NITK-IT-NLP@DravidianLangTech-2023: Impact of Focal Loss on Malayalam Fake News Detection using Transformers

Hariharan R L , Anand Kumar M
hariharanrl.197it003@nitk.edu.in
m_anandkumar@nitk.edu.in
Department of Information Technology
National Institute of Technology Karnataka

Abstract

Fake News Detection in Dravidian Languages is a shared task that identifies YouTube comments in the Malayalam language for fake news detection. In this work, we have proposed a transformer-based model with cross-entropy loss and focal loss, which classifies the comments into fake or authentic news. We have used different transformer-based models for the dataset with modifications in the experimental setup, out of which the fine-tuned model, which is based on MuRIL with focal loss, achieved the best overall macro F1-score of 0.87, and we got second position in the final leaderboard.

1 Introduction

Social media is becoming the most popular media through which people share information happening around them. Online social networks such as Twitter, Facebook, Instagram, Weibo, and many others are major channels for people to obtain an enormous amount of information; meanwhile, there can be unauthentic or fake information among them. The central aspect of fake news is that it fully caters to the audience's curiosity, which is why it spreads faster (Jing et al., 2023). Moreover, research proves that false information or fake news spreads faster than authentic news (Vosoughi et al., 2018). Thus it is very urgent to detect fake news and stop its propagation to prevent potential harm.

Fake news detection in Dravidian Languages is a shared task organized by DravidianLangTech, which focuses on detecting fake news from YouTube comments annotated as fake and original. The dataset was retrieved from YouTube comments which are in Malayalam language. We have proposed a transformer-based model that used cross-entropy and focal loss function.

The rest of the paper is organized as follows. First, in section 2, some related works are explained. Followed by the dataset description in section 3. Then in section 4 we explain the different models and the methodology we have used, followed by data preprocessing and model description in section 5 and 6. The result and discussions are explained in section 7 followed by the conclusions and future scope in section 8

2 Related Works

Fake news detection is one of the popular problems in natural language processing tasks. We can see from the literature that an enormous amount of research has been carried out to tackle the propagation of fake news. In this section, we provide some of the works done on mitigating fake in the Dravidian language, especially in Malayalam. We have also included some recent works developed on fake news detection for other languages.

(Sharma and Arya, 2023) has proposed a Hindi fake news detection model that utilizes linguistic feature-based word embedding. They mainly focus on 24 key features which are extracted and derived for the successful detection of fake news. (Verma et al., 2021) has proposed a method that uses both word embeddings and linguistic features for the identification of textual fake news. (Singhal et al., 2022) has created fact-checked data in 13 different Indian languages covering different news stories from 2013 to 2020. The data were mainly retrieved from social media websites. They have also analyzed the characterization of multilingual, multimedia, and multidomain. (Mirnalinee et al., 2022) has created a dataset for fake news detection in the Tamil language. They have scrapped news snippets from various news media and annotated them into binary, fake, and real labels. The news articles scrapped cover different regions like sports, politics, and science, which they manually annotate. (Hariharan and Anand Kumar, 2022) has studied the impact of different transformer-based models using the data they have created using translation. The dataset covers different domains retrieved from

Third Workshop on Speech and Language Technologies for Dravidian Languages,
pages 207–210, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-085-4_029

FakeNewsAMT and Celebrity Fake News, which are translated into Tamil and Malayalam languages using Google Translate.

It is evident from the literature the number of works for the Dravidian language, especially Malayalam, is much less than the other languages, the main reason being the lack of proper annotated data resources. Hence, we should develop a system that could address these challenges.

3 Dataset Description

The shared task provided data retrieved from YouTube comments and annotated for fake news detection (Subramanian et al., 2023). The dataset was mainly in Malayalam, and some had code-mixed content. The dataset was labeled into two classes, 'Fake' and 'Original', and the detailed statistics as shown in Table 1. The test data given had 1019 samples for which predictions will be made.

