CUETSentimentSillies@DravidianLangTech EACL2024: Transformer-based Approach for Detecting and Categorizing Fake News in Malayalam Language

Zannatul Fardaush Tripty, Md. Arian Al Nafis, Antu Chowdhury, Jawad Hossain, Shawly Ahsan and **Mohammed Moshiul Hoque**
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
{u1804030, u1804111, u1804112, u1704039, u1704057}@student.cuet.ac.bd
moshiul_240@cuet.ac.bd

Abstract

Fake news misleads people and may lead to real-world miscommunication and injury. Removing misinformation encourages critical thinking, democracy, and the prevention of hatred, fear, and misunderstanding. Identifying and removing fake news and developing a detection system is essential for reliable, accurate, and clear information. Therefore, a shared task was organized to detect fake news in Malayalam. This paper presents a system developed for the shared task of detecting and classifying fake news in Malayalam. The approach involves a combination of machine learning models (LR, DT, RF, MNB), deep learning models (CNN, BiLSTM, CNN+BiLSTM), and transformer-based models (Indic-BERT, XLM-R, Malayalam-BERT, m-BERT) for both subtasks. The experimental results demonstrate that transformer-based models, specifically m-BERT and Malayalam-BERT, outperformed others. The m-BERT model achieved superior performance in subtask 1 with macro F1-scores of 0.84, and Malayalam-BERT outperformed the other models in subtask 2 with macro F1-scores of 0.496, securing us the 5^{th} and 2^{nd} positions in subtask 1 and subtask 2, respectively.

1 Introduction

Social media has fundamentally transformed how we receive and exchange information in the digital era. But social media is also a source of false news, misinformation, and content emphasized by sensationalism, manipulation, and propaganda (Rohera et al., 2022). So the adoption of social media is very significant for awareness, but the authenticity of news is the cause of concern as some sources of news are not reliable (Choudhary and Arora, 2021). However, incorrect information may swiftly spread, sway public opinion, cause conflict, and advance agendas. Social media fake news undermines truth, democracy, and social unity (Bharathi et al., 2021). Propaganda raises public safety concerns. Financial losses and stock market fluctuations can result from rumors or false information about companies. So to maintain social cohesion, protect against cyber threats, and promote ethical journalism, it is important to detect fake news. Sometimes fake news can hamper the reputation of individuals, organizations, or businesses. So it is important to identify and correct false information to protect the integrity of affected people. Therefore, automated fake news identification is of utmost priority in today's digital age. Fake news detection has been a prominent subject of study, with academics examining different methodologies, databases, and NLP solutions to handle this issue (Oshikawa et al., 2018). This work aims to develop a system that can classify news into original and fake for subtask 1 and classify a text into four predefined categories for subtask 2. The key contributions of this work are illustrated in the following:

- Developed several ML and DL techniques to detect and categorize fake news.
- Investigated the performance of the models to find the right approach for the classification of social media text and performed in-depth error analysis, offering important insight into classifying text.

2 Related Work

Recent studies have made significant strides in detecting fake news in Dravidian languages. A Dravidian dataset was introduced by Raja et al. (2023), and they utilized unique adaptive learning to fine-tune transformer models. Their work demonstrated the effectiveness of transfer learning algorithms, with transformer models, particularly m-BERT and XLM-RoBERTa, outperforming other approaches. In another study, transformer models, including m-BERT, AL-BERT, BERT, and XLNet, were investigated by Balaji et al. (2023) to detect fraudulent

Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 245–251

content. Among these models, m-BERT exhibited the best performance.

Bala and Krishnamurthy (2023) employed Google's MuRIL model with a curated dataset of labeled Dravidian data to detect fake news. By leveraging fine-tuning techniques, their work showcased the effectiveness of the "mural-base-cased" model in identifying fake news. To detect fake news in Malayalam, Coelho et al. (2023) used LR, MNB, and an ensemble model (MNB, LR, and SVM). Among the three models, the ensemble model performed the best with a macro F1-score of 0.831.

