Akatsuki-CIOL@DravidianLangTech 2025: Ensemble-Based Approach Using Pre-Trained Models for Fake News Detection in Dravidian Languages

Mahfuz Ahmed Anik[1], Md. Iqramul Hoque[1], Wahid Faisal[1], Azmine Toushik Wasi[1†], Md Manjurul Ahsan[2]
[1]Shahjalal University of Science and Technology, Sylhet, Bangladesh
[2]University of Oklahoma, Norman, OK 73019, USA
[†]**Correspondence:** azmine32@student.sust.edu

Abstract

The widespread spread of fake news on social media poses significant challenges, particularly for low-resource languages like Malayalam. The accessibility of social platforms accelerates misinformation, leading to societal polarization and poor decision-making. Detecting fake news in Malayalam is complex due to its linguistic diversity, code-mixing, and dialectal variations, compounded by the lack of large labeled datasets and tailored models. To address these, we developed a fine-tuned transformer-based model for binary and multiclass fake news detection. The binary classifier achieved a macro F1 score of 0.814, while the multiclass model, using multimodal embeddings, achieved a score of 0.1978. Our system ranked 14th and 11th in the shared task competition, highlighting the need for specialized techniques in underrepresented languages. Our full experimental codebase is publicly available at: ciol-researchlab/NAACL25-Akatsuki-Fake-News-Detection.

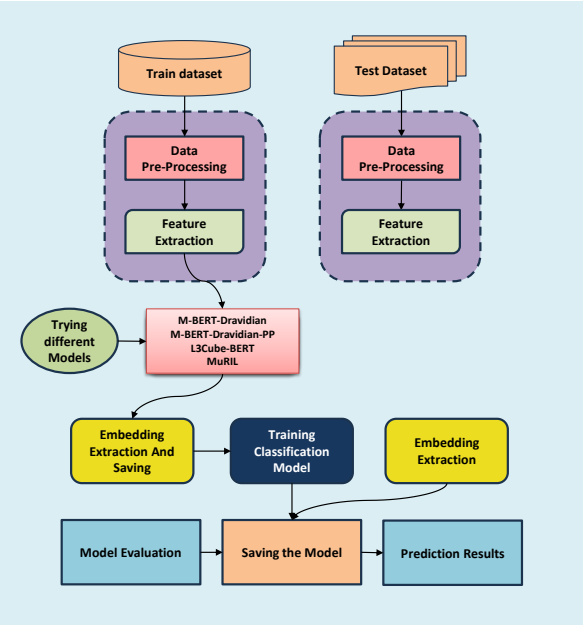

Figure 1: Model architecture, containing tokenizer, pre-trained model, classifier and other components

1 Introduction

Social media has transformed communication and information consumption, becoming a primary news source for many users worldwide. Platforms like Twitter, Facebook, and YouTube allow users to share and engage with information instantly, offering greater convenience, affordability, and timeliness compared to traditional news outlets (Kristian et al., 2024). These platforms solidify their role as preferred news mediums by enabling users to share, comment, and discuss news with their networks (Ku et al., 2019). However, this ease of sharing has also contributed to the widespread spread of fake news—misleading or false information designed to harm individuals, distort public opinion, or increase societal tensions (Fowler, Aug 22, 2022).

The spread of fake news on social media leads to emotional distress, societal polarization, and poor decision-making fueled by misinformation (De Paor and Heravi, 2020). Reports indicate that nearly 50% of Facebook referrals direct users to fake news sites (Pandey, 2018), while only 20% lead to reliable sources (Purcell et al., 2010). Furthermore, only 25% of individuals are confident in distinguishing real from fake news (Lyons et al., 2021), highlighting the need for scalable, automated solutions to address this growing issue.

