International Journal of Artificial Intelligence Research and Development (IJAIRD)
Volume 2, Issue 2, July-Dec 2024, pp. 122-128. Article ID: IJAIRD_02_02_011
Available online at https://iaeme.com/Home/issue/IJAIRD?Volume=2&Issue=2
Impact Factor (2024): 6.98 (Based on Google Scholar Citation)
Journal ID: 234A-56Z1; DOI: https://doi.org/10.5281/zenodo.14190096

COMPARING MULTILINGUAL LANGUAGE MODELS ON INDIC NEWS HEADLINE CLASSIFICATION

Rahul Kavi
Independent Researcher, USA.

ABSTRACT

This work explores the problem of news headline classification in Natural Language Processing (NLP). This is a widely studied topic in the realm of NLP. However, limited work has been done on multilingual text classification (specific to Indic languages). Indic language models focus primarily on widely spoken Indian Languages. The performance of these multilingual language models is measured on the Indic News Headline dataset (iNLTK). This dataset also serves as a genre classification dataset (containing ten different genres/categories of headlines). This is done for languages such as Gujarati, Malayalam, Marathi, Tamil, and Telugu (non-Latin scripts). Indic languages are sometimes challenging as the data may contain English (Latin script) mixed with non-Latin script. The performance of recently released models such as Saravam-1 (an LLM launched by Saravam AI) is compared to traditional approaches (BERT-like models) such as DistilBERT, XLMRoBERTa, and IndicBERT. The Saravam-1 LLM model is fine-tuned using the PEFT LoRA approach. The performance is then compared using weighted precision, recall, and F1 scores with the fine-tuned version of the other three BERT-like models. The performance of Saravam-1 LLM stands out from other BERT models with a weighted F1 score of 0.87 on the test set. The other three fine-tuned models, XLMRoBERTa, DistilBERT, and IndicBERT, still perform reasonably with weighted F1 scores of 0.84, 0.82, and 0.79.

Keywords: Classification, Headline Classification, Language Models, Natural Language Processing, Transformers

Cite this Article: Rahul Kavi. Comparing Multilingual Language Models on Indic News Headline Classification. *International Journal of Artificial Intelligence Research and Development (IJAIRD)*, 2(2), 2024, pp. 122-128.

https://iaeme.com/MasterAdmin/Journal_uploads/IJAIRD/VOLUME_2_ISSUE_2/IJAIRD_02_02_011.pdf

I. INTRODUCTION

News classification is a well-studied problem in NLP. The task involves classifying an input text (news article/headline) into several categories (positive, negative, neutral, genre-type, etc.). In many cases, news cannot always be categorized into positive, negative, and neutral news only as the text may be subjective. However, these can be classified into genres where the genre type is the kind/category (business, sports, movies, entertainment, pop culture, etc.). This work focuses on classifying Indic languages (with mostly non-Latin script) into several categories. The ten categories include entertainment, state, sports, business, Tamil cinema, neutral, spirituality, positive, negative, and tech. The dataset mainly contains non-Latin script languages that includes Gujarati, Malayalam, Marathi, Tamil, and Telugu. In some cases, the headline text contains non-Latin characters (e.g., abbreviations, names of organizations, etc.). This is an Indic-multilingual dataset consisting of five Indic languages. Indic languages are challenging as they are linguistically diverse and have a complex script system. The availability of a large amount of quality data is also an essential factor. There are fewer datasets for NLP research in the Indian language context (unlike other Latin languages). Most of the support for Large Language Models (LLM) is focused mainly on non-Indic languages (English, French, Arabic, Chinese, etc.). Some LLMs support Hindi. However, many versions of popularly available LLMs don't support other major Indian languages, such as Gujarati, Malayalam, Marathi, Tamil, and Telugu. This work evaluates the performance of Smaller Language Models (SLM) like XLMRoBERTa, DistilBERT, and IndicBERT with Indic LLMs such as Sarvam-1. These models support the selected languages under evaluation in this study. The Sarvam-1 model is a 2-billion parameter model that supports eleven Indic languages (including English). This text generation model can be fine-tuned for several downstream tasks (including classification, etc.). The Saravam-1 model is fine-tuned for this study using the PEFT-LoRA approach. On the other hand, BERT-like models (e.g., XLMRoBERTa, DistilBERT, IndicBERT) have held their ground in size to performance ratio. Larger models need much more compute for training or fine-tuning than smaller models (BERT-like SLMs). These smaller BERT-like models perform very well on a wide variety of multilingual tasks. For the Indic multilingual headline classification task, the fine-tuned Saravam-1 model performs very well against a fine-tuned version of XLMRoBERTa, DistilBERT, and IndicBERT. The performance of these models is evaluated with a weighted F1 score along with precision, recall, and accuracy. The fine-tuned Sarvam-1 model outperforms the other SLMs with balanced performance on both train and test datasets.