Table 1: Dataset Statistics

Data	Comments		
	Original	Fake	Total
Train	1658	1599	3257
Validation	409	406	815

4 Methodology

The shared task was to classify the YouTube comment into corresponding labels correctly. We were supposed to get a prediction for the test data with the training and validation data provided. Moreover, we were allowed to use external data to train the model. We have employed transformer-based pre-trained models for training the proposed model. Even though the task was for fake news detection in the Dravidian language, and the dataset given was for Malayalam, we had code-mixed samples. Hence we used the multilingual version of the BERT (Devlin et al., 2019) model for training the data. We further experimented with the multilingual version of RoBERTa (Liu et al., 2019) and the MuRIL (Khanuja et al., 2021) model. We used the binary cross entropy loss function as the loss function in all of our models and have experimented with focal loss.

5 Data Preprocessing

As we were using Malayalam and code-mixed to train the model, we applied some basic preprocessing before training the model. The dataset from youtube comments was cleaned using the Python library cleantext, which helps remove the unknown characters and do the ASCII conversions. As mentioned previously, we were using a pre-trained model based on the transformer; hence we had to tokenize the data according to how the model expected. We use the tokenization method from the hugging face library, corresponding to the underlying pre-trained model.

6 Model Description

We mainly used three pre-trained models for training the multilingual BERT (m-BERT) and MuRIL from Google, multilingual RoBERTa (xlmroberta). The m-BERT and xlmroberta were trained on multilingual data, and the MuRIL model was trained specially for the Indian context. The MuRIL model has been pre-trained with multilingual representations from various Indian languages, and they have augmented data that used the translated and transliterated pairs of documents for training. We used recommended hyperparameters for all the models whose particulars are given in Table 2.

6.1 Focal Loss

We can see that most of the classification models used the Cross-Entropy loss (CE) function for their training and learning. The main idea is to have a function that predicts a high probability for a positive class and a low one for a negative class using some threshold. The equation for CE is as follows.

$$CELp_t = -\sum y_i \times log(p_t) \tag{1}$$

The CE loss focussed on penalizing the model based on how far the prediction is from the actual label by giving equal importance to classes. Moreover, CE loss will be the same for any class; given the predicted probability, it will not change based on the class.

Focal Loss (Lin et al., 2020) is a loss function that balances easy and hard examples or positive and negative samples. It is a dynamically scaled CE loss, where the scaling factor decays to zero as the confidence in the correct class increases. The equation for focal loss is as follows.

$$FL_{p_t} = -\alpha \times (1 - p_t)^{\gamma} \times log(p_t) \tag{2}$$

Focal loss is ultimately a weighted CE loss that

Table 2: Hyperparameter Values

Hyperparameters	m-BERT	XLMRoBERTa	MuRIL
Learning rate	2e-5	2e-5	2e-5
Epochs	10	10	10
Batch	32	32	32
Optimizer	AdamW	AdamW	AdamW
Max length	400	400	400

considers the contribution of each sample to the loss based on the classification error.

7 Results and Discussion

This section explains the different experiments we have conducted for the shared task and the results for the same. We used the pre-trained transformer-based models to build our system. We have fine-tuned the transformer-based models using the hyperparameters mentioned in Table, using AdamW as the optimizer. The experiments were conducted on Tesla P100 16 GB GPU on the Kaggle platform.

We initially begin training the data with the m-BERT model after the necessary preprocessing, as explained in the section. We used cross-entropy loss for the experiments. We used the experimental settings for training the model on XLMRoBERTa and MuRIL. The results are shown in Table 3; we can see that the MuRIL model gives the best result, which XLMRoBERTa and m-BERT follow. The result of MuRIL can be because of the better understanding of the Indian language context by the pre-trained model. Among these, the best model was submitted to the shared task.

