nlpt_malayalm@DravidianLangTech-RANLP 2023: Fake News Detection in Malayalam using Optimized XLM-RoBERTa Model

Eduri Raja **Badal Soni** **Samir Kumar Borgohain**
Department of Computer Science & Engineering
National Institute of Technology Silchar
Assam, India, 788010
{eduri_rs, badal, samir}@cse.nits.ac.in

Abstract

The paper demonstrates the submission of the team nlpt_malayalm to the Fake News Detection in Dravidian Languages-DravidianLangTech@LT-EDI-2023. The rapid dissemination of fake news and misinformation in today's digital age poses significant societal challenges. This research paper addresses the issue of fake news detection in the Malayalam language by proposing a novel approach based on the XLM-RoBERTa base model. The objective is to develop an effective classification model that accurately differentiates between genuine and fake news articles in Malayalam. The XLM-RoBERTa base model, known for its multilingual capabilities, is fine-tuned using the prepared dataset to adapt it specifically to the nuances of the Malayalam language. A thorough analysis is also performed to identify any biases or limitations in the model's performance. The results demonstrate that the proposed model achieves a remarkable macro-averaged F-Score of 87% in the Malayalam fake news dataset, ranking 2nd on the respective task. This indicates its high accuracy and reliability in distinguishing between real and fake news in Malayalam.

1 Introduction

Fake news and misinformation proliferation have become pervasive problems in today's digital age. The rapid spread of false information through social media platforms and online news sources has the potential to mislead and deceive individuals, leading to harmful consequences for individuals, communities, and even entire nations.

In recent years, numerous research efforts have focused on developing effective solutions for fake news detection in various languages Hariharan and Anand Kumar (2023). However, even though millions of people worldwide and in the Indian state of Kerala speak Malayalam, a dearth of research is specifically focused on it. As a result, there is a pressing need to address this gap and develop robust models tailored for detecting fake news in Malayalam.

This research paper aims to tackle the challenge of fake news detection in Malayalam by proposing an innovative approach based on the XLM-RoBERTa base model. The XLM-RoBERTa Conneau et al. (2019) model is chosen due to its exceptional performance in multilingual tasks and ability to capture contextual information effectively. By leveraging the power of this model, we aim to develop a classification system capable of accurately identifying fake news articles written in Malayalam.

The contributions of this research paper include the following:

- Fine-tuning the XLM-RoBERTa base model on the Malayalam dataset to enhance its performance for fake news detection.
- Optimize the models' hyperparameters using Bayesian optimization to enhance their detection capabilities.
- Evaluating the proposed model's performance using various metrics, including precision, recall, and F1 score.

2 Related Work

Detecting fake news and misinformation has garnered significant attention from researchers in recent years. Numerous studies have explored various approaches and techniques to address this challenging problem. This section provides an overview of related work in fake news detection, focusing on methods applied to different languages and the specific challenges associated with detecting fake news in Malayalam. Sai and

Third Workshop on Speech and Language Technologies for Dravidian Languages, pages 186–191, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-085-4_026

Sharma (2021) used the XLM-R model for offensive language identification in Dravidian languages. Yasaswini et al. (2021) used the transfer learning technique to identify offensive language in Dravidian languages. To obtain optimal results, they used various classifiers like XLMR, mBERT, CNN, and ULMFiT. Ghanghor et al. (2021) used mBERT and XLM-R-based models with sauce loss and class weights for identifying the offensive language in Dravidian languages. Chen and Kong (2021) used mBERT with the TextCNN model to identify offensive language in Dravidian languages. K et al. (2021) used the CNN-BiLSTM hybrid model in shared tasks for identifying offensive language. Saha et al. (2021) proposed transformer-based ensembling strategies for identifying offensive language in Dravidian languages.

García-Díaz et al. (2022) proposed a knowledge integration method using fastText, BERT, and XLM-RoBERETa word embeddings and linguistic features. Palanikumar et al. (2022) used machine learning models and MuRIL for Transliteration as Data Augmentation for Abuse Detection in Tamil. S N et al. (2022) used TF-IDF with a random kitchen algorithm for abusive comment detection in Tamil. LekshmiAmmal et al. (2022) used MuRiL to identify offensive span in Tamil. Biradar and Saumya (2022) used IndicBERT and SVM classifier for identifying abusive comments in Dravidian code mixed data. Raja et al. (2023) used XLM-R and mBERT transformer models with adaptive fine-tuning for fake news detection in Dravidian languages. Transformer-based models have emerged as powerful tools for various NLP tasks, including fake news detection. These models leverage self-attention mechanisms to capture global dependencies in text, enabling them to learn contextual representations effectively. Transformer architectures such as BERT have demonstrated remarkable success in various NLP tasks.

