DOI: 10.21917/ijsc.2024.0476

COMPARISON OF EXPLAINABILITY OF MACHINE LEARNING BASED MALAYALAM TEXT CLASSIFICATION

S. Akshay and Manu Madhavan

Department of Computer Science and Engineering, Indian Institute of Information Technology, Kottayam, India

Abstract

Text classification is one of the primary NLP tasks where machine learning (ML) is widely used. Even though the applied machine learning models are similar, the classification task may address specific challenges from language to language. The concept of model explainability can provide an idea of how the models make decisions in these situations. In this paper, The explainability of different text classification models for Malayalam language, a morphologically rich Dravidian language predominantly spoken in Kerala, was compared. The experiments considered classification models from both traditional ML and deep learning genres. The experiments were conducted on three different datasets and explainability scores are formulated for each of the selected models. The results of experiments showed that deep learning models did very well with respect to performance matrices whereas traditional machine learning models did well if not better in the explainability part.

Keywords:

Model Explainability, Text Classification, Malayalam Language, Low Resource Languages, Natural Language Processing

1. INTRODUCTION

Amidst the rapid evolution of artificial intelligence (AI) and machine learning (ML), understanding how computer models work has become important. Artificial intelligence (AI) is used for lots of tasks, like understanding text. It is important to understand why an AI model makes specific choices, especially when it is working with written text. This understanding is key to trusting and using AI models effectively.

This study focuses on exploring how these machine learning models function, in Malayalam text classification and compare different deep learning architectures and machine learning models based on their explainability and performance. Malayalam, spoken primarily in the Indian state of Kerala, is a language that presents unique challenges in the field of text classification. Its morphological complexity, prevalence of code-mixing, and relatively free word order add layers of difficulty to the task. Furthermore, as a low-resource language with limited digital text corpora and NLP tools, the scarcity of datasets, coupled with the lack of specialized tools such as stemmers, lemmatizers etc., makes the development of robust models and facilitating machine understanding for Malayalam particularly challenging. So, this research looks into finding new ways to understand how different machine learning and deep learning models make decisions when dealing with Malayalam text. Sometimes described as "black boxes", most deep learning models and some the machine learning models (some exceptions being support vector machine, decision tree, random forest, Naive Bayes etc.) generate decisions without explicitly providing the reasoning behind them. By demystifying these black boxes, this research is directed at understanding the decision-making process and identifying the factors that influence their choices.

Understanding the model makes decisions can be crucial in model's Transparency and Trust, Algorithmic Fairness, Model Debugging and Improvement, Insights into Feature Importance and Human-AI Collaborations.

The work starts with classification of Malayalam text. Three publicly available Malayalam text dataset were selected for the experiments. Exploratory data analysis (EDA) was conducted which provided insights that guide topic modelling and word embedding. The text classification was performed using popular machine learning algorithms and custom deep learning models. The model explainability was analysed using LIME [1] and SHAP [2] and validated using the results of EDA topic modelling. In a more advanced analysis, the influence of a text in class selection for models are compared with key texts and relevant topics found in earlier stages of EDA, employing BERT and similarity measures. This in-depth assessment ensures a comprehensive understanding of model performance and transparency, making these phases crucial in evaluating overall effectiveness of different models.

The following are the main contributions of this paper.

- As per our best knowledge and belief this is the first attempt to compare the explainability of text classification in Malayalam language.
- The paper introduces a cosine similarity-based method for comparing the explainability of the text classification model using the results of topic modelling, LIME prediction and BERT embedding.

2. RELATED WORK

The survey of related works focused on two areas: (i) explainability of ML models (ii) text classification of Malayalam. Numerous studies have explored the concept of explainability in machine learning models, particularly in the realm of text classification tasks. However, relevant works related to the Malayalam text classification could not be found.

