byteSizedLLM@DravidianLangTech 2024: Fake News Detection in Dravidian Languages - Unleashing the Power of Custom Subword Tokenization with Subword2Vec and BiLSTM

Rohith Gowtham Kodali
ASRlytics
Hyderabad, India
rohitkodali@gmail.com

Durga Prasad Manukonda
ASRlytics
Hyderabad, India
mdp0999@gmail.com

Abstract

This paper focuses on detecting fake news in resource-constrained languages, particularly Malayalam. We present a novel framework combining subword tokenization, Sanskrit-transliterated Subword2vec embeddings, and a powerful Bidirectional Long Short-Term Memory (BiLSTM) architecture. Despite using only monolingual Malayalam data, our model excelled in the FakeDetect-Malayalam challenge, ranking 4th. The innovative subword tokenizer achieves a remarkable 200x compression ratio, highlighting its efficiency in minimizing model size without compromising accuracy. Our work facilitates resource-efficient deployment in diverse linguistic landscapes and sparks discussion on the potential of multilingual data augmentation. This research provides a promising avenue for mitigating linguistic challenges in the NLP-driven battle against deceptive content.

1 Introduction

The surge in online social media platforms has transformed communication dynamics, facilitating seamless information exchange, dialogue, and real-time awareness of current events. However, this unprecedented connectivity has also given rise to a worrisome proliferation of misinformation, commonly known as fake news (Subramanian et al., 2023).

In response to this escalating issue, we introduce the Fake News Detection in Dravidian Languages challenge—DravidianLangTech@EACL 2024[1]. This initiative addresses the critical need for robust fake news detection mechanisms, particularly in Dravidian languages. The challenge focuses on developing effective models capable of distinguishing between authentic and fake content across various social media platforms, such as Twitter and Facebook.

Task 2, the FakeDetect-Malayalam shared task, is a crucial platform for researchers addressing the challenge of identifying fake news within Malayalam-language articles. In an era of information overload, precise detection of misinformation is vital for trustworthy communication. The task's primary goal is to motivate participants to develop effective models for accurately classifying fake news into categories like False, Half True, Mostly False, Partly False, and Mostly True. This research paper, inspired by the task, explores innovative approaches for fake news detection in Dravidian languages, focusing on Malayalam, aiming to contribute novel insights and methodologies to advance the state of the art.

Our study focuses on Fake News Detection in South Indian social media, introducing an innovative approach featuring a custom-designed tokenizer and a Bidirectional Long Short-Term Memory (BiLSTM) architecture. A key innovation is our tokenizer, which reduces model sizes, improves efficiency, and addresses challenges related to real-time deployment, enhancing the practicality of deploying detection systems in dynamic online environments.

The implementation of BiLSTM, coupled with our specialized tokenizer, exhibits significant enhancements in fake news detection accuracy, demonstrating heightened sensitivity to linguistic nuances. The use of compact models effectively addresses inference challenges, ensuring swift real-time detection and practical deployment. Our work establishes a framework for Fake News Detection, providing valuable insights into model size, inference speed, and the challenges of real-time deployment in countering online misinformation within Dravidian languages.

This paper explains our approach, technical improvements, and findings, thoroughly investigating the detection of fake news in Dravidian languages. Our goal is to contribute not only to addressing the

[1] https://codalab.lisn.upsaclay.fr/competitions/16055

March 22, 2024 ©2024 Association for Computational Linguistics

challenges of fake news in Malayalam but also to enhancing our understanding of effective detection methods suitable for various linguistic contexts.

2 Related Work

The growing concern over disinformation and propaganda has led to increased research on fake news detection. In recent works, Raja et al. (2023) focused on detecting fake news in Dravidian languages using transfer learning with adaptive fine-tuning, emphasizing different techniques with Transformer-based models. Keya et al. (2022) employed a pretrained BERT model with data augmentation, comparing their results with twelve different models. Goldani et al. (2021) utilized capsule networks with varied architectures and n-gram levels for feature extraction. In the context of languages other than English, studies like Gereme et al. (2021) and Saghayan et al. (2021) explored fake news detection in Amharic and Persian, respectively. Chu et al. (2021) investigated the ability to generalize fake news detection models across languages, finding the BERT model effective. Faustini and Covões (2020) covered multiple languages, including Slavic, Latin, and German, emphasizing the need for fake news identification in resource-poor languages like Dravidian languages. Additionally, Vijjali et al. (2020) addressing COVID-19 fake news and proposing a two-stage automated pipeline employs BERT and ALBERT models for efficient verification. Notably, many studies primarily focus on English and other major languages, highlighting the potential for advancements in resource-poor languages.