II. RELATED WORK

News headline classification (or news genre classification) is a widely studied topic. One of the earlier models for news classification focused on user preferences and customization-based news story classification [1]. This approach used TF-IDF with a Naïve Bayes Classifier. Other earlier approaches include using Support Vector Machines with customized feature vectors [2]. Many earlier approaches included traditional preprocessing pipelines [3][4] that involve stemming, stop word removal, etc. These approaches are combined with conventional classification techniques (SVM, Random Forest, etc.). Preprocessing (tokenization, lemmatization, stop word removal) was a significant part of the data-cleaning process for creating feature vectors. Many widely used feature vectors included TF-IDF, Word-Frequency, Word Co-occurrence, and N-Gram approaches. These approaches were then combined with other classification approaches. Other approaches, such as Recurrent Neural Networks (RNNs), were also explored for news classification [5][6][7]. Similar Neural Network-based approaches include creating word embeddings using techniques such as Word2Vec [8][9]. Convolution Neural Networks (initially designed for images) were adapted to use with text for news

classification as well [10][11][12]. Most of the work done was focused on a single mono-lingual dataset. FastText is another popular model that has been explored to detect fake news and classify text in news articles [13][14][15]. BERT-like transformer models trained on large datasets showcased the power of pre-training [16]. This helps learn and understand common language patterns and knowledge transfer (based on the trained data). The model was tested on a wide variety of NLP tasks. After the introduction of BERT, a lot of interest was generated in training and using multilingual BERT models [17][18][19]. However, the most popular multilingual models successfully applied in various NLP tasks included XLMRoBERTa and DistilBERT. Meta introduced the XLMRoBERTa model [20]. This model was trained in a hundred languages on several cross-lingual tasks. This shows improved performance compared to Multilingual BERT (mBERT). The other model, DistilBERT [21], is smaller but more efficient (compared to the original BERT). For this paper, a multilingual version of DistilBERT was used. On the other hand, LLMs have taken off in the past few years. Widely used models such as Mistral-7B [22] and LLAMA3-8B [23] have shown human-level performance on several tasks, including classification, named entity recognition, question answering, text summarization, etc. However, these models are pretty large to train on a single GPU (or a CPU). Other techniques, such as PEFT-LoRA[24], have emerged to support such downstream fine-tuning tasks. These approaches help customize a generic text generation model to other tasks on custom datasets (e.g., classification) [25][26][27]. These models work pretty well on a wide variety of tasks. However, most of these LLMs focus on widely spoken non-Indic languages. Fewer models work well on other non-supported languages (with non-Latin scripts), such as Gujarati, Malayalam, Marathi, Tamil, and Telugu. Sarvam AI released a new LLM called Sarvam-1 [28]. This model has better token efficiency (this needs fewer tokens to represent a given word in native script). The Sarvam-1 LLM is also trained with high-quality data in ten Indic languages [29].

III. CURRENT APPROACH

This section contains the dataset description and the current approach taken. The dataset used for this experiment is described in terms of available languages in Table 1. The label description in terms of classes across languages is described in Table 2. Table 1 shows that most of the languages are well represented (except for Marathi, which has more samples than other languages in this dataset). However, in terms of class distribution, Entertainment, State, and Business samples have more examples than the rest of the classes. To account for this class imbalance, weighted cross entropy is used (where the weight for each class is calculated using inverse class frequency). This helps mitigate the over-sampling problem to a certain extent. A given train and test set contains data from all languages (represented in Table 1). The performance evaluation is done on news categories (presented in Table 2).