3 Dataset Description

The data sources are diverse social media platforms like Facebook, Twitter, etc. The objective of the shared task is to classify the given social media post as either fake or original news. The competition organizers have released the dataset for fake news detection in Dravidian languages in the Malayalam language S et al. (2022). The dataset's train, validation, and test set distributions are depicted in Table1.

Class	Train	Validation	Test
Original	1658	409	512
Fake	1599	406	507
Total	3257	815	1019

Table 1: Dataset statistics

We see that the given dataset is almost balanced[1]. It consists of transliterated, code-mixed, emojis, and Malayalam-scripted data.

4 System Description

In this section, we present the model architecture for detecting fake news in Malayalam. We experimented with three different transformer-based models to classify the text as fake or real. We used MultilingualBERT (mBERT) Pires et al. (2019) and DistilBERT Sanh et al. (2019) as the base models, and we leveraged the optimized XLM-RoBERTa (XLM-R) as the proposed model, a transformer-based language model known for its effectiveness in various NLP tasks, including text classification.

4.1 mBERT

mBERT is an extension of the original BERT Devlin et al. (2018) model that is designed to handle multiple languages. While the original BERT model was explicitly trained for English, mBERT is trained on a large corpus of text data from various languages, enabling it to understand and process text in different languages. mBERT leverages a masked language modelling (MLM) objective during training, where it learns to predict masked words in a sentence based on the surrounding context. By training in diverse languages, mBERT learns to generate contextualized representations that capture language patterns and relationships across multiple languages.

One of the critical benefits of mBERT is its ability to perform cross-lingual transfer learning. This means that the knowledge learned from one language can be transferred to another, even if the amount of labelled data in the target language is limited. By leveraging the shared representations across languages, mBERT can improve the performance of various NLP tasks for multiple languages, such as text classification, named entity recognition, sentiment analysis, and more. It has become a popular choice for multilingual applications and

[1] https://codalab.lisn.upsaclay.fr/competitions/11176

research in NLP due to its versatility and effectiveness in handling diverse languages.

4.2 DistilBERT

DistilBERT is a variant of the BERT model, a state-of-the-art NLP model developed by Google. DistilBERT is designed to be a smaller and faster version of the original BERT model while maintaining a similar performance level. The ”Distil” in DistilBERT stands for ”distillation,” which refers to training a smaller model to mimic the behaviour and performance of a larger model. It aims to compress the original BERT model using knowledge distillation, parameter sharing, and removing unnecessary components. By distilling the knowledge from BERT, DistilBERT achieves a significantly smaller model size, which leads to faster inference times and reduced memory requirements.

DistilBERT retains most of the critical characteristics of BERT, including its ability to perform a wide range of NLP tasks such as question answering, text classification, named entity recognition, and more. It learns contextual phrases of terms and sentences by training on enormous quantities of unlabeled text data, which allows it to seize intricate language patterns and semantic relationships.

4.3 XLM-RoBERTa

The XLM-RoBERTa model is a multilingual variant of the RoBERTa model based on the transformer architecture. It is pre-trained on a large corpus of multilingual data, allowing it to capture contextual information and semantic representations across multiple languages effectively.

The XLM-RoBERTa model consists of multiple layers of self-attention mechanisms, which enable it to learn contextualized word embeddings. Considering the surrounding context, these embeddings capture the relationships between words in a sentence. By leveraging the transformer's attention mechanism, the model can effectively encode the semantic information of the input text.

4.4 Adaptation for Malayalam

To adapt the XLM-RoBERTa model for detecting fake news in Malayalam, fine-tuning is performed using the annotated Malayalam news dataset. Fine-tuning involves training the model on the labeled data, allowing it to learn the specific patterns and linguistic characteristics of fake news in Malayalam.

During fine-tuning, the XLM-RoBERTa model is augmented with a classification layer on top. This layer maps the contextualized word embeddings generated by the model to a binary classification output, indicating whether the news article is genuine or fake. The parameters of the classification layer and the pre-trained XLM-RoBERTa model are updated simultaneously during the fine-tuning process.