Detecting fake news in low-resource languages like Malayalam is more challenging due to its linguistic diversity, including dialects, code-mixing, and idiomatic expressions (Thara and Poornachandran, 2021). The scarcity of structured datasets and pre-trained models for Malayalam compounds the problem (Elankath and Ramamirtham, 2023). Social media content, often containing code-mixed text in mixed scripts, poses further challenges for traditional NLP methods. While fake news detection has seen progress in high-resource languages

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 132–138

like English and Spanish, it remains underexplored for low-resource languages like Malayalam (Harris et al., 2024; Wang et al., 2024). Addressing misinformation in Dravidian languages is crucial, as they are spoken by millions in South India and Sri Lanka. The linguistic challenges, including code-switching and dialectal variations, necessitate tailored AI solutions for fairness and accuracy in fake news detection (Subramanian et al., 2025; Devika et al., 2024; Subramanian et al., 2023, 2024). Previous research has shown the effectiveness of ensemble-based models and feature fusion techniques in related tasks (Pillai and Arun, 2024).

In this study, we aim to bridge the gap in fake news detection for Malayalam by leveraging advanced NLP techniques and designing models that address its linguistic diversity and code-mixed nature, solving the first shared task of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages (DravidianLangTech-2025) at NAACL 2025. We use fine-tuned transformer-based architectures and hyperparameter optimization to develop robust solutions for binary and multiclass fake news classification tasks. Our approach tackles Malayalam's unique challenges, such as mixed scripts and dialectal variations, while ensuring scalability and effectiveness. This work provides valuable insights into the potential of transformer-based models for misinformation detection in low-resource languages, laying the foundation for future advancements in this area.

2 Problem Description

Problem Statement. The Fake News Detection in Dravidian Languages shared task focuses on addressing the critical issue of misinformation in low-resource languages, specifically in Tamil and Malayalam. The task is divided into two subtasks, each targeting a unique dimension of fake news detection:

Task 1, aims to classify social media posts from platforms like Twitter, Facebook, and YouTube into one of two categories: fake or original. This task operates at the comment or post level, challenging participants to build models that can effectively discern the authenticity of social media content.

Task 2, titled FakeDetect-Malayalam, targets the identification and classification of fake news within Malayalam-language news articles. Participants are tasked with categorizing news articles into one of five classes: False, Half True, Mostly False, Partly False, or Mostly True. This subtask emphasizes the nuanced detection of misinformation in a language with significant linguistic and cultural diversity.

Dataset. The datasets for the shared task are designed to support the development and evaluation of fake news detection models in Tamil and Malayalam. They are structured to address the unique requirements of the two subtasks.

The dataset for Task 1 consists of social media posts from platforms like Twitter, Facebook, and YouTube, labeled as either fake or original. This binary classification task aims to evaluate the authenticity of posts. The training set contains 3,257 labeled samples, while the validation set has 815 labeled samples for fine-tuning. The test set includes 1,019 unlabeled samples for model evaluation which is shown in 1.

The Task 2 dataset comprises Malayalam news articles categorized into five classes: False, Half True, Mostly False, Partly False, and Mostly True. This multiclass classification task focuses on detecting varying degrees of misinformation. The training dataset includes labeled articles, while the test dataset consists of unlabeled articles for evaluating participant systems.

Table 1: Dataset distribution for Task 1 and Task 2.

	Training	Development	Testing
Task 1	3,257	815	1,019
Task 2	1,615	285	200

3 System Description

3.1 Data Pre-processing

For both Task 1 and Task 2, we converted non-numerical labels into numerical representations for compatibility with machine learning models. Text sequences were tokenized using pre-trained tokenizers to retain domain-specific linguistic patterns and were truncated or padded to uniform lengths (512 tokens for Task 1 and 128 tokens for Task 2). Missing or invalid entries were removed to maintain data integrity. For Task 1, the label distribution was balanced, with 1,658 original and 1,599 fake samples in the train dataset. In contrast, Task 2 had significant class imbalance, with the following distribution: False (976), Mostly False (236), Half True (46), Partly False (129), and Mostly True (133). To address this, we applied a custom oversampling technique, replicating minority class samples to balance the dataset, reducing bias toward the

majority class and improving model generalization. Embeddings for both tasks were extracted from the [CLS] token of the BERT model to preserve contextual representations.