Kumari et al. (2023) utilized fine-tuning techniques on the IndicBERT model (macro F1 score of 0.78) for detecting misinformation in Dravidian languages. They employed SBERT sentence embedding, DNN-based classification, and an ensemble classifier to accurately categorize text. Chakravarthi et al. (2023) focused on categorizing code-mixed social media comments and posts into offensive or not offensive at different levels and presented a multilingual MPNet and CNN fusion model with weighted average F1-scores of 0.85, 0.98, and 0.76 for Tamil, Malayalam, and Kannada, respectively.

Kaliyar et al. (2021) proposed FakeBERT, a BERT-based deep learning strategy, to identify bogus news. They also employed deep learning-based models, including CNN and LSTM. The proposed FakeBERT model outperformed the other models with an accuracy of 0.989. Hossain et al. (2022) employed Logistic Regression to detect the abusive language in Tamil text. The LR and CNN+BiLSTM models outperformed the others, with LR achieving a higher recall value (0.44) than CNN+BiLSTM (0.36).

For the fake news detection task in the Urdu language, Kalra et al. (2022) utilized an ensemble of transformer models. Tula et al. (2021) proposed a multilingual ensemble-based model for identifying offensive content in low-resource Dravidian languages. The mode achieved an F1-score of 0.97, 0.75, and 0.70 for the Malayalam, Tamil, and Kannada datasets, respectively. Monti et al. (2019) proposed a novel automatic fake news detection model based on geometric deep learning. The authors achieved high accuracy for fake news detection with an ROC AUC score of 92.7%.

3 Task and Dataset Description

This shared task Subramanian et al. (2024) was organized by the organizers to detect and classify fake news. The shared task[1] included two subtasks: subtask 1 focused on classifying text as 'Original' or 'Fake' news and subtask 2 targeted to categorize texts into 'False', 'Half True', 'Mostly False', 'Partly False' and 'Mostly True'. For subtask 1, a system was developed to classify texts as fake or original using a corpus created by Malliga et al. (2023). The dataset included 5091 texts from YouTube comments of varying lengths in the Malayalam language. The training, validation, and test sets contained 3257, 815, and 1019 texts, respectively, divided into 'Original' and 'Fake' categories. 'Original' texts comprise 14031 words, while fake texts contain 23198 words (Table 1).

Classes	Train	Valid	Test	Total Words
Original	1658	409	512	14031
Fake	1599	406	507	23198
Total	3257	815	1019	37229

Table 1: Dataset statistics of subtask 1

The aim of subtask 2 was to classify texts into five categories, each defined by the degree of misinformation. The dataset consisted of 1919 texts from Malayalam language YouTube comments. The training set had 1669 texts and the test set had 250 texts (Table 2). Text lengths varied from 3 to 36 words, with an average of 10 words.

Classes	Train	Test	Total Words
False	1246	149	12185
Mostly False	239	63	2380
Half True	141	24	1462
Partly False	42	14	363
Mostly True	1	0	8
Total	1669	250	16398

Table 2: Dataset statistics of subtask 2

4 Methodology

We developed a framework for detecting and classifying fake news in the Malayalam language. Initially, data preprocessing was conducted to clean the data. Features were extracted using TF-IDF

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

(Nayel, 2020) for the machine learning (ML) models, while FastText (Joulin et al., 2016) embeddings were utilized for deep learning (DL) models. Various ML, DL, and transformer-based techniques were subsequently employed for classification purposes. The graphical representation of our methodology is depicted in Figure 1.

