Dravidian Fake News Detection with Gradient Accumulation based Transformer Model

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

Abstract

The proliferation of fake news poses a significant challenge in the digital era. Detecting false information, especially in non-English languages, is crucial to combating misinformation effectively. In this research, we introduce a novel approach for Dravidian fake news detection by harnessing the capabilities of the MuRIL transformer model, further enhanced by gradient accumulation techniques. Our study focuses on the Dravidian languages, a diverse group of languages spoken in South India, which are often underserved in natural language processing research. We optimize memory usage, stabilize training, and improve the model's overall performance by accumulating gradients over multiple batches. The proposed model exhibits promising results in terms of both accuracy and efficiency. Our findings underline the significance of adapting state-of-the-art techniques, such as MuRIL-based models and gradient accumulation, to non-English languages to address the pressing issue of fake news.

1 Introduction

In the digital age, the rapid proliferation of fake news has emerged as a formidable challenge, permeating every facet of society and eroding trust in information sources (Raja et al., 2022). Addressing this issue is paramount, and the ability to detect false information, particularly in non-English languages, represents a critical frontier in the combat against misinformation. This research endeavors to present a novel approach to tackle the problem of fake news detection within the linguistic landscape of Dravidian languages.

Dravidian languages, a diverse and linguistically rich group primarily spoken in South India, often find themselves underserved in the realm of natural language processing research (Raja et al., 2023b). The unique characteristics and linguistic nuances of these languages necessitate dedicated attention. Our study aims to bridge this gap by introducing a pioneering solution that harnesses the capabilities of the multilingual representations for Indian languages (MuRIL) (Khanuja et al., 2021) transformer model, a state-of-the-art language representation model designed for multilingual tasks.

Furthermore, we enhance the MuRIL Transformer model's performance by incorporating the gradient accumulation (Lamy-Poirier, 2021) technique. This approach offers multiple advantages, such as optimized memory usage, stabilized training processes, and improvements in overall model efficiency. By accumulating gradients over various batches, we mitigate memory constraints while maintaining high accuracy.

Additionally, we incorporated adaptive learning rate scheduling, a dynamic technique that adjusts the learning rate during training, to ensure our model trains stably and efficiently, eventually improving its ability to converge to a better solution. Our findings highlight the substantial impact of adapting cutting-edge techniques to non-English languages, specifically Dravidian languages (Raja et al., 2024), in the acute and crucial quest to combat the dissemination of fake news. Additionally, we compared the proposed model with other three multilingual models, like multilingualBERT (mBERT) (Devlin et al., 2018), cross-lingual language model for robustly optimized BERT (XLM-RoBERTa) (Conneau et al., 2019), and IndicBERT(Kakwani et al., 2020). This study not only showcases the potential of our approach but also contributes valuable insights to the broader field of natural language processing and misinformation detection, offering a robust foundation for developing tools and strategies to combat misinformation in linguistically diverse landscapes. The main contributions of this research are:

- We introduced a novel approach by combining the MuRIL transformer model with gradient accumulation, effectively reducing GPU mem-

ory usage during training.

- We implemented an adaptive learning rate strategy to improve the training stability and convergence of the model.
- We conducted an in-depth comparative analysis of the proposed model against other multilingual transformer models, such as mBERT, XLM-RoBERTa, and IndicBERT.

2 Related Works

The prevalence of fake news, misinformation, and disinformation in the digital age has attracted substantial research attention in recent years. While the majority of existing studies concentrate on fake news detection and classification in widely spoken languages, there needs to be more research focusing on regional languages, such as those within the Dravidian language family. This section provides a comprehensive overview of the literature on fake news detection, with a specific emphasis on studies pertaining to regional languages.

Till now, limited research has been conducted on the detection of fake news in low-resource languages, and the challenges associated with this issue have been addressed in (Magueresse et al., 2020). A BERT-based multilingual model is employed (Kar et al., 2021) to identify fake news within the context of COVID datasets in Indic languages, resulting in an F1 score of 89%. (Huang and Chen, 2020) introduced an ensemble model for the purpose of fake news detection and explored the challenges associated with cross-domain intractability. Their approach involved a combination of models, including the embedding LSTM network, depth LSTM, linguistic inquiry, and word count in conjunction with a CNN classifier. Additionally, n-gram models were utilized in combination with a CNN classifier to extract relevant features from the news data. To optimize the ensemble learning model, the authors employed the self-adaptive harmony search algorithm, allowing them to determine the most suitable weights for the constituent models.