3.2 Models

For **Task 1**, we used "mdosama39/malayalam-bert-FakeNews-Dravidian" (M-BERT-Dravidian), a BERT-based model fine-tuned for Malayalam fake news detection, with 238 million parameters. It effectively captured contextual information for binary classification by extracting [CLS] token embeddings from the last hidden layer, which were then processed using a Multi-Layer Perceptron (MLP) classifier. The MLP had two hidden layers (786 and 512 dimensions) and a softmax output layer, refining the pre-trained features to distinguish between fake and original news.

For **Task 2**, we applied a multimodal embedding strategy, combining embeddings from two additional pre-trained models, "l3cube-pune/malayalam-bert" (L3Cube-BERT) and "Hate-speech-CNERG/hindi-abusive-MuRI" (MuRIL), to capture diverse linguistic and contextual features. This integration leveraged the strengths of multiple models to enable robust predictions for the complex multiclass classification task.

3.3 Implementation Details

We processed text sequences for both tasks using tokenizers from pre-trained models, truncating or padding inputs to 512 tokens for compatibility. For Task 1, we utilized the domain-specific tokenizer of M-BERT-Dravidian, extracting [CLS] token embeddings from the last hidden layer, which were fed into an MLP classifier with hidden dimensions of 786 and 512. For Task 2, we enhanced classification performance using a multimodel embedding strategy by combining embeddings from L3Cube-BERT and MuRIL. These concatenated embeddings were processed through the same MLP architecture, leveraging the linguistic and contextual diversity of the combined models.

Both tasks employed a batch size of 16, the Adam optimizer (learning rate: 0.0001, betas: 0.9, 0.999), and linear learning rate scheduling. Figure 1 illustrates the overall system architecture. A dropout rate of 20% was applied for Task 1 to mitigate overfitting, while no dropout was used for Task 2 to maintain the integrity of combined embeddings. Training and evaluation pipelines were implemented in a GPU-enabled environment using PyTorch (v2.0.0) and Hugging Face Transformers (v4.35.0), ensuring efficient computation. Table 2 summarizes the hyperparameters, hidden dimensions, batch sizes, and performance metrics for all models.

4 Experimental Findings

4.1 Training and Validation Results

For **Task 1**, our best-performing model, M-BERT-Dravidian, achieved a training F1 score of 0.9794 and a validation F1 score of 0.8304, as shown in Table 2. These results demonstrate strong performance with effective generalization from the training data to the validation set. The minimal gap between training and validation scores highlights the robustness of the model, indicating no significant overfitting during training.

For **Task 2**, the best validation F1 score was achieved by the ensemble of MuRIL and L3Cube-BERT, which obtained a training F1 score of 0.9890 and a validation F1 score of 0.4115. This result highlights the ensemble's ability to handle the linguistic diversity and class imbalance challenges inherent to Task 2. The slight performance improvement over other models indicates the potential benefits of combining features from multiple pre-trained models for low-resource languages.

4.2 Test Results

As shown in Table 3, our model achieved a macro F1 score of 0.814 for Task 1 and 0.1978 for Task 2 on the test dataset. For Task 1, our score is close to the highest score of 0.898, and above the mean (0.7805) and median (0.832), demonstrating strong performance in binary classification. The minimum score of 0.334 further highlights the model's effectiveness. In contrast, Task 2 presented greater challenges, with a score of 0.1978, below the mean (0.3244) and median (0.2593), and far behind the top-performing system's score of 0.6283. These results emphasize the model's robustness in Task 1 and reveal the complexities of multiclass classification in low-resource, code-mixed settings.