Table 1: Language distribution count

Language	Train Count	Test Count
Gujarati	5269	659
Malayalam	5036	630
Marathi	9672	1210
Tamil	5346	669
Telugu	4328	541
Total	**29651**	**3709**

Table 2: Class distribution count

Class	Train Count	Test Count
Entertainment	6743	842
State	6028	759
Business	4985	591
Sports	2710	355
Tamil-Cinema	2260	282
Neutral	1966	255
Spirituality	1519	194
Positive	1196	158
Negative	1166	128
Tech	1078	145

The DistilBERT, XLMRoBERTa, and IndicBERT models were initialized from their respective repositories on the HuggingFace hub. These three models were trained with a learning rate of 0.00005 and a batch size of 4. The IndicBERT and Sarvam-1 LLM models were trained for five epochs. Since the two Indic models were trained on a large corpus of Indic languages, these were fine-tuned and performed for fewer epochs (as it has an inherent understanding of major Indic languages in non-Latin scripts). The Saravam-1 LLM model was trained with the PEFT-LoRA package available on HuggingFace APIs.

IV. RESULTS AND CONCLUSION

This section contains the results obtained from this approach, along with the observations. The results are summarized with a weighted average of precision, recall, and F1 scores (this was chosen due to some class imbalance across the dataset). The results are presented in Table 3. One can see less of a delta between the weighted F1 scores of the train and test sets (which indicates that the model is robust and generalized well). For other models, such as XLMRoBERTa, the performance difference b/w Train and Test sets is significant (indicating some degree of overfitting). Based on the evaluation presented in Table 3 and Table 4, the Sarvam-1 model is a better choice for this dataset. Most original LLAMA2, LLAMA3, and Mistral LLMs are often huge and don't support major Indic languages (non-finetuned versions). Sarvam-1 LLM excels in this area. This is an excellent choice for major Indic languages-based NLP downstream tasks.

Table 3: Weighted Metrics of different models on Train set

Model	Precision	Recall	F1
IndicBERT	0.84	0.88	0.85
XLMRoBERTa	0.89	0.92	0.90
DistilBERT	0.91	0.90	0.90
Sarvam-1	0.88	0.88	0.88

Table 4: Weighted Metrics of different models on Test set

Model	Precision	Recall	F1
IndicBERT	0.78	0.81	0.79
XLMRoBERTa	0.84	0.85	0.84
DistilBERT	0.83	0.82	0.82
Sarvam-1	0.87	0.87	0.87

REFERENCES

[1] Billsus, D., & Pazzani, M. J. (1999). A hybrid user model for news story classification. In UM99 User Modeling: Proceedings of the Seventh International Conference (pp. 99-108). Springer Vienna.

[2] Paaß, G., Kindermann, J., & Leopold, E. (2004). Text classification of news articles with support vector machines. In Text Mining and its Applications: Results of the NEMIS Launch Conference (pp. 53-64). Berlin, Heidelberg: Springer Berlin Heidelberg.

[3] Liliana, D. Y., Hardianto, A., & Ridok, M. (2011). Indonesian news classification using support vector machine. World Academy of Science, Engineering and Technology, 57, 767-770.

[4] Liparas, D., HaCohen-Kerner, Y., Moumtzidou, A., Vrochidis, S., & Kompatsiaris, I. (2014). News articles classification using random forests and weighted multimodal features. In Multidisciplinary Information Retrieval: 7th Information Retrieval

[5] Du, C., & Huang, L. (2018). Text classification research with attention-based recurrent neural networks. International Journal of Computers Communications & Control, 13(1), 50-61.Facility Conference, IRFC 2014, Copenhagen, Denmark, November 10-12, 2014, Proceedings 7 (pp. 63-75). Springer International Publishing.

[6] Kandhro, I. A., Jumani, S. Z., Kumar, K., Hafeez, A., & Ali, F. (2020). Roman Urdu headline news text classification using RNN, LSTM and CNN. Advances in Data Science and Adaptive Analysis, 12(02), 2050008.

[7] Wu, X., & He, J. (2020, September). Character-level recurrent neural network for text classification applied to large scale Chinese news corpus. In Proceedings of the 2020 3rd International Conference on Machine Learning and Machine Intelligence (pp. 83-87).