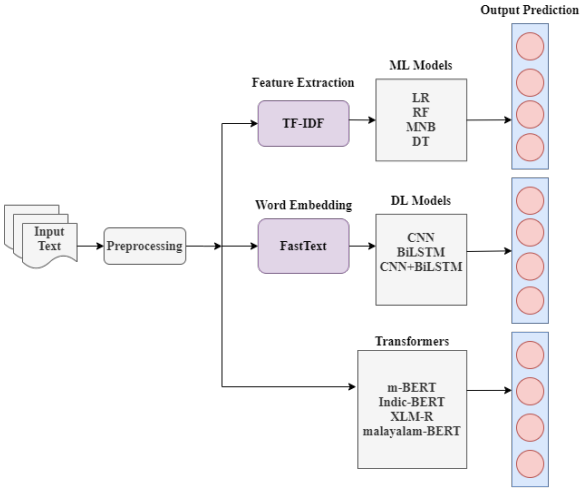

Figure 1: Proposed methodology for fake news detection and classification

4.1 Data Augmentation

In subtask 2, a noticeable class imbalance existed, particularly in the 'Mostly True' class, which contained only one sample. To tackle this, we adopted the back translation technique to augment all classes except the class of 'False'. This technique enhanced dataset balance by iteratively translating sentences from Malayalam to another language and back, as detailed in Table 3.

Classes	Train	Total Words
False	1246	12185
Mostly False	671	6819
Half True	399	4148
Partly False	122	1074
Mostly True	3	21
Total	2441	24247

Table 3: Training set statistics of subtask 2 after augmentation

4.2 Preprocessing

For effective training and evaluation, we conducted preprocessing on datasets, like removing emojis, punctuation, extra spaces, URLs, and numerical texts. We considered the five most frequent stopwords and removed them. For subtask 1, English stopwords in the corpus were also removed. This streamlined preprocessing ensured standardized and refined textual datasets for analysis.

4.3 Training

In this section, we provide a detailed overview of the architectures of various models. The first step in both cases was to extract features using different feature extraction techniques and then apply various machine learning (ML) and deep learning (DL) algorithms. Furthermore, as depicted in Figure 1, the system development also utilized different transformer models.

4.3.1 ML Baseline

TF-IDF values for unigram features have been used as features for training ML models. Various conventional machine learning methods were employed for the detection of fake news. These methods include Logistic Regression (LR), Random Forest (RF), Decision Tree (DT), and Multinomial Naive Bayes (MNB). In the LR and DT models, the regularization parameter (C value) was set to 2. For Random Forest, we implemented 100 estimators (n_estimators = 100) to enhance its predictive performance.

4.3.2 DL Baseline

Three deep learning models CNN, BiLSTM, and CNN+BiLSTM were employed for fake news detection and classification. In the CNN model, the embedding layer was followed by three convolutional layers featuring 64, 32, and 16 filters, each activated by ReLU. The convolution layers were followed by MaxPooling layers for feature reduction. For the BiLSTM model, the embedding layer was followed by two bidirectional LSTM layers with 32 and 16 units. The resulting sequences were flattened and directed into a dense layer with softmax activation for classification. In the CNN+BiLSTM hybrid model, the embedding layer was followed by a convolutional layer with 128 filters and a kernel of 5 and a BiLSTM layer with 32 units with a dropout rate of 0.2.

4.3.3 Transformers

Considering the current trend of transformers, we also utilized pre-trained transformer-based models including XLM-R (Conneau et al., 2019), m-BERT (Joshi, 2022), Indic-BERT (Kakwani et al., 2020),

and Malayalam-BERT (Joshi, 2022). The learning rate was $2e^{-5}$ with a 0.1 warm-up ratio, and stability was improved by doubling gradient accumulation steps to 2. We applied a weight decay of 0.01 and used a linear learning rate scheduler over a 10-epoch training period. We employed the Adafactor optimizer and used a batch size of 16 for both training and evaluation.

5 Experiments and Results

The performance of various methods on the test set is presented in Table 4 and Table 5 for subtask 1 and subtask 2, respectively. From the results displayed in Table 4, it's evident that transformer-based models outperformed ML and DL models in subtask 1, with the m-BERT model achieving the highest macro F1 score of 0.84. Among the DL models, BiLSTM exhibited the highest macro F1 score of 0.782.