4.5 Training and Optimization

The fine-tuning process involves training the adapted XLM-RoBERTa model on the annotated Malayalam news dataset. The training is performed using a batch-wise approach, where a subset of the dataset is processed at each iteration. The optimizer, Adam, updates the model's parameters based on the computed loss. Tuning hyperparameters is an essential step in developing deep-learning-based models. The implementation of a deep-learning-based model is highly dependent on the choosing of the optimal hyperparameters. Bayesian optimization can help find the best set of hyperparameters that maximize the model's performance, such as precision, recall, and F score. We used a Bayesian optimization Snoek et al. (2015) method to find the optimal hyperparameters for the proposed model and the remaining models. Hyperparameters of the proposed model are depicted in Table 2.

Parameter	Value
Train_batch_size	9
Number of training epochs	10
Max sequence length	128
Learning_rate	1e-5

Table 2: Parameters of the model

To prevent overfitting the model, early stopping was employed. Training would be stopped if the model's performance on the validation set did not improve for a certain number of consecutive epochs.

4.6 Inference and Prediction

Once trained, the model can predict the authenticity of new, unseen Malayalam news articles. The input text undergoes the preprocessing steps, including tokenization and normalization. The preprocessed text is then fed into the trained model, which predicts whether the news article is genuine or fake.

Model	Accuracy	Recall	Precision	F1-score	AUC
DistilBERT	0.8464	0.8753	0.8295	0.8518	0.9068
mBERT	0.8580	0.8845	0.8411	0.8623	0.9127
XLM-RoBERTa	**0.8687**	**0.8973**	**0.8495**	**0.8728**	**0.9384**

Table 3: Performance of the models over the training data

Model	Precision	Recall	F1-score
DistilBERT	0.76	0.82	0.79
mBERT	0.81	0.83	0.82
XLM-RoBERTa	**0.85**	**0.90**	**0.87**

Table 4: Performance of the models over the test data

The prediction can be based on the probability output of the classification layer, where a threshold can be applied to determine the final label.

5 Results and Evaluation

In this section, we present the results and evaluation of the proposed fake news detection model for the Malayalam language. We analyze the model's performance based on the evaluation metrics and provide insights into its effectiveness in distinguishing between real and fake news.

5.1 Performance Metrics

The enactment of the fake news detection model is evaluated using various metrics, including macro-averaged recall, precision, and F1 score. These metrics comprehensively comprehend the model's performance across different aspects.

The precision metric measures the model's ability to accurately identify real and fake news, while recall estimates the model's ability to seize all instances of each class. The F1 score delivers a balanced criterion by assessing both precision and recall.

The performance metrics are calculated for each class (original and fake), and the overall performance is typically reported as macro-averaged or weighted averages of the class-wise metrics.

5.2 Comparative Analysis

To provide a comprehensive evaluation, the performance of the proposed fake news detection model is compared with the existing approaches, such as mBERT and DistilBERT. This comparison helps assess the model's effectiveness and highlights its strengths and weaknesses. Table 3 depicts the performance of the models over the training data. Table 4 represents the performance of the models over the test data.

The proposed optimized XLM-RoBERTa model achieved an impressive macro-averaged F-Score of 87%. It demonstrated superior performance to both the mBERT and DistilBERT models, showcasing its effectiveness in detecting fake news in Malayalam. The mBERT model achieved a macro-averaged F-Score of 82%. While it performed well, it lagged behind the XLM-RoBERTa model in terms of overall performance. The DistilBERT model obtained a macro-averaged F-Score of 79%. It exhibited slightly lower performance compared to both XLM-RoBERTa and mBERT models.

The macro-averaged F-Scores comprehensively evaluate the model's precision and recall performance across all classes. The results highlight the superiority of the XLM-RoBERTa model, which attained the highest macro-averaged F-Score of 87%. This suggests that XLM-RoBERTa is particularly well-suited for fake news detection in Malayalam. The comparative analysis of macro F1 scores indicates that XLM-RoBERTa outperforms both mBERT and DistilBERT models. The XLM-RoBERTa model's ability to capture contextual dependencies and linguistic nuances in the Malayalam language contributes to its superior performance. This finding emphasizes the importance of language-specific modelling approaches for accurate fake news detection in languages with unique characteristics like Malayalam. The results exhibit the potential of transformer-based models for addressing the challenges of fake news detection in Malayalam. By achieving a high macro F1 score, the XLM-RoBERTa model can identify and combat misinformation in the Malayalam news ecosystem.

