MUCS@DravidianLangTech2023: Malayalam Fake News Detection Using Machine Learning Approach

Sharal Coelho[a], **Asha Hegde**[b],
Kavya G[c], **Hosahalli Lakshmaiah Shashirekha**[d]
Department of Computer Science, Mangalore University, Mangalore, India
{[a]sharalmucs,[b]hegdekasha,[c]kavyamujk}@gmail.com
[d]hlsrekha@mangaloreuniversity.ac.in

Abstract

Social media is widely used to spread fake news which can lead to individuals or group making wrong judgments based on fake news. The fake news can even create confusion, panic, anxiety, etc., leading to taking inappropriate actions by the individuals. Fake news creators with their tactics may use legitimate sources and mimic the style of reputed publications to create fake news, making it difficult and challenging to identify such content. To address the challenges of detecting fake news in this paper, we - team MUCS, describe the Machine Learning (ML) models submitted to ”Fake News Detection in Dravidian Languages” at DravidianLangTech@RANLP 2023 shared task. Three different models, namely: Multinomial Naive Bayes (MNB), Logistic Regression (LR), and an Ensemble model (MNB, LR, and Support Vector Machine (SVM)) with hard voting, are trained using Term Frequency - Inverse Document Frequency (TF-IDF) of word unigrams, to detect fake news in code-mixed Malayalam text. Among the three models ensemble model performed better with a macro F1-score of 0.831 and placed 3rd rank in the shared task.

1 Introduction

With the overwhelming growth of social media like Twitter, Facebook, YouTube, etc., and the ease with which the information can be shared widely and quickly on these social media platforms (Ahmed et al., 2017; Chakravarthi et al., 2022a,b; Chakravarthi, 2023), creation and sharing of fake news has seen the unprecedented growth. The anonymity of users on social media has given a chance to fake news spreaders to divert people’s beliefs, trust, and opinions by intentionally spreading fake information. Usually rumors and fake news spread fast and damage personal relationships and social connections (Kaliyar et al., 2021). Further, they may also cause anxiety and emotional distress through unfavorable perceptions, scrutiny from the public, and social isolation (Sadeghi et al., 2022). In order to prevent harm and discomfort from fake news, to the users, organizations, and communities, identifying and filtering out such fake news automatically has become the need for the day (Khanam et al., 2021).

The majority of the fake news detection systems have focused on high-resource languages like Spanish and English (Hegde et al., 2022c) giving no or less importance for low-resource Dravidian languages, such as Tulu, Malayalam, Tamil, Telugu, and Kannada, due to lack of resources (Hegde et al., 2022a). Among the low-resource languages, Malayalam is relatively spoken by a smaller population in Indian states of Kerala and the Lakshadweep Islands (Thara and Poornachandran, 2022). Unlike other Dravidian languages, Malayalam has its own linguistic complexities, including dialect variations, word semantics, idiomatic expressions, and so on. These complexities can make it harder to process and analyze the Malayalam text.

As there are no guidelines to create any post/comment on social media, users usually combine words and sub-words belonging to more than one language they know, leading to code-mixed text. Further, they may use more than one script to create the post/comment. These factors make it difficult to process the code-mixed texts. Learning approaches that work well for monolingual text may not give good results for code-mixed texts. Further, there are no pretrained models/specific techniques for code-mixed texts in low-resource languages.

The ”Fake News Detection in Dravidian Languages” shared task organised at DravidianLangTech@RANLP 2023[1] (Subramanian et al., 2023) promotes fake news detection in code-mixed Malayalam text. To address the challenges of de-

[1] https://codalab.lisn.upsaclay.fr/competitions/11176

Third Workshop on Speech and Language Technologies for Dravidian Languages, pages 288–292, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-085-4_042

tecting fake news in Malayalam, in this shared task, we - team MUCS, implemented three distinct models: i) MNB, ii) LR, and iii) Ensemble models (LR, MNB, and SVM) with hard voting, trained with TF-IDF of word unigrams.