[8] Rahmawati, D., & Khodra, M. L. (2016, August). Word2vec semantic representation in multilabel classification for Indonesian news article. In 2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA) (pp. 1-6). IEEE.

[9] Wensen, L., Zewen, C., Jun, W., & Xiaoyi, W. (2016, October). Short text classification based on Wikipedia and Word2vec. In 2016 2nd IEEE International Conference on Computer and Communications (ICCC) (pp. 1195-1200). IEEE.

[10] Jang, B., Kim, I., & Kim, J. W. (2019). Word2vec convolutional neural networks for classification of news articles and tweets. PloS one, 14(8), e0220976.

[11] Georgakopoulos, S. V., Tasoulis, S. K., Vrahatis, A. G., & Plagianakos, V. P. (2018, July). Convolutional neural networks for toxic comment classification. In Proceedings of the 10th hellenic conference on artificial intelligence (pp. 1-6).

[12] Conneau, A., Schwenk, H., Barrault, L., & Lecun, Y. (2016). Very deep convolutional networks for text classification. arXiv preprint arXiv:1606.01781.

[13] Taher, Y., Moussaoui, A., & Moussaoui, F. (2022). Automatic fake news detection based on deep learning, FasText and news title. International Journal of Advanced Computer Science and Applications, 13(1).

[14] Dharma, E. M., Gaol, F. L., Warnars, H. L. H. S., & Soewito, B. E. N. F. A. N. O. (2022). The accuracy comparison among word2vec, glove, and fasttext towards convolution neural network (cnn) text classification. J Theor Appl Inf Technol, 100(2), 31.

[15] Umer, M., Imtiaz, Z., Ahmad, M., Nappi, M., Medaglia, C., Choi, G. S., & Mehmood, A. (2023). Impact of convolutional neural network and FastText embedding on text classification. Multimedia Tools and Applications, 82(4), 5569-5585.

[16] Koroteev, M. V. (2021). BERT: a review of applications in natural language processing and understanding. arXiv preprint arXiv:2103.11943.

[17] Wang, Z., Mayhew, S., & Roth, D. (2019). Cross-lingual ability of multilingual bert: An empirical study. arXiv preprint arXiv:1912.07840.

[18] Wu, S., & Dredze, M. (2020). Are all languages created equal in multilingual BERT?. arXiv preprint arXiv:2005.09093.

[19] Wang, Z., Mayhew, S., & Roth, D. (2020). Extending multilingual BERT to low-resource languages. arXiv preprint arXiv:2004.13640.

[20] Conneau, A. (2019). Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116.

[21] Sanh, V. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

[22] Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. D. L., ... & Sayed, W. E. (2023). Mistral 7B. arXiv preprint arXiv:2310.06825.

[23] Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., ... & Ganapathy, R. (2024). The llama 3 herd of models. arXiv preprint arXiv:2407.21783.

[24] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685.

[25] Joshi, S., Khan, M. S., Dafe, A., Singh, K., Zope, V., & Jhamtani, T. (2024, July). Fine Tuning LLMs for Low Resource Languages. In 2024 5th International Conference on Image Processing and Capsule Networks (ICIPCN) (pp. 511-519). IEEE.

[26] Kavi, R., Anne, J. (2024). Cyberbullying Tweet Classification Using Language Models, International Journal of Artificial Intelligence Research and Development (IJAIRD), 2(2), 2024, pp. 105-110. IAEME publication.

[27] Kavi, R., Anne, J. (2024). Comparing Efficiency of Large and Small Language Models for Spam Text Detection, Journal of Scientific and Engineering Research (JSAER), 2024, pp. 35-38. Leon Publications.

[28] Sarvam-1, https://huggingface.co/sarvamai/sarvam-1

[29] https://www.sarvam.ai/blogs/sarvam-1

Citation: Rahul Kavi. Comparing Multilingual Language Models on Indic News Headline Classification. International Journal of Artificial Intelligence Research and Development (IJAIRD), 2(2), 2024, pp. 122-128.

Abstract Link: https://iaeme.com/Home/article_id/IJAIRD_02_02_011

Article Link:
https://iaeme.com/MasterAdmin/Journal_uploads/IJAIRD/VOLUME_2_ISSUE_2/IJAIRD_02_02_011.pdf

✉ **editor@iaeme.com**