Doshi-Velez and Kim [3] introduced the concept of "interpretable machine learning" and classified it into two dimensions: transparency and post-hoc interpretability. Transparency encompasses inherently interpretable models like decision trees and rule-based classifiers. Post-hoc interpretability, on the other hand, involves generating explanations for complex models after making predictions. Techniques such as LIME (Local Interpretable Model-Agnostic Explanations) [1] and SHAP (Shapely Additive Explanations) [2] have gained popularity for their ability to provide post-hoc interpretability.

Within the domain of text classification, Ribeiro et al. [4] proposed an approach called "anchors that generates rule-based explanations for individual predictions. They demonstrated its effectiveness in explaining text classifiers for tasks such as sentiment analysis and topic classification. The works by Kurasinski et al. [5], Mahoney [6] and Zhao et al. [7] were used explainability techniques for various text classification tasks such as fake news detection, legal document classification etc. Amin Nayebi et al. [8] discussed explainable AI methods for clinical and Biomedical text documents. This dealt with LIME, SHAP anchors among others and gave an insight into which works well. It is worth noting that existing works on explainability in text classification predominantly focus on English language models. In Malayalam, limited research has been conducted. Given the unique characteristics of the Malayalam language, including its script, morphology, and specific linguistic nuances, it is crucial to explore explainability techniques tailored specifically to Malayalam text classification. Therefore, this research paper aims to contribute to the field by investigating and comparing the explainability of machine learning-based Malayalam text classification models using a diverse set of techniques.

3. METHODOLOGY

In this section, machine learning pipeline which includes dataset selection, preprocessing, model selection, evaluation is followed, then the explainability part is done and analyzed. The Fig.1 depicts the detailed methodology of the work.

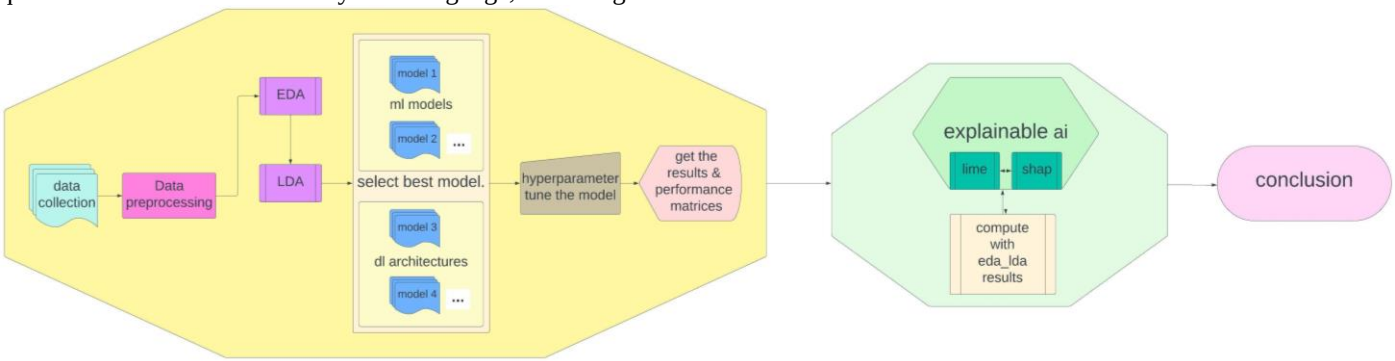

Fig.1. Methodology

3.1 DATASET

The following Three datasets were used for the tests:

- Malayalam news headings which have Malayalam news articles and their news categories (i.e., business, entertainment and sports) as the target variable.

There are 4958 samples.

- Utterences corresponding to the top three frequent intents('Calender_set', 'play_music', 'weather_query') from the malayalam utterences in 'amazon massive dataset' (taking utterances as the feature and their intents as target).

There are 2022 samples.

- a translated version of chat sentiment dataset (contains texts and their corresponding sentiments (positive, negative and neutral) as target variable)

There are 546 samples.

All these datasets are available in Kaggle. The class labels and data distribution of these three datasets are shown in Fig.2.