The rest of the paper is organised as follows: Section 2 gives a brief description of the related work. While Section 3 explains the methodology, Section 4 is about experiments and outcomes. Finally, the paper concludes in Section 5 with future work.

2 Related Work

Fake news detection in low-resource languages and code-mixed low-resource languages are getting prominence gradually. Researchers have tried to explore several techniques to identify fake news using available benchmarked corpora in low-resource languages. A brief description of few of the relevant works are given below:

A novel Kurdish fake news corpus created by Azad et al. (2021) to detect fake news in Kurdish language consists of two datasets (i) Crawled fake news and (ii) Texts that are altered from real news. TF-IDF of words is used to train ML models (Naïve Bayes (NB), SVM, LR, Decision Tree (DT), and Random Forest (RF)) to detect fake news. The SVM classifier outperformed all other classifiers with an accuracy of 88.71% for 'Crawled fake news' and LR classifier outperformed all the other algorithms on 'Texts that are altered from real news' with an accuracy of 83.26%. Hegde and Shashirekha (2021) explored ensemble (RF, MLP, Gradient Boosting (GB), and Adaptive Boosting) model with soft voting for Urdu fake news detection. Using a combination of TF-IDF of word unigrams, character n-grams in the range (2, 3), and fastText vectors, to train the ensemble model, they obtained a macro F1-score of 0.552 and an accuracy of 0.713%. To detect fake news in Malayalam language, Bijimol and Santhosh (2022) proposed a model using Passive Aggressive classifier - an online learning algorithm, trained with TF-IDF of words and achieved an accuracy of 98.4%.

Balouchzahi et al. (2021) developed an ensemble model (LinearSVM, LR, Multilayer Perceptron (MLP), XGB, and RF) with soft voting, trained with TF-IDF of char and word n-grams in the range (1, 3) and (2, 5) respectively. They applied feature selection techniques (Chi-square, Mutual Information Gain (MIG), and f_classif) to select the relevant features and obtained a macro F1-score of 0.592. In another work, Balouchzahi and Shashirekha (2020) developed an ensemble model (MNB, LR, and MLP) with hard voting for the binary classification of fake news in Urdu. Using term frequency of word and character n-grams in the range (1, 2) and (1, 5) respectively to train the ensemble model, they obtained a macro F1-score of 0.770. The approach to detect fake news in the political domain on the "Liar" dataset is proposed by Khanam et al. (2021). They trained ML models (XGB, RF, NB, k-Nearest Neighbors (k-NN), DT, and SVM) using TF-IDF of word n-gram features and obtained an accuracy of 75% using XGB and 73% accuracy using both SVM and RF models.