3 Dataset

3.1 Embedding Malayalam Dataset

We employed a substantial corpus extracted from the AI4Bharath Malayalam dataset[2], comprising the initial 11,512,628 lines of a 1GB text corpus. This dataset serves as a rich source of diverse linguistic content, fostering the development of robust embeddings for our research endeavors. The dataset exhibits linguistic variety and covers a spectrum of topics relevant to the scope of our research.

3.2 Fake news DravidianLangTech@EACL 2024 Malayalam dataset

The dataset utilized in this challenge is derived from the corpus provided by the workshop organizers (Subramanian et al., 2024). For Task 2, the dataset is divided into 1,669 training samples, 250 development samples, and 250 test samples. This division enables a balanced approach to model training, validation, and evaluation, specifically addressing Task 2's requirements.

4 Methodology

This section unveils the intricate details of our innovative architecture, which seamlessly integrates two crucial components: a dynamic Subword Embeddings module and a robust BiLSTM Tagger module. We embark on a journey through the art of data preprocessing, the finesse of subword tokenization, the mastery of embedding training, and ultimately, the orchestration of our advanced classifier.

4.1 Preprocessing and Tokenization

This section outlines the data preprocessing and tokenization procedures employed for the Shared Task on Homophobia/Transphobia Detection in social media comments, aiming to prepare the data for effective model training and enhance the performance of the homophobia/transphobia detection system.

Our comprehensive preprocessing involved normalization, cleaning (eliminating noise such as URLs and hashtags), and transliteration using the `indic_transliteration` library[3]. This ensured uniform processing across the dataset.

In the dataset, we identified 6,161,116 unique words and 19,155 distinct subword units. After establishing a minimum frequency for embedding training, 14,190 subword units were finalized, showcasing the linguistic diversity within the AI4Bharath text. Utilizing a meticulous preprocessing pipeline and transliterator, we applied subword tokenization to enhance the granularity of linguistic representation, paving the way for a more nuanced embedding model.

4.1.1 Subword Tokenization Details

Our subword tokenizer (VowelToken), utilizes universal linguistic principles derived from vowel boundaries, ensuring accurate segmentation across diverse languages.

Engineered with a rule-based design, VowelToken focuses on consistent vowel boundary patterns

[2] https://github.com/AI4Bharat/indicnlp_corpus

[3] https://github.com/indic-transliteration/indic_transliteration_py

observed in multiple languages, facilitating accurate identification and segmentation of compound words. This design enhances precision in subword tokenization across languages.

4.2 Subword Embeddings Module

The subword embeddings (Subword2Vec) module is responsible for obtaining subword embeddings using the Word2Vec method by Mikolov et al. (2013) from a given corpus. It operates as follows:

The module's initialization involves specifying critical parameters, starting with the vocabulary size (V) that sets the upper limit for subword consideration. Additionally, the minimum frequency parameter (f_{min}) serves as the threshold for subword inclusion based on frequency. The embedding dimension ($d_{subword}$), characterizing the dimensionality of subword embeddings, is also defined. These parameters collectively configure the module during the initialization process, a pivotal aspect of our research.

Subword counts are collected from the corpus to construct a subword vocabulary (S). The subword splitting process is executed based on vowels, excluding subwords with counts below f_{min}. This process is mathematically expressed as:

S = {s | $sisasubword, count(s) \geq f_{min}, |S| \leq V\}$
$= \{s \in \mathcal{W} \mid count(s) \geq f_{min}, |S| \leq V\}$

The subword splitting process involves dividing the input word into subwords based on vowel boundaries. Consonant prefixes and suffixes are included in the subwords when applicable, and special tokens "_" (start of subword) are added to the first letter. Subword embeddings ($E_{subword}$) are initialized as a random matrix with dimensions ($|S|$, $d_{subword}$).

The training phase employs Stochastic Gradient Descent (SGD) (Tian et al., 2023) to train subword embeddings. The objective is to minimize the Mean Squared Error (MSE) loss (L) between subword pairs. The SGD update is expressed as:

$$E^{(t+1)}_{subword} = E^{(t)}_{subword} - \eta \nabla L(E^{(t)}_{subword})$$

Here, t represents the training iteration, η is the learning rate, and ∇L is the gradient of the loss function. Training subword embeddings is a crucial step in refining the model's representation of subword relationships.

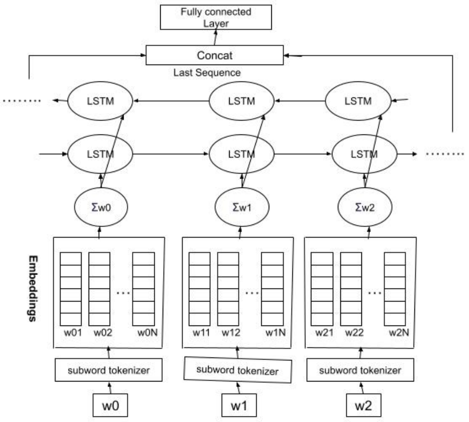

Figure 1: The unfolded architecture of BiLSTM classifier with three 3 word example sample.

