Tayyab@DravidianLangTech 2024:Detecting Fake News in Malayalam LSTM Approach and Challenges

M. T. Zamir[1], M. S Tash[2], Z. Ahani[3], A. Gelbukh[4] and G. Sidorov[5]
Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC)
{[1]mzamir2023,[2]mshahikit2022,[3]z.ahani2023,[4]gelbukh,[5]sidorov}@cic.ipn.mx

Abstract

Global communication has been made easier by the emergence of online social media, but it has also made it easier for "fake news," or information that is misleading or false, to spread. Since this phenomenon presents a significant challenge, reliable detection techniques are required to discern between authentic and fraudulent content. The primary goal of this study is to identify fake news on social media platforms and in Malayalam-language articles by using LSTM (Long Short-Term Memory) model. This research explores this approach in tackling the DravidianLangTech@EACL 2024 tasks.[1]. Using LSTM networks to differentiate between real and fake content at the comment or post level, Task 1 focuses on classifying social media text. To precisely classify the authenticity of the content, LSTM models are employed, drawing on a variety of sources such as comments on YouTube. Task 2 is dubbed the FakeDetect-Malayalam challenge, wherein Malayalam-language articles with fake news are identified and categorized using LSTM models. In order to successfully navigate the challenges of identifying false information in regional languages, we use lstm model. This algoritms seek to accurately categorize the multiple classes written in Malayalam. In Task 1, the results are encouraging. LSTM models distinguish between orignal and fake social media content with an impressive macro F1 score of 0.78 when testing. The LSTM model's macro F1 score of 0.2393 indicates that Task 2 offers a more complex landscape. This emphasizes the persistent difficulties in LSTM-based fake news detection across various linguistic contexts and the difficulty of correctly classifying fake news within the context of the Malayalam language.

1 Introduction

Online social media has made communication easier on a global level, which allows people to seamlessly interact and share information with each other across the globe. In the realm of Natural Language Processing (NLP) (bad, 2021), diverse tasks hold significance, ranging from detecting hate speech(Shahiki-Tash et al., 2023a) and hopeful (Yigezu et al., 2023a; Shahiki-Tash et al., 2023b) sentiments (Tash et al., 2023) to language identification (Tash et al., 2022) and combatting fake news. Researchers leverage various models tailored to these tasks' intricacies. For hate speech detection (Yigezu et al., 2023b), models like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), traditional machine learning (Kanta and Sidorov, 2023), and Transformer-based (Tonja et al., 2022) architectures such as BERT and its variants have been instrumental. On the other hand, the spread of false information and fake news (yig, 2023) is a serious problem brought about by this increased connectivity. Significant doubts about the veracity and credibility of online content have been raised by the purposeful spread of inaccurate or misleading information through a variety of social media platforms. This trend has major ramifications for society cohesion, public confidence, and democratic discourse in addition to endangering the veracity of information.

The DravidianLangTech@EACL 2024 (Subramanian et al., 2024) task was created for this problem, with a specific focus on countering fake news in Dravidian languages. The goal of this innovative project is to use advanced technology, specifically LSTM (Long Short-Term Memory) models (Yigezu et al., 2022), to address the complex problems related to identifying and categorizing fake news(Fazlourrahman et al., 2022; Balouchzahi and Shashirekha, 2020). This initiative aims to achieve two main goals. Initially, Task 1 focuses on social media content classification, with a focus on distinguishing between accurate and false information. The challenge for participants is to create LSTM-based systems that can operate at the

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

Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 113–118
March 22, 2024 ©2024 Association for Computational Linguistics

comment or post level on social media sites like Facebook, YouTube, and Twitter and can distinguish between authentic content and fake content. Second, the goal of Task 2, also referred to as the FakeDetect-Malayalam challenge is to recognize and classify fake news in articles written in Malayalam. The task for participants is to create LSTM-based models that can effectively identify false information in Malayalam and categorize articles into groups such as Mostly True, False, Half True, Mostly False, and Partly False. The initiative aims to progress fake news detection, particularly in the context of Dravidian languages, through these tasks. The results and developments that come from this work will not only help create strong detection systems but also promote credibility, trustworthiness, and dependability in online content, forming a more genuine and informed digital environment.