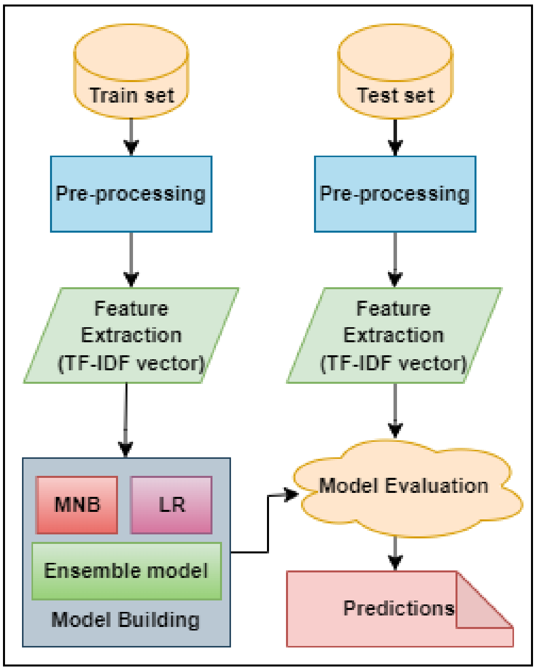


Figure 1: The proposed framework of ML classifiers

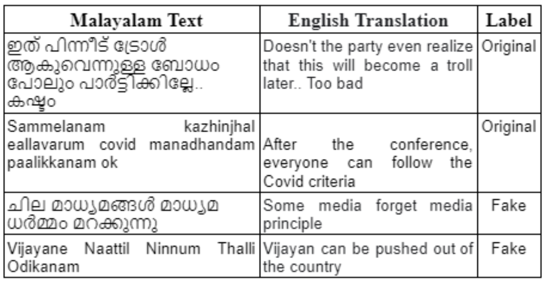
The related work reveals that several ML algorithms trained with TF-IDF of word and char n-grams are explored to detect the fake news in low-resource Dravidian languages. However, as the performance of many of the existing works are low, there is scope to develop models to detect fake news in code-mixed low-resource Dravidian languages.

3 Methodology

The framework of the proposed ML models to identify fake news detection in code-mixed Malayalam text is shown in Figure 1 and the steps involved in the proposed methodology are described below:

Classes	Train set	Development set	Test set
Original	1658	409	512
Fake	1599	406	507
Total	3257	815	1019

Table 1: Class-wise distribution of the dataset



Malayalam Text	English Translation	Label
ഇത് പിന്നീട് ട്രോൾ ആകുവെന്നുള്ള ബോധം പോലും പാർട്ടിക്കില്ലേ.. കഷ്ടം	Doesn't the party even realize that this will become a troll later.. Too bad	Original
Sammelanam kazhinjhal eallavarum covid manadhandam paalikkanam ok	After the conference, everyone can follow the Covid criteria	Original
ചില മാധ്യമങ്ങൾ മാധ്യമ ധർമ്മം മറക്കുന്നു	Some media forget media principle	Fake
Vijayane Naattil Ninnum Thalli Odikanam	Vijayan can be pushed out of the country	Fake

Table 2: Samples of code-mixed Malayalam text from the given dataset

3.1 Pre-processing

Malayalam code-mixed data consists of noise such as punctuation, alphanumeric, and special characters (slash, brackets, ampersands, etc.) which are removed during pre-processing. Text written in Roman script is lowercased and emojis are converted to their corresponding English text as they convey emotions which will be useful for classification. The pre-processed data is used for feature extraction.

3.2 Feature Extraction

TF-IDF is used to preserve the relative importance of a word within a document (Hegde et al., 2022b). A higher TF-IDF score indicates that a term is important within a specific document while being relatively less common in the entire corpus. In the proposed work, TF-IDF vectors of word unigrams is obtained from the pre-processed data using TfidfVectorizer[2]. 15,280 word unigrams are obtained from Train set to train the classifiers.

3.3 Model Building

The three ML models: i) MNB, ii) LR, and iii) Ensemble of ML classifiers (MNB, LR, and SVM) with hard voting, are proposed to identify fake news in code-mixed Malayalam text. The strength of MNB model is its capacity to effortlessly handle word occurrences and distribution, capturing distinctive patterns in the text (Abbas et al., 2019). From the training set, the classifier learns the frequency of each class and each word within a class and applies this to the test sample to determine the most likely class from the words it contains. LR model works by transforming the linear combination of extracted features from the given samples through the logistic function, yielding a probability score representing the given data point belonging to a certain class. Ensemble models are a group of diversified classifiers designed with the aim of overcoming the weakness of one classifier with the strength of the others. An Ensemble of ML classifiers (MNB, LR, and SVM) with hard voting is applied to obtain the benefits of SVM classifier to handle complex decision boundaries and high-dimensional data, LR classifier for its simplicity and probabilistic interpretation, and MNB classifier for its efficiency in text-based categorization. This improves the performance of the ensemble models as compared to individual classifiers.