4.3 Ablation Studies

We evaluated several models for Task 1 and Task 2, as shown in Table 2. For Task 1, M-BERT-Dravidian achieved the best validation F1 score of 0.8304, showing strong generalization with balanced precision and recall. The combined model using embeddings from M-BERT-Dravidian,

Table 2: Hyperparameter Settings and Performance Metrics for Task 1 and Task 2 Train and Validation Dataset.

Task	Model	Max Length	Batch Size	Hidden Dim	LR	Dropout	Train Acc	Train F1	Val Acc	Val F1
Task 1	M-BERT-Dravidian	512	16	[786, 512]	0.0001	0.2	0.9794	0.9794	0.8307	0.8304
Task 1	M-BERT-Dravidian, L3Cube-BERT, MuRIL	512, 512, 512	8	[786, 512]	0.0001	0.0	0.9942	0.9942	0.827	0.827
Task 1	L3Cube-BERT	786	16	[786, 512]	0.0001	0.3	0.9975	0.9975	0.8258	0.8255
Task 2	M-BERT-Dravidian-PP	512	16	[768, 512]	0.0001	0.35	0.7771	0.7720	0.5474	0.3916
Task 2	M-BERT-Dravidian, L3Cube-BERT	512	16	[768, 512]	0.0001	0.35	0.8758	0.8743	0.5921	0.4097
Task 2	MuRIL, L3Cube-BERT	512, 512	8	[786, 512]	0.0001	0.5	0.9890	0.9890	0.6351	0.4115

Table 3: F1 Score (Macro) on Test Dataset

Macro F1	Maximum	Minimum	Mean	Median	Our Score
Task 1	0.898	0.334	0.7805	0.832	0.814
Task 2	0.6283	0.1667	0.3244	0.2593	0.1978

L3Cube-BERT, and MuRIL had a slightly lower validation F1 score of 0.827, with a high training F1 score of 0.9942, indicating overfitting. L3Cube-BERT achieved a validation F1 score of 0.8255, demonstrating good training performance but less robustness. For Task 2, traditional methods like Bag of Words and TF-IDF (Dai et al., 2024; Deo et al., 2024) were less effective in complex multiclass tasks. Among individual models, M-BERT-Dravidian-PP achieved the best validation F1 score of 0.3916 and a training F1 score of 0.7720. Combining models improved performance, with the L3Cube-BERT and MuRIL ensemble achieving a validation F1 score of 0.4115 and a training F1 score of 0.9890. Another combination, M-BERT-Dravidian-PP and L3Cube-BERT, achieved a validation F1 score of 0.4068, despite strong training performance (F1: 0.9783). These results highlight the effectiveness of model ensembles for multiclass classification in Task 2.

For Task 1, the best validation F1 score of 0.8304 was achieved at epoch 46. Early in training, the model showed overfitting with a high training F1 score of 0.9874 at epoch 1, while the validation F1 score lagged at 0.8110. However, generalization improved, reaching 0.8232 at epoch 38 before peaking at epoch 46. For Task 2, M-BERT-Dravidian-PP achieved its best validation F1 score of 0.3916 at epoch 19, with a training F1 of 0.7720. Combined models performed better, with the L3Cube-BERT and MuRIL ensemble achieving the highest validation F1 score of 0.4115 at epoch 27. Another combination, M-BERT-Dravidian-PP and L3Cube-BERT, peaked at epoch 31 with a validation F1 score of 0.4068, but was slightly less robust than the top ensemble. These results emphasize the importance of training duration and model synergy for optimal performance.

5 Discussion

This study explored fake news detection in Malayalam through binary and multiclass classification. In Task 1, our fine-tuned transformer model achieved a macro F1 score of 0.814, effectively distinguishing fake from authentic posts with limited labeled data. Task 2 was more challenging, with a macro F1 score of 0.1978, requiring classification into nuanced categories like "Half True" and "Partly False." This task highlighted the complexities of dialectal variations, class imbalance, and code-mixed text in Malayalam. The performance gap between binary and multiclass classification shows the need for larger datasets, enhanced augmentation, and class-aware loss functions. Our study demonstrates the potential of transformer models and multimodal embeddings for Malayalam's linguistic diversity.