(Dave et al., 2021) developed a model to identify offensive language in code-mixed YouTube comments across three Dravidian languages. This modal employs TF-IDF character n-grams, leverages pre-trained MuRIL embeddings and utilizes logistic regression and linear SVM for classification tasks. They achieved impressive weighted F1 scores of 64%, 95%, and 71% in Kannada, Malayalam, and Tamil, respectively. (Patankar et al., 2022) employed three methods to detect and classify YouTube comments in Tamil and Tamil-English code-mixed formats: ensemble models, recurrent neural networks, and transformers. Their evaluation of these methods in the Tamil dataset revealed that MuRIL and XLM-RoBERTA were the top-performing models, achieving a macro-averaged F1 score of 43%. Furthermore, MuRIL and M-BERT produced outstanding results in the context of code-mixed data, yielding a macro-averaged F1 score of 45%.

(Hariharan and Anand Kumar, 2023) analyzed to assess the influence of various transformer-based models, such as multilingual BERT, XLM-RoBERTa, and MuRIL, on a dataset created as a component of their research concerning multilingual low-resource fake news classification. They executed a range of experiments encompassing language-specific approaches and the utilization of diverse models, intending to discern the distinct impact of these models on the task at hand. There has been a noticeable gap in addressing reducing the GPU memory usage in transformer models for low-resource fake news detection in languages like Dravidian. To address this gap, we propose a novel solution, integrating gradient accumulation with transformer models for improved fake news detection in Dravidian languages.

3 Methodology

3.1 Dataset

In our research, we conducted extensive validation of the effectiveness of our model using a meticulously curated dataset known as the Dravidian_Fake news dataset. This dataset has been made publicly available on the IEEE Dataport (Raja et al., 2023a). The Dravidian_Fake news dataset is a comprehensive compilation of both genuine and fabricated news articles, primarily in Dravidian languages. Creating this dataset involved a thorough extraction process from various sources, including news websites, government portals, social media platforms, and fact-checking websites. It comprises a substantial collection of 26,000 news articles spanning Tamil, Telugu, Kannada, and Malayalam languages.

To ensure the robustness of our analysis, we initially worked with the original dataset sizes, which contained 6,277, 6,481, 6,278, and 6,311 articles

for Tamil, Telugu, Kannada, and Malayalam, respectively. To mitigate potential biases and ensure a fair evaluation, we selected 6,000 samples from each language subset. This balanced approach was adopted to maintain fairness and counteract any inherent biases that might skew our analysis.

3.2 Model Architecture

In this research, we harnessed the power of the gradient accumulation-based MuRIL Transformer model for fake news detection in Dravidian languages. The model, built upon the cutting-edge BERT architecture, captured linguistic subtleties and contextual information specific to Dravidian languages. Figure 1 shows the training process with gradient accumulation for the transformer model.

Gradient accumulation optimization, often referred to as gradient accumulation over multiple small batches, is a valuable technique applied in deep learning to address the constraints posed by GPU memory when training complex neural networks. This approach enables the model to accumulate gradients over several small batches before executing updates to the model's parameters. Consequently, it offers the dual advantage of reducing memory consumption during training while simultaneously enhancing the model's accuracy (Hermans et al., 2017).

The fundamental principle underlying gradient accumulation optimization involves a sequence of forward and backward passes executed on smaller data batches before the initiation of parameter updates. For instance, when dealing with a batch size denoted as B and the intention to accumulate gradients over N batches, the initial batch is subdivided into N smaller batches, each with a size of B/N. Subsequently, forward and backward passes are conducted on each of these smaller batches. Gradients computed during each of these backward passes are collected and integrated across the N batches. This accumulation of gradients significantly influences the parameter updates of the model. This strategic approach adeptly manages the computational burden and memory requirements during training while facilitating improvements in the model's performance.