2 Related Work

Trends in technology and extensive research have been made in the field of fake news detection in response to the spread of misinformation and fake news on online platforms in recent years. Various approaches, strategies, and methods have been investigated in a number of studies to address the widespread problem of fake news, which includes social media platforms and multilingual environments. One popular area of research has been identifying fake news on social media. By using neural networks to detect fake news on Twitter, (Shuaibo et al., 2022) invented the application of deep learning(Ahani et al., 2024) techniques (Zervopoulos et al., 2022).They showed how machine learning models can effectively separate false information from real content by focusing on feature extraction and classification in their study. Furthermore, Kumar et al. (2018) suggested a method for detecting fake news by analyzing the text of social media posts using natural language processing (NLP) techniques (Murugesan, 2019). Their research highlighted the value of sentiment analysis and linguistic characteristics in precisely detecting false information.

Moreover, studies have begun to focus on multilingual settings, recognizing the difficulties presented by false information in tongues other than English. Ruchansky et al. (2017) examined the transferability of models across languages in their investigation of cross-lingual fake news detection. (Ruchansky et al., 2017; Bade and Seid, 2018) Research on the identification of fake news in Dravidian languages is beginning to emerge. In their 2020 study, Vigneshwaran and Soman investigated the detection of fake news in Tamil-language news articles by using machine learning algorithms to categorize the authenticity of the (Huang, 2022). Their research highlighted how crucial Tamil-specific linguistic subtleties are to creating precise detection models.

Furthermore, the use of LSTM model has become more popular in the identification of fake news. Recurrent neural networks are effective at capturing temporal dependencies and contextual information in text; Ma et al. (2019) used LSTM networks to detect fake news in Chinese social media. (Zhang et al., 2018; Bade and Afaro, 2018) Furthermore, the focus has been directed to initiatives aimed at addressing the detection of fake news in languages like Malayalam. Kunnath and Jayaraman (2021) used machine learning models in conjunction with lexical and syntactic features to study the detection of fake news in Malayalam (Mirnalinee et al., 2022) . Their research demonstrated the importance of using language-specific strategies to effectively counteract misinformation.

In support of this research work, the DravidianLangTech@EACL 2024 initiative seeks to expand the reach of fake news detection to Dravidian languages. This initiative addresses a research gap by integrating the LSTM model concentrating on social media content and Malayalam-language articles. Overall, these studies present an overview of the DravidianLangTech@EACL 2024 initiative's pursuit of combating fake news within Dravidian languages by highlighting the various methodologies and approaches used in fake news detection across social media platform.

3 Task Description

This work aims to identify Fake News Detection in Dravidian Languages, as mentioned in the introduction . There are two sub-tasks in this work.

3.1 Task 1

This task requires binary classification of content, both in the native language and Roman, into two groups: Original and Fake. All participants must create systems that can reliably classify content authenticity into these two categories—Original and fake information—regardless of language.

3.2 Task 2

The purpose of Task 2, FakeDetect-Malayalam, is to identify fake news in Malayalam content. To accurately classify articles into False, Half True, Mostly False, Partly False, and Mostly True categories, participants develop language-specific models, highlighting the importance of precise identification within regional languages like Malayalam.

4 Methodology

Due to the complex nature of data for both tasks, it is quite obvious that the proposed model must have different aspects to precisely and accurately predict the fake and original content and similarly for multi-class classification for the second task.

4.1 Data sets

The data sets are obtained from (Subramanian et al., 2023) for both tasks, Task 1 involves training, validation, and test data sets and Task 2 has training and test data sets. Task 1 appears to involve binary classification to differentiate between original and fake content. Task 2 has multi-class classification having 5 classes False, Half True, Mostly False, Partly False, and Mostly True.Figure 1 shows the training samples and figure 2 shows the validation data samples for task1.Figure 3 shows the training data set labels for task2.