6 Conclusion

This study investigated fake news detection in Malayalam through binary and multiclass classification tasks. The binary classifier achieved a strong macro F1 score of 0.814, showcasing the effectiveness of transformer-based models in identifying misinformation in social media posts. However, the lower performance in multiclass classification (macro F1 score of 0.1978) underscores the challenges of categorizing nuanced misinformation, which requires deeper contextual understanding and tailored strategies. These findings highlight the importance of diverse training data, robust preprocessing, and context-aware approaches to address linguistic complexities such as code-mixing and dialect variations. Future research should aim to advance fake news detection in low-resource languages for greater effectiveness and inclusivity.

Limitations

Despite promising results, limitations persist. The small dataset size, especially for multiclass classification, restricts the model's ability to capture Malayalam's dialects and script variations. Although oversampling addressed class imbalance, it introduced risks of bias and overfitting (Gosain and Sardana, 2017). Differentiating similar misinformation categories remains challenging, requiring architectures capable of finer semantic distinctions. Dependence on pre-trained models risks propagating biases from their training data. While ensemble methods boosted performance, they increased computational complexity, limiting scalability. Lastly, curated datasets may not fully reflect real-world social media complexities, emphasizing the need for diverse data and adversarial training to enhance generalization (De Paor and Heravi, 2020).

Broader Impact Statement

Developing robust fake news detection systems for low-resource languages like Malayalam can significantly curb misinformation in underrepresented communities, fostering informed decision-making and social harmony. Addressing linguistic challenges like code-mixing and dialectal diversity contributes to inclusive AI solutions, bridging resource gaps in NLP. These advancements promote media literacy and trust in digital platforms, mitigating societal polarization and ensuring equitable access to reliable information.

Acknowledgement

We express our sincere gratitude to Computational Intelligence and Operations Laboratory (CIOL) for their invaluable guidance, unwavering support, and continuous assistance throughout this journey. We are deeply appreciative of their efforts in organizing the CIOL Winter ML Bootcamp (Wasi et al., 2024), which provided an enriching learning environment and a strong foundation for collaborative research. The research mentoring and structured support offered by CIOL played a pivotal role in shaping this work, fostering innovation, and empowering participants to contribute meaningfully to the field of computational linguistics.

References

Shuying Dai, Keqin Li, Zhuolun Luo, Peng Zhao, Bo Hong, Armando Zhu, and Jiabei Liu. 2024. AI-based nlp section discusses the application and effect of bag-of-words models and tf-idf in nlp tasks. *Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023*, 5(1):13–21.

Saoirse De Paor and Bahareh Heravi. 2020. Information literacy and fake news: How the field of librarianship can help combat the epidemic of fake news. *The Journal of Academic Librarianship*, 46(5):102218.

Saoirse De Paor and Bahareh Heravi. 2020. Information literacy and fake news: How the field of librarianship can help combat the epidemic of fake news. *The Journal of Academic Librarianship*, 46(5):102218.

Tula Kanta Deo, Rajesh Keshavrao Deshmukh, and Gajendra Sharma. 2024. Comparative study among term frequency-inverse document frequency and count vectorizer towards k nearest neighbor and decision tree classifiers for text dataset. *Nepal Journal of Multidisciplinary Research*, 7(2):1–11.

K Devika, B Haripriya, E Vigneshwar, B Premjith, Bharathi Raja Chakravarthi, et al. 2024. From dataset to detection: A comprehensive approach to combating malayalam fake news. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 16–23.

Syam Mohan Elankath and Sunitha Ramamirtham. 2023. Sentiment analysis of malayalam tweets using bidirectional encoder representations from transformers: a study. *Indonesian Journal of Electrical Engineering and Computer Science*, 29(3):1817–1826.