Classifier	P	R	F
LR	0.83	0.82	0.82
DT	0.75	0.74	0.74
RF	0.79	0.77	0.76
MNB	0.83	0.83	0.83
CNN	0.714	0.650	0.622
BiLSTM	0.785	0.782	0.782
CNN + BiLSTM	0.714	0.650	0.622
m-BERT	**0.84**	**0.84**	**0.84**
Indic-BERT	0.763	0.747	0.743
XLM-R	0.837	0.837	0.837

Table 4: Performance of various models for the subtask 1, where P, R, and F denote precision, recall, and macro F1-score, respectively

In subtask 2, Malayalam-BERT achieved the highest macro F1 score of 0.496 among transformer models, followed closely by m-BERT (0.467) and IndicBERT (0.309). Among machine learning models, Random Forest demonstrated the best macro F1 score of 0.476. Furthermore, among the deep learning models, the Convolutional Neural Network (CNN) attained the highest macro F1 score of 0.463 compared to the other models.

6 Error Analysis

6.1 Quantitative Analysis:

We utilized a confusion matrix for error analysis for both subtask 1 and subtask 2. The confusion matrix of subtask 1 (Figure 2) showed us a True Positive Rate (TPR) of 82.64% and 85.54% for the

Classifier	P	R	F
LR	0.785	0.360	0.384
DT	0.482	0.451	0.461
RF	**0.796**	0.426	0.476
MNB	0.663	0.366	0.386
CNN	0.466	0.463	0.463
BiLSTM	0.485	0.476	0.441
CNN+BiLSTM	0.353	0.369	0.109
m-BERT	0.529	0.453	0.467
Indic-BERT	0.382	0.314	0.309
Malayalam-BERT	0.589	**0.456**	**0.496**

Table 5: Performance of various models for the subtask 2, where P, R, and F denote precision, recall, and macro F1-score, respectively

'Fake' and 'Original' classes, respectively, which is an indicator that our applied model performed well overall in identifying both the original and fake cases.

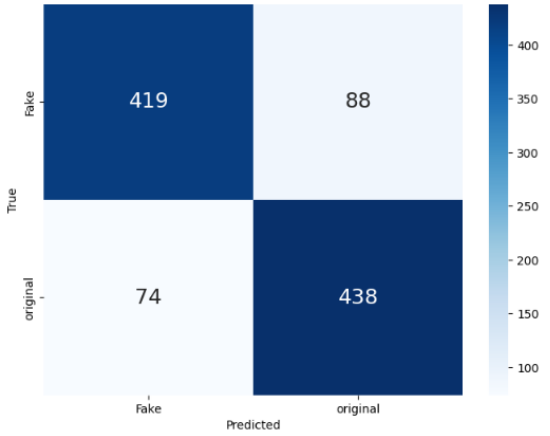

Figure 2: Confusion matrix of m-BERT for subtask 1

By analyzing the confusion matrix of subtask 2 (Figure 3), we found that the False class had the highest TPR of 79.86% due to an adequate amount of data. However, the classes 'Mostly False' and 'Partly False' had lower TPR of 32.26% and 28.57%, respectively. Since the texts of the classes 'Mostly False' and 'False' were similar in context, the model had a tendency to misclassify 'Mostly False' as 'False' and vice versa.

Furthermore, upon analyzing Table 3, we observed that the dataset for subtask 2 was imbalanced. This imbalance caused our model to misclassify instances with the wrong class.