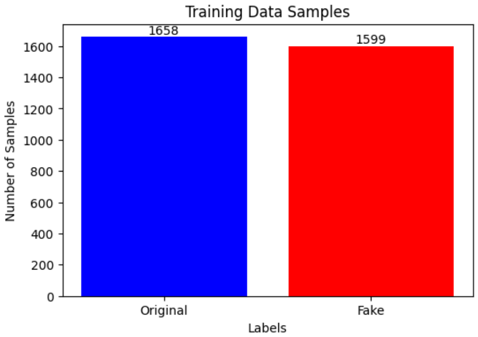

Figure 1: Validation Data Set samples

4.2 Data Preprocessing

Preprocessing includes removing HTML tags, numbers, and symbols (emojis included) from the data once it has been obtained for both tasks. These elements could add unnecessary noise to the data sets, which would impact the analysis. Removing them makes the corpus of texts cleaner, which makes natural language processing (NLP) jobs more accurate. In order for models to concentrate on relevant language patterns and features for better performance and robustness in classification or analysis, this cleaning process is essential to improving the quality of data being used.

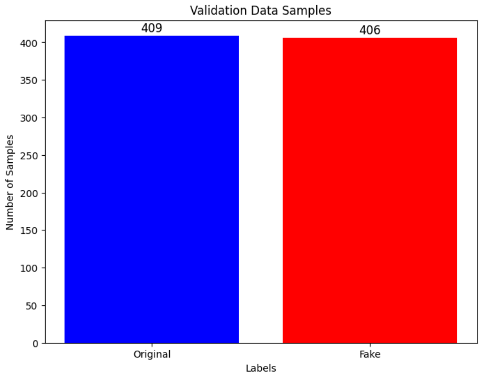

Figure 2: Validation Data Set samples

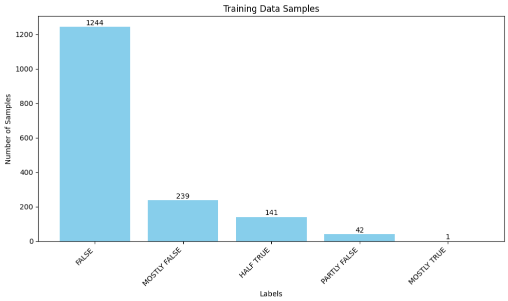

Figure 3: Training Data Set samples

4.3 Model Evaluation

This ability of recurrent neural networks (RNNs) to preserve long-term dependencies makes Long Short-Term Memory (LSTM) networks particularly good at processing sequential data. Input, forget, and output gates are examples of gates that are incorporated into LSTMs to make learning from sequential data more efficient and to mitigate the disappearing or expanding gradient problems that are frequently seen in traditional RNNs.

Task 1 used an LSTM model with a post-RNN dropout layer for Real vs. Fake News Detection (Binary Classification). By controlling excessive learning from the training data and concentrating on identifying patterns within textual sequences, this design avoided overfitting. To prepare the data for the analysis of the LSTM model, the methodology included preliminary processes such as tokenization, sequence padding, and text preprocessing.

In Task 2, focused on Multi class Classification, the LSTM model was configured to handle multiple output categories, enabling the classification of news articles into various classes (e.g., True, false, partially true). Leveraging the LSTM's proficiency in understanding sequential data, similar prepossessing methods were applied, and the LSTM architecture was adapted to accommodate the multiple output classes.

Both tasks showcased the LSTM's efficacy as the primary architecture for text data processing. The LSTM's adeptness in capturing long-term dependencies and intricate patterns within sequences effectively fulfilled the objectives of differentiating between real and fake news in Task 1 and categorizing text into multiple classes in Task 2. The tailored preprocessing steps and LSTM configurations highlighted the versatility and success of LSTM networks in addressing various text classification challenges.

5 Results and Discussions

Different performance outcomes were found when our LSTM-based models were evaluated for both tasks. With an macro f1-score of 0.78, our model shown encouraging performance in Task 1, which focused on differentiating between Real and Fake News (Binary Classification). Due to the presence of dropout layers and proper pre processing, the model was able to identify unique patterns within textual sequences, which resulted in a balanced performance that showed notable results.