Gary Fowler. Aug 22, 2022. Council Post: Fake News, Its Impact And How Tech Can Combat Misinformation — forbes.com. [Accessed 26-01-2025].

Anjana Gosain and Saanchi Sardana. 2017. Handling class imbalance problem using oversampling techniques: A review. In *2017 international conference on advances in computing, communications and informatics (ICACCI)*, pages 79–85. IEEE.

Sheetal Harris, Hassan Jalil Hadi, Naveed Ahmad, and Mohammed Ali Alshara. 2024. Fake news detection revisited: An extensive review of theoretical frameworks, dataset assessments, model constraints, and forward-looking research agendas. *Technologies*, 12(11):222.

Natalia Kristian, Dana Indra Sensuse, and Sofian Lusa. 2024. The role of social media functions in enhancing knowledge sharing with user engagement and information quality. *Jurnal Indonesia Sosial Teknologi*, 5(10).

Kelly YL Ku, Qiuyi Kong, Yunya Song, Lipeng Deng, Yi Kang, and Aihua Hu. 2019. What predicts adolescents' critical thinking about real-life news? the roles of social media news consumption and news media literacy. *Thinking Skills and Creativity*, 33:100570.

Benjamin A Lyons, Jacob M Montgomery, Andrew M Guess, Brendan Nyhan, and Jason Reifler. 2021. Overconfidence in news judgments is associated with false news susceptibility. *Proceedings of the National Academy of Sciences*, 118(23):e2019527118.

Neha Pandey. 2018. Fake news: A manufactured deception, distortion and disinformation is the new challenge to digital literacy. *Journal of Content, Community and Communication*, 4(8):15–21.

Aditya R Pillai and Biri Arun. 2024. A feature fusion and detection approach using deep learning for sentimental analysis and offensive text detection from code-mix malayalam language. *Biomedical Signal Processing and Control*, 89:105763.

Kristen Purcell, Lee Rainie, Amy Mitchell, Tom Rosenstiel, and Kenny Olmstead. 2010. Understanding the participatory news consumer. *Pew Internet and American Life Project*, 1:19–21.

Malliga Subramanian, , B Premjith, Kogilavani Shanmugavadivel, Santhia Pandiyan, Balasubramanian Palani, and Bharathi Raja Chakravarthi. 2025. Overview of the Shared Task on Fake News Detection in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Malliga Subramanian, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Santhiya Pandiyan, Prasanna Kumar Kumaresan, Balasubramanian Palani, B Premjith, K Vanaja, S Mithunja, K Devika, et al. 2024. Overview of the second shared task on fake news detection in dravidian languages: Dravidianlangtech@ eacl 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 71–78.

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.

S Thara and Prabaharan Poornachandran. 2021. Transformer based language identification for malayalam-english code-mixed text. *IEEE Access*, 9:118837–118850.

Xinyu Wang, Wenbo Zhang, and Sarah Rajtmajer. 2024. Monolingual and multilingual misinformation detection for low-resource languages: A comprehensive survey. *arXiv preprint arXiv:2410.18390*.

Azmine Toushik Wasi, MD Shakiqul Islam, Sheikh Ayatur Rahman, and Md Manjurul Ahsan. 2024. Ciol presnts winter ml bootcamp. 6 December, 2024 to 6 February, 2025.

A Appendix

A.1 Error Analysis

For **Task 1**, fig 2 shows that 433 Fake samples were correctly labeled, with 74 misclassified as Original. Conversely, 443 Original samples were correct, but 69 were wrongly labeled as Fake. For **Task 2**, fig 3 142 out of 149 False samples were accurately identified. However, the model struggled with more nuanced classes: 16 out of 24 Half True were misclassified, 78.57% of Partly False were incorrectly labeled, and Only 41.27% of Mostly False samples were correct. This bias toward False likely stems from its dominance (three-fourths) in the development data. Sample prediction errors and actual labels are included in table 4. Incorrect predictions largely arise from limited context in short or slang-laden Malayalam text, data imbalance where minority classes (e.g., "Mostly False") are often misclassified, and subtle semantic overlaps between similar labels (e.g., "Fake" vs. "Original"). Such subtleties are challenging for the model to detect without sufficient training examples or language-specific fine-tuning, highlighting the need for data augmentation, balanced class distribution, and more extensive contextual cues to improve classification accuracy.