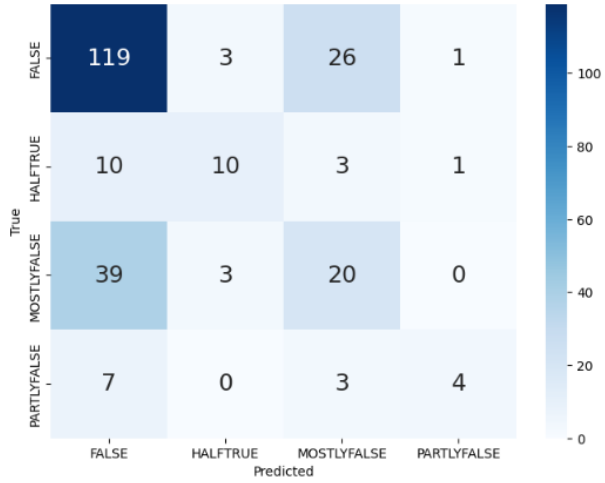

Figure 3: Confusion matrix of Malayalam-BERT for subtask 2

Text Sample	Actual	Predicted
Sample 1. അരഭവിക്കുന്നത് പാവപ്പെട്ട സാധാരണക്കാരായ ജനങ്ങൾ ഇനി പുതിയ പ്രോട്ടോകോൾ വരും (It is the poor common people who are suffering now that the new protocol will come)	Original	Original
Sample 2. ഇവിടെ ഇപ്പോഴും നേരം വെളുക്കാത്ത അന്തങ്ങൾ ഉണ്ട് There are still dawning ends here	Fake	Original

Figure 4: A few examples of predicted outputs by the proposed (m-BERT) model for subtask 1 (here, corresponding english texts were translated using 'Google Translator')

6.2 Qualitative Analysis:

We analyzed some samples to understand the misclassifications made by our model. In Figure 4, the model demonstrated accurate prediction for sample 1, while sample 2 was misclassified. Further analysis of the confusion matrix in subtask 1, as depicted in Figure 2, revealed a lower TPR for the 'Fake' class than the 'Original' class. The model's inability to effectively detect fake news may be attributed to the semantic depth of the content, where the nuanced meanings closely resemble those found in the 'Original' news. Figure 5 illustrates the predicted labels and actual labels generated by the proposed model for subtask 2. Notably, the model demonstrated accurate classification for text samples 1 and 4. However, it exhibited challenges in correctly categorizing text samples 2 and 3. Specifically, text sample 2 was predicted as 'Partly False' instead of its true class, 'Half True', while text sample 4 was predicted as 'Mostly False' instead of its actual class, 'Partly False'. This misclassification can be attributed to a class imbalance within the dataset. The dataset was comprised of a limited number of examples for the 'Half True' (399 samples) and 'Partly False' (122 samples) classes, even after augmentation. In comparison, the classes 'False' and 'Mostly False' were more abundant. This scarcity of samples for 'Half True' and 'Partly False' may pose challenges for the model to effectively learn and generalize patterns associated with these classes, contributing to the observed misclassification.

Text Sample	Actual	Predicted
Sample 1. ചന്ദനക്കുറിയണിഞ്ഞ് വിഎസ് അച്യുതാനന്ദൻ. (VS Achuthanandan dressed in sandalwood.)	False	False
Sample 2. ടി പി ചന്ദ്രശേഖരൻ വധത്തിന് പിന്നിൽ സി.പി.ഐ.എം ആണെന്ന് കെ.ടി ജലീൽ ഏറ്റുപറയുന്നു (KT Jalil confesses that CPIM is behind the assassination of TP Chandrasekaran	Half True	Partly False
Sample 3. വിവിധ വാഹന ടാക്സ് നിരക്കുകൾ സംസ്ഥാന സർക്കാർ കൂട്ടി. (The state government has increased various vehicle tax rates.)	Partly False	Mostly False
Sample 4. ബഹ്റൈനിലെ ഇസ്രയേൽ എംബസിക്ക് പലസ്തീൻ അനുകൂലികൾ തീയിട്ടു. (Palestinian supporters set fire to Israel's embassy in Bahrain.)	Mostly False	Mostly False

Figure 5: A few examples of predicted outputs by the proposed (Malayalam-BERT) model for subtask 2 (here, corresponding english texts were translated using 'Google Translator')

7 Conclusion and Limitations

Our study explored a diverse range of models for detecting and classifying fake news. Through the investigation of four ML models, three DL models, and four transformer models, we gained valuable insights into their performance and effectiveness in these tasks. In subtask 1, m-BERT outperformed other transformer models, including ML and DL models, with a macro F1 score of 0.84, but surprisingly, the LR model with TF-IDF feature extraction came close to 0.82. In subtask 2, Malayalam-BERT outperformed the other ML, DL, and transformer models with a macro F1 score of 0.496. Some DL and ML models came close to this result. CNN with FastText feature extraction came close to it with a macro F1 score of 0.463. Although the system demonstrated strong performance in detecting Malayalam fake news, it faced a significant challenge in classifying multi-class fake news due to a potential data imbalance. To address this limitation, further research and strategies, such as advanced algorithms tailored for imbalanced datasets, are needed to enhance classification accuracy.