Classifier	Precision	Recall	Accuracy	Macro F1-score
MNB	0.831	0.831	0.831	0.830
LR	0.820	0.819	0.819	0.819
Ensemble model	**0.831**	**0.831**	**0.831**	**0.830**

Table 3: Performance of the proposed models

4 Experiments and Results

The goal of the shared task is to classify the given Malayalam code-mixed text into ”Original” or ”Fake” news. The statistics of the Malayalam code-mixed dataset for fake news detection provided by the shared task organisers is shown in Table 1. This dataset contains user-generated text extracted from various social media platforms such as Twitter, Facebook, etc. These texts in Malayalam and/or English will be written in Malayalam and/or Roman scripts. The sample texts from the dataset along with the English translation are shown in Table 2. Predictions on the Test set are evaluated by the organizers of the shared task based on macro F1-scores. The performance of the proposed models for the Development set and Test sets in terms of precision, recall, accuracy, and macro F1-score are shown in Table 3. Among the proposed models, MNB and Ensemble model obtained better results, both with a F1-score of 0.831 securing 3rd rank in the shared task. The comparison of the macro F1-scores of all the participating teams of the shared

[2] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

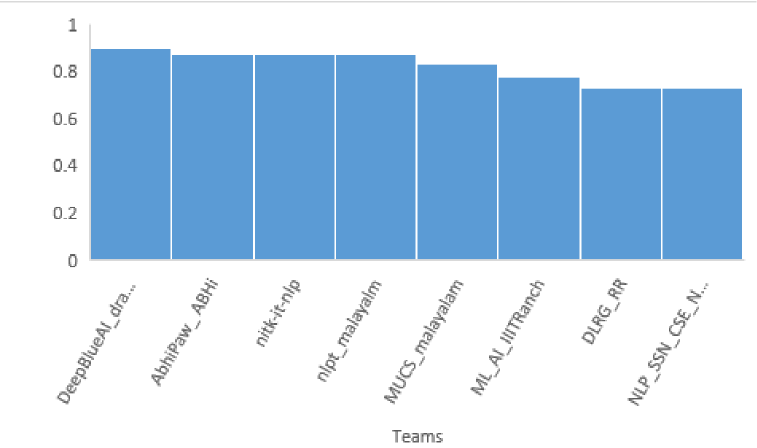
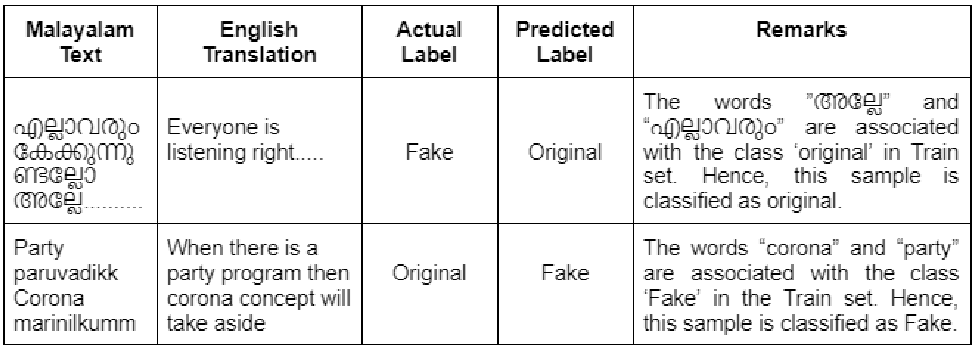


Figure 2: Comparison of macro F1-scores of the proposed ensemble model with other participants' models



Malayalam Text	English Translation	Actual Label	Predicted Label	Remarks
എല്ലാവരും കേക്കുന്നു ണ്ടല്ലോ അല്ലേ..........	Everyone is listening right.....	Fake	Original	The words "അല്ലേ" and "എല്ലാവരും" are associated with the class 'original' in Train set. Hence, this sample is classified as original.
Party paruvadikk Corona marinilkumm	When there is a party program then corona concept will take aside	Original	Fake	The words "corona" and "party" are associated with the class 'Fake' in the Train set. Hence, this sample is classified as Fake.

Table 4: Sample misclassified texts from the Test set with predictions generated by ensemble model

task are shown in Figure 2. The misclassified samples along with their English translation, remarks, true and predicted labels for ensemble model are shown in Table 4.