Table 4: Incorrect Predictions in Text Classification

Task	Text Sample	Predicted	Actual
Task-1	Sample:പരാജയം	Fake	original
Task-1	Sample: ചുവന്ന ട്രൗസർ ഇട്ടാൽ കൊറോണ വരില്ല എന്ന് അറിയില്ലേ ഗമ്മേ	original	Fake
Task-1	Sample: താബിലീഗ് ക്കാർ പിന്നെ അവർ ആവർ ത്തിച്ചി രു ന്നങ്കിൽ നമുക് പറയാൻ മായിരുന്നു. എന്നഈ ഗുണ്ടി മേളം നിയന്ത്രി ക്കാൻ കഴിഞ്ഞില	Fake	original
Task-2	Sample: മഞ്ഞ് ഉരുകുന്നില്ല, കറുത്തിരുണ്ടു പൊള്ളുന്നു	Mostly False	False
Task-2	Sample: ബഫല്ലോ മഞ്ഞുവീഴ്ചയെത്തുടർന്ന് കുടുങ്ങിയ ഒരു കാർ ഫോട്ടോ കാണിക്കുന്നു	Half True	False
Task-2	Sample: ബിബിൻ ജോർജിനെ കോളേജിൽ നിന്നും അപമാനിച്ച് ഇറക്കിവിട്ട സംഭവം ചർച്ചയാകാത്തത് ഇറക്കി വിട്ടയാൾ മുസ്ലിമായതിനാൽ.	False	Mostly False

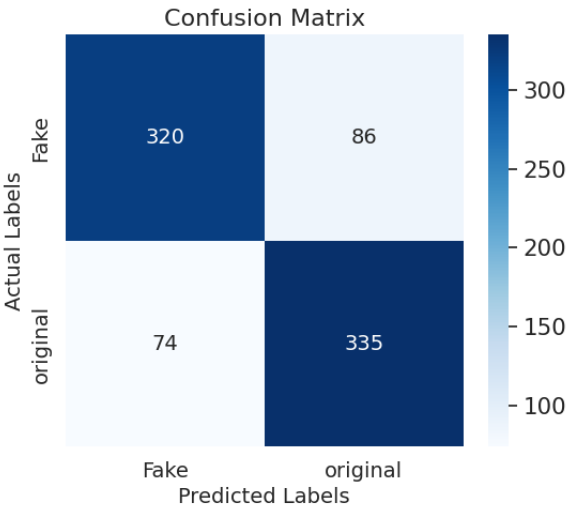

Figure 2: Confusion Matrix of task 1

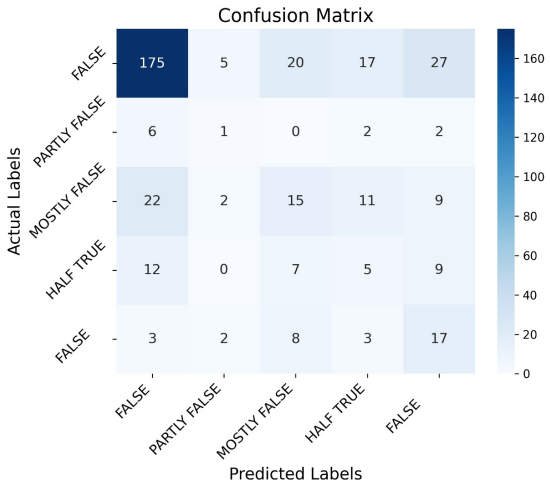

Figure 3: Confusion Matrix of task 2