References

Abhinaba Bala and Parameswari Krishnamurthy. 2023. Abhipaw@ dravidianlangtech: Fake news detection in dravidian languages using multilingual bert. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 235–238.

Varsha Balaji, B Bharathi, et al. 2023. Nlp_ssn_cse@ dravidianlangtech: Fake news detection in dravidian languages using transformer models. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 133–139.

B Bharathi et al. 2021. Ssncse_nlp@ dravidianlangtech-eacl2021: Offensive language identification on multilingual code mixing text. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 313–318.

Bharathi Raja Chakravarthi, Manoj Balaji Jagadeeshan, Vasanth Palanikumar, and Ruba Priyadharshini. 2023. Offensive language identification in dravidian languages using mpnet and cnn. *International Journal of Information Management Data Insights*, 3(1):100151.

Anshika Choudhary and Anuja Arora. 2021. Linguistic feature based learning model for fake news detection and classification. *Expert Systems with Applications*, 169:114171.

Sharal Coelho, Asha Hegde, G Kavya, and Hosahalli Lakshmaiah Shashirekha. 2023. Mucs@ dravidianlangtech2023: Malayalam fake news detection using machine learning approach. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 288–292.

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.

Alamgir Hossain, Mahathir Bishal, Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022. Combatant@ tamilnlp-acl2022: Fine-grained categorization of abusive comments using logistic regression. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 221–228.

Raviraj Joshi. 2022. L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages. *arXiv preprint arXiv:2211.11418*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

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*.

Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. 2021. Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimedia tools and applications*, 80(8):11765–11788.

S Kalra, P Verma, Y Sharma, and GS Chauhan. 2022. Ensembling of various transformer based models for the fake news detection task in the urdu language.

Kirti Kumari, Shirish Shekhar Jha, Zarikunte Kunal Dayanand, and Praneesh Sharma. 2023. Ml&ai_iiitranchi@ dravidianlangtech: Leveraging transfer learning for the discernment of fake news within the linguistic domain of dravidian language. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 198–206.

S Malliga, Bharathi Raja Chakravarthi, SV Kogilavani, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, and Muskaan Singh. 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*, pages 59–63.

Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M Bronstein. 2019. Fake news detection on social media using geometric deep learning. *arXiv preprint arXiv:1902.06673*.

Hamada A Nayel. 2020. Nayel at semeval-2020 task 12: Tf/idf-based approach for automatic offensive language detection in arabic tweets. *arXiv preprint arXiv:2007.13339*.

Ray Oshikawa, Jing Qian, and William Yang Wang. 2018. A survey on natural language processing for fake news detection. *arXiv preprint arXiv:1811.00770*.

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.

Dhiren Rohera, Harshal Shethna, Keyur Patel, Urvish Thakker, Sudeep Tanwar, Rajesh Gupta, Wei-Chiang Hong, and Ravi Sharma. 2022. A taxonomy of fake news classification techniques: Survey and implementation aspects. *IEEE Access*, 10:30367–30394.

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.

Debapriya Tula, Prathyush Potluri, Shreyas Ms, Sumanth Doddapaneni, Pranjal Sahu, Rohan Sukumaran, and Parth Patwa. 2021. Bitions@ dravidianlangtech-eacl2021: Ensemble of multilingual language models with pseudo labeling for offence detection in dravidian languages. In *Proceedings of the first workshop on speech and language technologies for dravidian languages*, pages 291–299.