On the other hand, our LSTM model faced a more difficult environment in Task 2, which faced Multi class Classification. With an F1-score of 0.2393, the model encountered difficulties in accurately categorizing news articles into multiple classes.

6 Error Analysis

The LSTM model for Malayalam fake news detection demonstrates remarkable accuracy, particularly in true positive identification. However, a notable challenge arises with false positives, mislabeling instances as fake news even in a balanced dataset. This pattern warrants meticulous analysis and adjustment in the model's discriminatory capabilities. Comprehensive evaluations on validation and test sets are crucial for assessing the model's adaptability. Proposed strategic modifications involve fine-tuning parameters and scrutinizing false positive occurrences to enhance overall accuracy and efficacy.

7 Limitations

Utilizing the model LSTM in fake news detection offers improved textual comprehension, but effectiveness may be limited by corpus specificity. Finetuning is crucial to address potential mismatches with unique Malayalam fake news characteristics. The linguistic complexities of Malayalam may hinder the model's ability to discern subtle patterns, requiring further investigation and refinement.

8 Conclusion

In conclusion, Task 1 and Task 2 evaluation of our LSTM-based model highlights both achievements and weaknesses. Task 1 showed excellent performance, obtaining a noteworthy macro F1-score of 0.78 in binary classification, successfully differentiating Real News from Fake News. This success confirms the LSTM model's ability to identify distinct patterns in textual sequences and shows its ability to accurately classify binary data.

With an F1-score of 0.2393 Task 2, which involved Multi-class Classification, highlighted weaknesses. The model had difficulties correctly classifying content multiple classifications, indicating that it was not able to differentiate between different categories. This emphasizes that in order to increase multi-class classification skills, feature representation, data balance, or model refinement changes are required.

The differences in the tasks show the effectiveness of LSTMs in binary classification as well as the challenges that arise in multi class classification. Resolving these complexities requires concerted efforts, such as advanced feature engineering, possible data balancing techniques, or model improvements targeted at improving multi class classification. As we proceed, an iterative process that includes extensive testing, enhanced feature representations, and model optimizations is important.

9 Future work

In order to increase sequence understanding, future work will augment the text classification tasks using LSTM models by combining transformer-based architectures. Furthermore, using larger and large-scale data sets and advanced data balancing techniques to rectify class imbalances may improve the resilience of the model. In order to understand model decisions, more research will need to incorporate interpretability techniques. We will investigate how to enhance and customize models for particular applications by combining domain-specific embeddings with ensemble techniques and transfer learning from pre-trained models. The goal of this multimodal strategy is to improve LSTM models' adaptability and performance in text classification tasks, especially in multi-class settings.

Ethics Statement

We affirm our commitment to ethical research practices and compliance with ACL guidelines in conducting and presenting our study. No ethical concerns or conflicts of interest arose during the course of this research.

Acknowledgments

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONACYT, Mexico, grants 20242138, 20241567, and 20242080 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

References

2021. Natural Language Processing and Its Challenges on Omotic Language Group of Ethiopia, author=Bade, Girma Yohannis. *Journal of Computer Science Research*, 3(4):26–30.

2023. Evaluating the Effectiveness of Hybrid Features in Fake News Detection on Social Media, author=Yigezu, Mesay Gemeda and Mehamed, Moges Ahmed and Kolesnikova, Olga and Guge, Tadesse Kebede and Gelbukh, Alexander and Sidorov, Grigori. In *2023 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, pages 171–175. IEEE.

Zahra Ahani, Moein Shahiki Tash, Yoel Ledo Mezquita, and Jason Angel. 2024. Utilizing deep learning models for the identification of enhancers and super-enhancers based on genomic and epigenomic features. *arXiv preprint arXiv:2401.07470*.

Girma Yohannis Bade and Akalu Assefa Afaro. 2018. Object Oriented Software Development for Artificial Intelligence. *American Journal of Software Engineering and Applications*, 7(2):22–24.

Girma Yohannis Bade and Hussien Seid. 2018. Development of Longest-Match Based Stemmer for Texts of Wolaita Language. *vol*, 4:79–83.