In gradient descent (Von Oswald et al., 2023), the core objective is to optimize the loss function L concerning the model parameters denoted as θ. Gradients of the loss are computed using the chain rule of calculus, represented as:

$$\nabla L = \left(\frac{\partial L}{\partial \theta_1}, \frac{\partial L}{\partial \theta_2}, \ldots, \frac{\partial L}{\partial \theta_n} \right)$$

In the traditional approach of batch gradient descent, model parameters are updated as per the equation:

$$\theta \leftarrow \theta - \alpha \left(\frac{1}{\text{batch\_size}} \right) \sum \nabla L$$

Here, the symbol α represents the learning rate, and batch_size signifies the number of data points within each mini-batch. To introduce the concept of gradient accumulation, gradients are accumulated over N batches, where N corresponds to the 'accumulate_grad_batches' hyperparameter. The parameter update equation is suitably adjusted to accommodate this accumulation:

$$\theta \leftarrow \theta - \alpha \left(\frac{1}{\text{batch\_size} \cdot N} \right) \sum \sum \nabla L$$

This modified equation scales the learning rate to account for the accumulation factor, effectively simulating a larger batch size. This adjustment enhances our comprehension of how gradient accumulation optimizes training stability and efficiency in transformer models.

4 Training

In this research, we used the PyTorch Lightning framework (Helwe et al., 2022) with an 8GB GPU. In comparison to PyTorch, the PyTorch Lightning framework showcased notable improvements in terms of speed and efficiency. This significant advancement can be credited to the implementation of mixed precision within the PyTorch Lightning framework. Mixed precision leverages both 32 and 16-bit floating-point formats, effectively curbing memory consumption during model training. The outcome is a remarkable boost in performance, resulting in a noteworthy up-to-threefold increase in speed on contemporary GPUs (Sawarkar, 2022). We employed a transformer model with gradient accumulation to enhance the efficiency of our fake news detection system. Figure 2 shows the four multilingual transformer models with and without gradient accumulation (GA) technique over the Dravidian_Fake dataset, and it clearly describes that the multilingual transformer model with gradient accumulation has less GPU memory usage

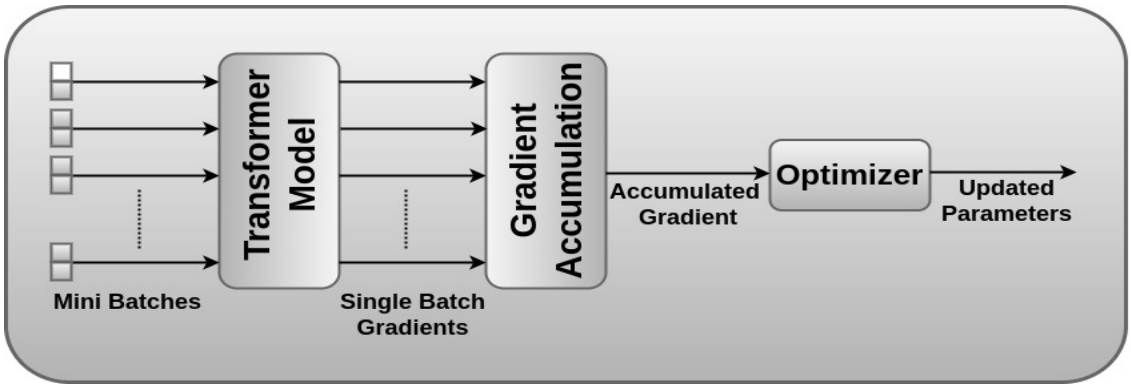

Figure 1: Training process with Gradient Accumulation

compared to the model without gradient accumulation technique. The hyperparameters used in this research are shown in Table 1. In the training process, the gradient accumulation size was configured to be 4, and the batch size was set to 8. This choice of hyperparameters plays a crucial role in optimizing the training of deep neural network models, such as the transformer model used in this study.

Gradient accumulation is a technique used to address memory constraints when training deep learning models, particularly on GPUs with limited memory capacity. Setting the gradient accumulation size to 4 means that instead of updating the model's parameters after processing each batch of data, the gradients are accumulated over four consecutive batches before performing a parameter update. This accumulation effectively simulates a larger batch size, which can lead to more stable training and better convergence. It helps in managing GPU memory efficiently during training, making it feasible to train complex models on hardware with limited memory.

The batch size refers to the number of data samples that are processed together in one forward and backward pass during training. In this case, a batch size of 8 indicates that eight data samples are processed simultaneously in each iteration. A smaller batch size can reduce memory usage but may lead to noisy gradients and slower convergence, while a larger batch size can provide more stable gradients but require more memory. The combination of a batch size of 8 and a gradient accumulation size of 4 allows the training process to effectively simulate a batch size of 32 (8 * 4), striking a balance between memory efficiency and training stability.