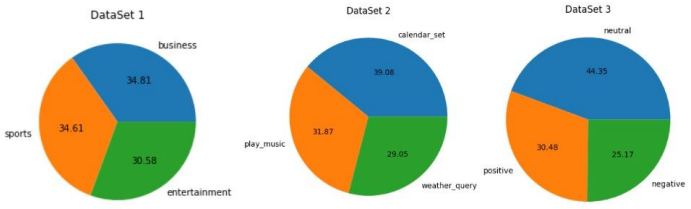

Fig.2. Distribution of data in three datasets. **Dataset 1**: Malayalam news headlines; **Dataset 2**: Amazon massive dataset; **Dataset 3**: Translated chat sentiment dataset.

3.2 PREPROCESSING

The preprocessing stage involves data cleaning, tokenization, stop word removal, and stemming. These steps are essential to prepare the raw text data for subsequent analysis, ensuring that the data is in a standardized format and that noise is minimized. In Malayalam text, preprocessing presents specific challenges due to the complex script, the presence of compound words and absence of much relevent work (no standardized set of stopwords for example). To tackle this, a manually manually curated list of Malayalam stopwords were used for stopword removal and the stemming was performed using a python library 'morph-gen' [9].

3.3 EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is performed to understand the distribution of data across different classes, word lenght and character length. Most common words in each category excluding trivial stopwords are computed and kept as lists. The words commonly appearing in all labels are excluded form all three lists (since it does not hold any significance towards any of the labels). Thus, the list corresponding to the respective category would contain words that would potentially be a decisive pivot for the models in predicting the class. Guided-LDA [10] based topic modelling is performed to understand the topic distribution of words in each dataset.

3.4 TEXT EMBEDDING

Embedding is done to represent the given text as vectors. Different word or sentence embedding techniques such as BOW, N-grams, TF-IDF, Word2vec, and BERT are applied to find and

take the one which is best for the preprocessed data. How well the embedding captures linguistic nuances and its impact on model performance is considered (BERT was found best at handling complex sentence structures and long-range dependencies common in Malayalam.)

3.5 MODEL ARCHITECTURES AND TEXT CLASSIFICATION

After applying the preprocessing techniques on the data, the performance of the data with different machine learning models and selected deep learning model architectures are compared.

The machine learning models used for comparison include logistic regression, random forest, support vector machines, bagging classifier, and gradient boosting algorithms (XGBoost tree algorithm and LightGBM tree algorithm).Each of these models have their own perks for instance logistic regression is viewed here as an interpretable baseline for comparison, random forest is robust against overfitting, XGBoost and LightGBM Captures complex patterns efficiently and optimize performance through advanced regularization techniques. Custom models are also utilized by selecting the best performing models from the previous list and applying as a stacking classifier and voting classifiers. Custom models using stacking and voting classifiers leverage the strengths of individual models for improved accuracy.

Under deep learning models, various architectures combining LSTM layers, fully connected layers, bidirectional LSTM layers, Convolution layers etc are utilized. This is done to figure out which works best with respect to Malayalam based text classification. The details of Deep-Learning architectures used are given in Table.1.

The hyperparameters and these different architectures are tuned to tweak model performance for better results.

Table.1. Deep Learning Architectures

Model	Model Architecture
CNN-LSTM	Embedding - Dropout - Conv1D - MaxPooling 1D - LSTM - Dense
Feed Forward	Embedding - flatten - Dense - Dropout - Dense
LSTM	Embedding - LSTM - Dense - Dense
BiLSTM	Embedding - bidirectional LSTM - Dense - Dense
CNN-BiLSTM	Embedding - Conv1D - MaxPooling 1D - bidirectional LSTM - Dense - Dense

3.6 MODEL EXPLAINABILITY

To explore the explainability of the text classification models, LIME and SHAP techniques were employed. LIME is used to provide insight into the criteria behind the choices or results by respective models.

- Using the explainer module available in LIME and SHAP implementations, the components of the text that contributed in the model selecting the corresponding class(keywords) for the text is identified.
- This process is repeated for all texts in the dataset and a list of most important words(keywords) for each of the class is generated.
- Finally, the results of explainer were validated using the outcomes of EDA. Thus, an explainability score is calculated for each specific model.