Table 3: Validation Data Results on different Models using CE Loss

Model	Precision	Recall	F1-score
m-BERT	0.8604	0.8601	0.8601
XLMRoBERTa	0.8761	0.8761	0.8761
MuRIL	0.8853	0.8846	**0.8846**

We started experimenting with focal loss to see how it can perform on complex samples, as we mentioned in the section 6.1 while explaining the focal loss. Moreover, while training the model with CE loss, we could see that the loss did not reduce below 30-35%. Thus we explored the possibility of focal loss whose results are given in Table 4. However, the model did not perform better than the model with cross-entropy loss. We can see that the MuRIL model performed better than the other two models. MuRIL gave 0.8758 for the focal loss model and 0.8846 for the cross-entropy model. Similar is the case for the rest of the models. As we have to give predictions for test data, we included both model results for the final submission.

Table 4: Validation Data Results on different Models using Focal Loss

Model	Precision	Recall	F1-score
m-BERT	0.8587	0.8550	0.8548
XLMRoBERTa	0.8582	0.8577	0.8576
MuRIL	0.8793	0.8759	**0.8758**

Once the test data results and the labels were released, we tested the same with all our models, whose results are given in Tables 5,6. The model performed almost the same for both cross-entropy and focal loss functions; from the tables, it is clear that there is a difference of nearly 0.5% with them. Moreover, XLMRoBERTa and MuRIL performed the same with a small difference in the scores. The final leaderboard result reflected our MuRIL-based model with focal loss function with an F1 score of 0.8692 (which was given as 0.87).

Table 5: Test Data Results on different Models using CE Loss

Model	Precision	Recall	F1-score
m-BERT	0.8450	0.8449	0.8449
XLMRoBERTa	0.8640	0.8637	**0.8636**
MuRIL	0.8640	0.8635	0.8635

Table 6: Test Data Results on different models using Focal Loss

Model	Precision	Recall	F1-score
m-BERT	0.8408	0.8389	0.8388
XLMRoBERTa	0.8728	0.8697	0.8691
MuRIL	0.8728	0.8693	**0.8692**[a]

[a] 2^{nd} Position in the Leaderboard (Score given as 0.87)

8 Conclusions and Future Scope

Currently, social media is the primary influence among people to deliver whatever information is happening around them. Thus it is essential to address the problem of fake information being circulated on online social media platforms. In this work, we have developed a system for fake news identification on Malayalam data retrieved from YouTube comments. We have leveraged the transformer-based model and the focal loss function to address the fake news detection problem. We achieved an F1-score of 0.87 on test data with the proposed model and got second place in the final leaderboard. In the future, we would improve the model's performance by focusing on the code-mixed aspect and language-based fine-tuning.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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 Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

RamakrishnaIyer LekshmiAmmal Hariharan and Madasamy Anand Kumar. 2022. Impact of transformers on multilingual fake news detection for tamil and malayalam. In *International Conference on Speech and Language Technologies for Low-resource Languages*, pages 196–208. Springer.

Jing Jing, Hongchen Wu, Jie Sun, Xiaochang Fang, and Huaxiang Zhang. 2023. Multimodal fake news detection via progressive fusion networks. *Information Processing and Management*, 60(1):103120.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, et al. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. 2020. Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(2):318–327.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

TT Mirnalinee, Bhuvana Jayaraman, A Anirudh, R Jagadish, and A Karthik Raja. 2022. A novel dataset for fake news detection in tamil regional language. In *International Conference on Speech and Language Technologies for Low-resource Languages*, pages 311–323. Springer.

Richa Sharma and Arti Arya. 2023. Lfwe: Linguistic feature based word embedding for hindi fake news detection. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(6).

Shivangi Singhal, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. 2022. Factdrill: A data repository of fact-checked social media content to study fake news incidents in india. In *Proceedings of the international AAAI conference on web and social media*, volume 16, pages 1322–1331.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, Muskaan Singh, Sandhiya Raja, Vanaja, and Mithunajha S. 2023. Overview of the shared task on fake news detection from social media text. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Pawan Kumar Verma, Prateek Agrawal, Ivone Amorim, and Radu Prodan. 2021. WELFake: Word Embedding over Linguistic Features for Fake News Detection. *IEEE Transactions on Computational Social Systems*, 8(4):881–893.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.