6 Conclusion

In this research paper, we propose a fake news detection model for the Malayalam language. The model utilizes the XLM-RoBERTa architecture and is finetuned on a dataset of real and fake news articles in Malayalam. We demonstrated through extensive experimentation and evaluation that the proposed model achieves promising results in detecting fake news in Malayalam. The model achieved an impressive macro F1 score of 87%, indicating its ability to balance precision and recall. We compared the proposed model with existing approaches such as mBERT and DistilBERT to fake news detection and discussed its performance, strengths, and limitations. The use of the XLM-RoBERTa model, combined with finetuning, proved advantageous in capturing contextual information and handling the linguistic nuances of the Malayalam language.

References

Shankar Biradar and Sunil Saumya. 2022. IIITDWD@TamilNLP-ACL2022: Transformer-based approach to classify abusive content in Dravidian code-mixed text. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 100–104, Dublin, Ireland. Association for Computational Linguistics.

Shi Chen and Bing Kong. 2021. cs@DravidianLangTech-EACL2021: Offensive language identification based on multilingual BERT model. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 230–235, Kyiv. Association for Computational Linguistics.

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. *CoRR*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

José García-Díaz, Manuel Valencia-Garcia, and Rafael Valencia-García. 2022. UMUTeam@TamilNLP-ACL2022: Abusive detection in Tamil using linguistic features and transformers. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 45–50, Dublin, Ireland. Association for Computational Linguistics.

Nikhil Ghanghor, Parameswari Krishnamurthy, Sajeetha Thavareesan, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2021. IIITK@DravidianLangTech-EACL2021: Offensive language identification and meme classification in Tamil, Malayalam and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 222–229, Kyiv. Association for Computational Linguistics.

RamakrishnaIyer LekshmiAmmal Hariharan and Madasamy Anand Kumar. 2023. Impact of transformers on multilingual fake news detection for tamil and malayalam. In *Speech and Language Technologies for Low-Resource Languages*, pages 196–208, Cham. Springer International Publishing.

Sreelakshmi K, Premjith B, and Soman Kp. 2021. Amrita_CEN_NLP@DravidianLangTech-EACL2021: Deep learning-based offensive language identification in Malayalam, Tamil and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 249–254, Kyiv. Association for Computational Linguistics.

Hariharan LekshmiAmmal, Manikandan Ravikiran, and Anand Kumar Madasamy. 2022. NITK-IT_NLP@TamilNLP-ACL2022: Transformer based model for toxic span identification in Tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 75–78, Dublin, Ireland. Association for Computational Linguistics.

Vasanth Palanikumar, Sean Benhur, Adeep Hande, and Bharathi Raja Chakravarthi. 2022. DE-ABUSE@TamilNLP-ACL 2022: Transliteration as data augmentation for abuse detection in Tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 33–38, Dublin, Ireland. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Eduri Raja, Badal Soni, and Samir Kumar Borgohain. 2023. Fake news detection in dravidian languages using transfer learning with adaptive finetuning. *Engineering Applications of Artificial Intelligence*, 126:106877.

Kayalvizhi S, Thenmozhi Durairaj, Bharathi Raja Chakravarthi, and Jerin Mahibha C. 2022. Findings of the shared task on detecting signs of depression from social media. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 331–338, Dublin, Ireland. Association for Computational Linguistics.

Prasanth S N, R Aswin Raj, Adhithan P, Premjith B, and Soman Kp. 2022. CEN-Tamil@DravidianLangTech-ACL2022: Abusive comment detection in Tamil using TF-IDF and random kitchen sink algorithm. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 70–74, Dublin, Ireland. Association for Computational Linguistics.

Debjoy Saha, Naman Paharia, Debajit Chakraborty, Punyajoy Saha, and Animesh Mukherjee. 2021. Hate-alert@DravidianLangTech-EACL2021: Ensembling strategies for transformer-based offensive language detection. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 270–276, Kyiv. Association for Computational Linguistics.

Siva Sai and Yashvardhan Sharma. 2021. Towards offensive language identification for Dravidian languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 18–27, Kyiv. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Jasper Snoek, Oren Rippel, Kevin Swersky, Ryan Kiros, Nadathur Satish, Narayanan Sundaram, Md. Mostofa Ali Patwary, Prabhat, and Ryan P. Adams. 2015. Scalable bayesian optimization using deep neural networks.

Konthala Yasaswini, Karthik Puranik, Adeep Hande, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. IIITT@DravidianLangTech-EACL2021: Transfer learning for offensive language detection in Dravidian languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 187–194, Kyiv. Association for Computational Linguistics.