of Transformers on Clinical Text Classification.* DOI: 10.1109/JBHI.2021.3062322.
15. Tazroute, S. (2020). *Text classification for the juridical field.* DOI: 10.13140/RG.2.2.32548.91527.
16. Vajrala, A. (2019). *Text Classification*. Retrieved from DOI:10.13140/RG.2.2.20041.70246

Article received: 2022-07-28