4.3 BiLSTM Classifier

The BiLSTM architecture, inspired by Ghosh et al. (2020), plays a crucial role in the fake news classification task. It consists of two essential components: a subword embedding layer and a bi-directional LSTM layer.

The Sub-Word Embedding Layer operates on an input word sequence $x = [w_1, w_2, ..., w_n]$ utilizing a subword embedding function. Each word w_i is mapped to its corresponding subword embeddings, denoted as $w_{i1}, w_{i2}, ..., w_{in}$, where n represents the number of subwords for the i-th word. The final word embedding for w_i, denoted as $\mathbf{e}_i$, is obtained by summing the embeddings of its constituent subwords:

$$\mathbf{e}_i = w_{i1} + w_{i2} + \ldots + w_{in}$$

The output of this layer is a tensor X_{embed} of dimensions $1 \times n \times d_{embed}$, where d_{embed} signifies the size of each word embedding.

$$X_{embed} = [\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_n]$$

Here, $\mathbf{e}_i$ represents the word embedding for the i-th word in the sequence, and n is the length of the sequence.

The subsequent Bi-directional LSTM Layer engages with the embedded sequence X_{embed} to adeptly capture contextual information. Configured with an input size of 100 (matching the embedding size) and a hidden size of 128, the bidirectional LSTM ensures the seamless flow of information both in forward and backward directions. The resulting output, denoted as $blstm\_out$, manifests as a tensor of shape $1 \times n \times 256$.

	Precision	**Recall**	**F1-Score**	**Support**
HALF TRUE	0.68	0.88	0.77	149
MOSTLY FALSE	0.47	0.29	0.36	24
FALSE	0.67	0.35	0.46	63
PARTLY FALSE	0.40	0.29	0.33	14
MOSTLY TRUE	0.0	0.0	0.0	0
Macro Avg	0.55	0.45	0.48	250
Weighted Avg	0.64	0.66	0.63	250
Accuracy			0.66	250

Table 1: Classification Report of the Task2 test set

In essence, the model navigates the input sequence through an embedding layer, harnesses the bidirectional LSTM layer to adeptly capture contextual nuances, and formulates predictions utilizing a streamlined process. The forward pass of the model can be succinctly expressed in mathematical terms as follows:

$$h_i^{(f)}, h_i^{(b)} = BiLSTM(e_{1:i}, e_{i:n}), \quad \forall i \in \{1, ..., n\}$$

$$y = Wh_n^{(f)} + b$$

Here, $h_i^{(f)}$ and $h_i^{(b)}$ represent the forward and backward hidden states at position i, respectively. The BiLSTM function operates on subword embeddings $e_{1:i}$ and $e_{i:n}$ for each i in the sequence. The final prediction y is obtained by a linear transformation with weights matrix W and bias b.

Figure 1 depicts the unfolded architecture of the BiLSTM Classifier module, illustrating a three-word example sample. This design seamlessly integrates subword embeddings with a BiLSTM-based approach, yielding promising results across diverse natural language processing applications. It underscores the model's adaptability and potential in a wide array of contexts.

5 Experimental Setup

Our experimental setup aims to showcase the effectiveness of our proposed approach for fake news detection in Dravidian languages, with a specific focus on Malayalam. Utilizing the custom subword tokenizer "VowelToken," which considers vowel boundaries for efficient tokenization, and Subword2Vec for generating 100-dimensional subword embeddings, we meticulously trained the embeddings on a 1GB monolingual Malayalam text corpus. A Sanskrit transliterator was employed to aid the training process. Despite limiting the training to a single epoch, exceptional results were achieved, with a perplexity of 1.07 on a 10MB development dataset, highlighting the effectiveness of our approach. The final 100-dimensional embedding size after training was 5.92MB, striking a balance between efficiency and accuracy.

To assess the effectiveness of the subword embeddings, we seamlessly incorporated them into our BiLSTM-based model architecture. The ClassificationModel comprises a Sub-Word Embedding Layer, a Bi-directional LSTM Layer, and a Linear Classification Layer, leveraging subword embeddings obtained from VowelToken. The BiLSTM layer is configured with an input size of 100 and a hidden size of 128. Moreover, the model employs the Adam optimizer with a learning rate of 0.001 throughout the training process.

Our model underwent rigorous training on Task 2, the Multi-Class Classifier, to comprehensively assess its performance. This experiment aimed to demonstrate the robustness of our custom subword tokenizer and explore the contribution of subword embeddings within the BiLSTM-based classification model to fake news detection in the context of Dravidian languages, especially Malayalam. Evaluation metrics, including recall, precision, F1 score, and accuracy, were used to measure the model's effectiveness.