We employed the *scheduler* returned by *get_linear_schedule_with_warmup* for adaptive learning. This scheduler adjusts the learning rate during training to help improve model convergence. Specifically, it increases the learning rate linearly during the warm-up phase and then decreases it linearly during the remaining training steps. The dataset was divided into training and testing sets using an 80-20% split, maintaining the integrity of the data distribution. Furthermore, the testing set was further partitioned into a 90-10% split to create a separate validation subset, which was crucial for monitoring the model's performance and preventing overfitting.

Table 1: Hyperparameters of the Proposed Model

Hyperparameters	Values
Accumulate_grad_batches	4
Batch Size	8
Learning rate	0.00001
Optimizer	AdamW
Max_sequence_length	128
Number of epochs	4

5 Experimental Results and Discussions

In this research, we utilized the MuRIL transformer model with gradient accumulation and adaptive learning. We conducted a comparative analysis of our proposed model against three other transformer models: mBERT, XLM-RoBERTa, and IndicBERT. Table 2 describes the models used in our study, along with their accuracy and F1 scores for four Dravidian languages. In the table, Acc, F1, and GA denote accuracy, F1 score, and gradient accumulation, respectively.

In Table 2, the first four models are trained without gradient accumulation but with an adaptive learning rate, while the subsequent four models incorporate gradient accumulation alongside an adap-

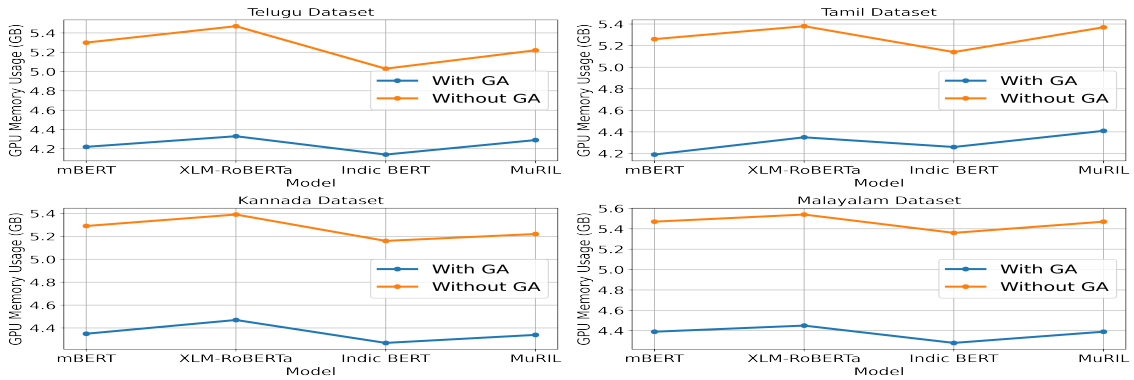

Figure 2: Training process with and without Gradient Accumulation

Table 2: Accuracy & F1-Score results of proposed model and other multilingual models over Dravadian_Fake Dataset

	Telugu		**Tamil**		**Kannada**		**Malayalam**	
Models	Acc	F1	Acc	F1	Acc	F1	Acc	F1
mBERT	88.94	87.59	87.79	86.77	88.80	87.30	86.02	86.90
XLM-RoBERTa	90.62	89.12	88.97	87.48	89.68	88.22	87.15	86.39
IndicBERT	91.39	90.42	89.27	88.69	90.18	89.48	89.96	88.74
MuRIL	92.84	91.35	91.85	91.27	92.91	90.69	91.33	91.05
mBERT_GA	90.21	88.38	88.94	87.82	89.75	87.95	88.93	87.36
XLM-RoBERTa_GA	91.78	90.54	90.83	90.76	91.07	90.21	90.49	89.55
IndicBERT_GA	92.46	90.89	90.97	90.53	91.77	90.38	91.02	90.47
Proposed Model	**93.65**	**93.78**	**92.36**	**92.29**	**93.28**	**92.83**	**92.05**	**91.61**

tive learning rate. Notably, the models with gradient accumulation demonstrated slightly higher accuracy than their non-gradient accumulation counterparts. The primary objective of this research is to reduce GPU memory usage during transformer model training, all while maintaining or improving accuracy and F1 scores.