5 Conclusion and Future work

In this paper, we describe the three models: MNB, LR, and Ensemble (LR, RF, and SVM) classifiers with hard voting, submitted to the ”Fake News Detection in Dravidian Languages” at DravidianLangTech@RANLP 2023 shared task. The proposed models are trained with TF-IDF of word unigrams, for detecting fake news in code-mixed Malayalam texts. Among the three models, ensemble model obtained 3rd rank with a macro F1-score of 0.831. As a future work, fake news detection in low-resource languages like Tulu, Kannada, and other Dravidian languages will be explored.

References

Muhammad Abbas, K Ali Memon, A Aleem Jamali, Saleemullah Memon, and Anees Ahmed. 2019. Multinomial Naive Bayes Classification model for Sentiment Analysis. *IJCSNS Int. J. Comput. Sci. Netw. Secur*, 19(3):62.

Hadeer Ahmed, Issa Traore, and Sherif Saad. 2017. Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. In *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments: First International Conference, ISDDC 2017, Vancouver, BC, Canada, October 26-28, 2017, Proceedings 1*, pages 127–138. Springer.

Rania Azad, Bilal Mohammed, Rawaz Mahmud, Lanya Zrar, and Shajwan Sdiqa. 2021. Fake News Detection in Low-resourced Languages “Kurdish language” using Machine Learning Algorithms. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(6):4219–4225.

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

Fazlourrahman Balouchzahi, Hosahalli Lakshmaiah Shashirekha, and Grigori Sidorov. 2021. Ensembled Feature Selection for Urdu Fake News Detection. In *CEUR Workshop Proceedings*, volume 3159, pages 1117–1126. CEUR-WS.

TK Bijimol and Anit Sara Santhosh. 2022. Malayalam Fake News Detection using Machine Learning. In

National Conference on Emerging Computer Applications, volume 4.

Bharathi Raja Chakravarthi. 2023. Detection of homophobia and transphobia in Youtube comments. *International Journal of Data Science and Analytics*, pages 1–20.

Bharathi Raja Chakravarthi, Adeep Hande, Rahul Ponnusamy, Prasanna Kumar Kumaresan, and Ruba Priyadharshini. 2022a. How can we detect homophobia and transphobia? experiments in a multilingual code-mixed setting for social media governance. *International Journal of Information Management Data Insights*, 2(2):100119.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Thenmozhi Durairaj, John McCrae, Paul Buitelaar, Prasanna Kumaresan, and Rahul Ponnusamy. 2022b. Overview of the shared task on homophobia and transphobia detection in social media comments. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 369–377, Dublin, Ireland. Association for Computational Linguistics.

Asha Hegde, Shubhanker Banerjee, Bharathi Raja Chakravarthi, Ruba Priyadharshini, Hosahalli Shashirekha, John Philip McCrae, et al. 2022a. Overview of the Shared Task on Machine Translation in Dravidian Languages. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 271–278.

Asha Hegde, Sharal Coelho, Ahmad Elyas Dashti, and Hosahalli Shashirekha. 2022b. MUCS@ Text-LT-EDI@ ACL 2022: Detecting Sign of Depression from Social Media Text using Supervised Learning Approach. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 312–316.

Asha Hegde, Sharal Coelho, and Hosahalli Shashirekha. 2022c. MUCS@DravidianLangTech@ACL2022: Ensemble of Logistic Regression Penalties to Identify Emotions in Tamil Text. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 145–150.

Asha Hegde and Hosahalli Lakshmaiah Shashirekha. 2021. Urdu Fake News Detection Using Ensemble of Machine Learning Models.

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.

Z Khanam, BN Alwasel, H Sirafi, and Mamoon Rashid. 2021. Fake News Detection Using Machine Learning Approaches. In *IOP conference series: materials science and engineering*, volume 1099, page 012040. IOP Publishing.

Fariba Sadeghi, Amir Jalaly Bidgoly, and Hossein Amirkhani. 2022. Fake News Detection on Social Media Using A Natural Language Inference Approach. *Multimedia Tools and Applications*, 81(23):33801–33821.

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. 2022. Social Media Text Analytics of Malayalam–English Code-Mixed using Deep Learning. *Journal of big Data*, 9(1):45.