Fazlourrahman Balouchzahi and HL Shashirekha. 2020. Learning Models for Urdu Fake News Detection. In *FIRE (Working Notes)*, pages 474–479.

B Fazlourrahman, BK Aparna, and HL Shashirekha. 2022. Coffitt-covid-19 fake news detection using fine-tuned transfer learning approaches. In *Congress on Intelligent Systems: Proceedings of CIS 2021, Volume 2*, pages 879–890. Springer.

Xiaolei Huang. 2022. Easy adaptation to mitigate gender bias in multilingual text classification. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 717–723, Seattle, United States. Association for Computational Linguistics.

Selam Kanta and Grigori Sidorov. 2023. Selam@ DravidianLangTech: Sentiment Analysis of Code-Mixed Dravidian Texts using SVM Classification. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 176–179.

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.

Manoj Kumar Murugesan. 2019. *Comparative Analysis of Machine learning Algorithms using NLP Techniques in Automatic Detection of Fake News on Social Media Platforms*. Ph.D. thesis, Dublin, National College of Ireland.

Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. Csi: A hybrid deep model for fake news detection. pages 797–806.

Moein Shahiki-Tash, Jesús Armenta-Segura, Zahra Ahani, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023a. Lidoma at homomex2023@ iberlef: Hate speech detection towards the mexican spanish-speaking lgbt+ population. the importance of preprocessing before using bert-based models. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023)*.

Moein Shahiki-Tash, Jesús Armenta-Segura, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023b. LIDOMA at HOPE2023IberLEF: Hope Speech Detection Using Lexical Features and Convolutional Neural Networks. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023), co-located with the 39th Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), CEUR-WS. org*.

Wang Shuaibo, Di Hui, Huang Hui, Lai Siyu, Ouchi Kazushige, Chen Yufeng, and Xu Jinan. 2022. Supervised contrastive learning for cross-lingual transfer learning. In *Proceedings of the 21st Chinese National Conference on Computational Linguistics*, pages 884–895, Nanchang, China. Chinese Information Processing Society of China.

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.

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.

M Shahiki Tash, Z Ahani, Al Tonja, M Gemeda, N Hussain, and O Kolesnikova. 2022. Word Level Language Identification in Code-mixed Kannada-English Texts using traditional machine learning algorithms. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON): Shared Task on Word Level Language Identification in Code-mixed Kannada-English Texts*, pages 25–28.

Moein Tash, Jesus Armenta-Segura, Zahra Ahani, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023. LIDOMA@ DravidianLangTech: Convolutional Neural Networks for Studying Correlation Between Lexical Features and Sentiment Polarity in Tamil and Tulu Languages. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 180–185.

Atnafu Lambebo Tonja, Mesay Gemeda Yigezu, Olga Kolesnikova, Moein Shahiki Tash, Grigori Sidorov, and Alexander Gelbuk. 2022. Transformer-based model for word level language identification in code-mixed kannada-english texts. *arXiv preprint arXiv:2211.14459*.

Mesay Gemeda Yigezu, Girma Yohannis Bade, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023a. Multilingual Hope Speech Detection using Machine Learning.

Mesay Gemeda Yigezu, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023b. Transformer-Based Hate Speech Detection for Multi-Class and Multi-Label Classification.

Mesay Gemeda Yigezu, Atnafu Lambebo Tonja, Olga Kolesnikova, Moein Shahiki Tash, Grigori Sidorov, and Alexander Gelbukh. 2022. Word Level Language Identification in Code-mixed Kannada-English Texts using Deep Learning Approach. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON): Shared Task on Word Level Language Identification in Code-mixed Kannada-English Texts*, pages 29–33.

Alexandros Zervopoulos, Aikaterini Georgia Alvanou, Konstantinos Bezas, Asterios Papamichail, Manolis Maragoudakis, and Katia Kermanidis. 2022. Deep learning for fake news detection on Twitter regarding the 2019 Hong Kong protests. *Neural Computing and Applications*, 34(2):969–982.

Jiawei Zhang, Limeng Cui, Yanjie Fu, and Fisher B Gouza. 2018. Fake news detection with deep diffusive network model. *arXiv preprint arXiv:1805.08751*.