Comparing the models with and without gradient accumulation, the gradient accumulation models exhibited higher accuracy and F1 scores. Specifically, when contrasted with mBERT, the version with gradient accumulation displayed a 1.5% higher accuracy and a 0.75% higher F1 score. In the case of XLM-RoBERTa, the model with gradient accumulation outperformed its non-gradient accumulation counterpart with a 2% higher accuracy and a 2.5% higher F1 score. Furthermore, when compared to IndicBERT, the IndicBERT model with gradient accumulation achieved a 1.35% higher accuracy and a 1.23% higher F1 score. Finally, relative to MuRIL, the MuRIL model with gradient accumulation exhibited 0.6% higher accuracy and a 1.5% higher F1 score. Significantly, IndicBERT and MuRIL outperformed mBERT and XLM-RoBERTa in terms of accuracy and F1 scores, owing to their specialized training in Indian languages, including Dravidian languages.

6 Conclusion

The research presented in this study addresses the critical issue of fake news detection within the linguistic context of Dravidian languages. By developing the Dravidian fake news classifier, we have demonstrated the efficacy of leveraging state-of-the-art natural language processing techniques, specifically the MuRIL model with gradient accumulation for reducing GPU memory usage in training the model, to detect fake news in languages such as Tamil, Telugu, Malayalam, and Kannada. The experimental results reveal that the Dravidian fake news classifier can remarkably differentiate between fake and legitimate news articles in Dravidian languages. The proposed model's accuracy of 92.83% and F1 score of 92.62% on average of the four Dravidian languages in the test phase attest to its proficiency. The study underscores the importance of language-specific fake news detection by customizing the model for Dravidian languages.

It recognizes the linguistic nuances present in Dravidian languages and the necessity of considering these aspects when addressing misinformation. In the future, we will explore multimodal approaches that combine textual information with other modalities, such as images and videos, for fake news detection in Dravidian languages.

References

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.

Bhargav Dave, Shripad Bhat, and Prasenjit Majumder. 2021. IRNLP_DAIICT@DravidianLangTech-EACL2021:offensive language identification in Dravidian languages using TF-IDF char n-grams and MuRIL. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 266–269, Kyiv. Association for Computational Linguistics.

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.

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.

Chadi Helwe, Chloé Clavel, and Fabian Suchanek. 2022. Logitorch: A pytorch-based library for logical reasoning on natural language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 250–257.

Joeri R. Hermans, Gerasimos Spanakis, and Rico Möckel. 2017. *Accumulated Gradient Normalization*, volume 77 of *Proceedings of Machine Learning Research*, pages 439–454. Proceedings of Machine Learning Research.

Yin-Fu Huang and Po-Hong Chen. 2020. Fake news detection using an ensemble learning model based on self-adaptive harmony search algorithms. *Expert Systems with Applications*, 159:113584.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of EMNLP*.

Debanjana Kar, Mohit Bhardwaj, Suranjana Samanta, and Amar Prakash Azad. 2021. No rumours please! a multi-indic-lingual approach for covid fake-tweet detection. In *2021 grace hopper celebration India (GHCI)*, pages 1–5. IEEE.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages.

Joel Lamy-Poirier. 2021. Layered gradient accumulation and modular pipeline parallelism: fast and efficient training of large language models. *CoRR*, abs/2106.02679.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *CoRR*, abs/2006.07264.

Shantanu Patankar, Omkar Gokhale, Onkar Litake, Aditya Mandke, and Dipali Kadam. 2022. Optimize_prime@dravidianlangtech-acl2022: Abusive comment detection in tamil.

Eduri Raja, Badal Soni, and Samir Kumar Borgohain. 2023a. Dfnd : Dravidian_fake news data. *https://dx.doi.org/10.21227/nj13-t949*.

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

Eduri Raja, Badal Soni, Candy Lalrempuii, and Samir Kumar Borgohain. 2024. An adaptive cyclical learning rate based hybrid model for dravidian fake news detection. *Expert Systems with Applications*, 241:122768.

Eduri Raja, Badal Soni, Shivangi Mishra, and Ashish Dwivedi. 2022. Stand-alone bidirectional encoder representations from transformer-based fake news detection model. In *Proceedings of the International Conference on Computational Intelligence and Sustainable Technologies: ICoCIST 2021*, pages 647–655. Springer.

Kunal Sawarkar. 2022. *Deep Learning with PyTorch Lightning: Swiftly Build High-performance Artificial Intelligence (AI) Models Using Python*. Packt Publishing Ltd.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. 2023. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pages 35151–35174. PMLR.