6 Experimental results and discussions

Motivated by the challenges of obtaining large multilingual datasets, we intentionally trained our subword embeddings on a limited 1GB monolingual text corpus. Despite this constraint, the embeddings exhibited competitive performance, achieving remarkable accuracy on unseen data.

The classification report (see Table 1) provides a detailed analysis of the model's performance on each category within Task 2. The F1-Scores for each category vary, with "HALF TRUE" achieving the highest at 0.77, indicating strong precision and recall. Categories like "MOSTLY FALSE" and "PARTLY FALSE" have lower F1-Scores, suggesting challenges in correctly classifying instances in these categories.

The overall accuracy of 0.66 showcases the model's competence in classifying news articles into various categories. It is essential to note that the model achieves a reasonable performance, considering the complexity of the task and the limited size of the model (6.87MB), ensuring efficient deployment with approximately 10X faster inference compared to Large Language Models (LLMs).

Team	F1-Score (macro)
CUET_Binary_Hackers	0.5191
CUETSentimentSilles	0.4964
Quartet	0.4868
byteSizedLLM	0.4797
KEC_DL_KSK	0.4763

Table 2: Top 5 results in Fake News Detection - Task 2

It's noteworthy that acquiring training data for embedding training from social media poses a significant challenge. Despite these difficulties and limited data embeddings, our model still achieves top-notch performance, as demonstrated by our ByteSizedLLM team securing the 4th rank overall in the Task 2 competition (see Table 2). This highlights the crucial role of diverse and contextually relevant training data in achieving superior results.

The effectiveness of our subword tokenization is evident in low perplexity after just one epoch, showcasing the model's potential. Unlocking its full capabilities involves leveraging more extensive training data for heightened accuracy and generalization. While expanding the dataset may increase subword tokens, the model size is expected to remain manageable due to the saturation of subword tokens within the initial 1GB of text.

7 Conclusion and future work

Our innovative fake news detection approach for Dravidian languages leverages VowelToken, a custom subword tokenizer capturing vowel nuances within Dravidian languages. This granular understanding improves subword embedding generation, enhancing model performance. The BiLSTM architecture coupled with custom subword embeddings demonstrates efficient information extraction and classification, achieving promising results despite limited training data. The model's compactness facilitates implementation, further contributing to real-world applicability. Future work will explore larger social media and multilingual datasets and investigate advanced embedding techniques like contextualized embeddings, aiming to further improve the model's performance and unlock its full potential in diverse Dravidian language contexts.

References

Samuel Kai Wah Chu, Runbin Xie, and Yanshu Wang. 2021. Cross-language fake news detection. *Data and Information Management*, 5(1):100–109.

Pedro Henrique Arruda Faustini and Thiago Ferreira Covões. 2020. Fake news detection in multiple platforms and languages. *Expert Systems with Applications*, 158:113503.

Fantahun Gereme, William Zhu, Tewodros Ayall, and Dagmawi Alemu. 2021. Combating fake news in "low-resource" languages: Amharic fake news detection accompanied by resource crafting. *Information*, 12(1).

Koyel Ghosh, Dr. Apurbalal Senapati, and Dr. Ranjan Maity. 2020. Technical domain identification using word2vec and BiLSTM. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON): TechDOfication 2020 Shared Task*, pages 21–26, Patna, India. NLP Association of India (NLPAI).

Mohammad Hadi Goldani, Saeedeh Momtazi, and Reza Safabakhsh. 2021. Detecting fake news with capsule neural networks. *Applied Soft Computing*, 101:106991.

Ashfia Jannat Keya, Md. Anwar Hussen Wadud, M. F. Mridha, Mohammed Alatiyyah, and Md. Abdul Hamid. 2022. Augfake-bert: Handling imbalance through augmentation of fake news using bert to enhance the performance of fake news classification. *Applied Sciences*, 12(17).

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

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.

Masood Hamed Saghayan, Seyedeh Fatemeh Ebrahimi, and Mohammad Bahrani. 2021. Exploring the impact of machine translation on fake news detection: A case study on persian tweets about covid-19. In *2021 29th Iranian Conference on Electrical Engineering (ICEE)*, pages 540–544.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, Premjith B, Sandhiya Raja, Vanaja, Mithunajha S, Devika K, Hariprasath S.B, Haripriya B, and Vigneshwar E. 2024. Overview of the Second Shared Task on Fake News Detection in Dravidian Languages. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, Malta. European Chapter of the Association for Computational Linguistics.

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.

Yingjie Tian, Yuqi Zhang, and Haibin Zhang. 2023. Recent advances in stochastic gradient descent in deep learning. *Mathematics*, 11(3).

Rutvik Vijjali, Prathyush Potluri, Siddharth Kumar, and Sundeep Teki. 2020. Two stage transformer model for covid-19 fake news detection and fact checking.