For a specific Machine learning/Deep learning model, the scores were achieved by:

- Getting the top words (after removing common words between classes and stop words) appearing for the given class(in initial phases).
- The keywords or important words contributing in selection of a given class for a model while prediction is also achieved using lime and SHAP.
- Cosine similarity was calculated between the sets of keywords from LIME analysis and those from EDA for each class. This metric quantified the alignment between the model's explanations and the data-driven insights from EDA, thus the name 'explainability scores' for each class.
- The obtained explainability scores for each class is then aggregated to get an overall explainability score for that model.
- These steps are performed for each of the selected models and their explainability scores are noted.

These analyses since conducted across three different datasets, each representing a distinct use case (intent classification, sentiment analysis, and news genre classification), accounts for dataset bias and generalize the findings.

4. RESULTS AND DISCUSSIONS

4.1 FINDINGS FROM EDA

We conducted the EDA with following major objectives: (i) class distribution in each dataset (ii) distribution of words in different classes (iii) average word and character length in each class (iv) top words in each class. It is clear from the Fig.2 that the class distribution of instances in all dataset are almost balanced. Then character and word level analysis of dataset to obtain their properties was performed. The distribution of average word length and character length in Malayalam news headline dataset is shown in Table 2. Similar patterns were observed for other datasets too.

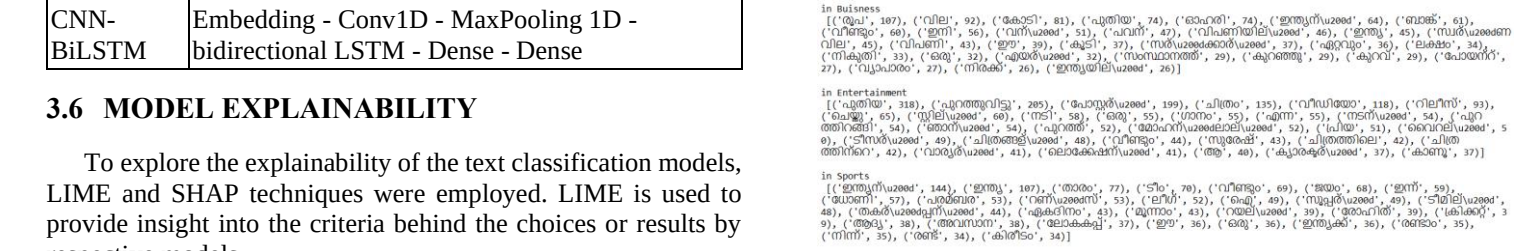

Fig.3. Sample of Most common Words for each class in Malayalam News headline dataset

The word distribution among the classes (as shown in Fig.3) gave an important insight on the feature words. The words distributed among all classes were removed from feature set, as

they carry less significance in representing a particular class. The top words identified from EDA was used later in the validation process of explainability module.

Table.2. Class wise distribution of word and character length in Malayalam news headlines dataset

Characteristic	Business		Entertainment		Sports	
	Mean	SD	Mean	SD	Mean	SD
num of characters	84.22	41.31	81.51	40.95	73.25	31.96
Word count	12.62	13.56	12.17	13.10	11.64	13.43

4.2 TOPIC MODELLING

The Latent Dirichlet Allocation (LDA) topic model was applied to understand the how the latent themes in the documents correlate with the class information. A variant of LDA called seed Guided LDA was used [10], where the keywords obtained from EDA were given as seed words for LDA. The number of topics were fixed to the number of classes in each dataset. Finally, the top words representing each topics (classes) were identified. A sample results for Malayalam news headline dataset is given in Fig.4.

```
Topic ID: 0
Words: 0.008*"ഇന്നി" + 0.006*"ഏപ്രിൽ" + 0.006*"പുതിയ" + 0.005
*"നികുതി" + 0.005*"എത്തുന്നു" + 0.004*"എ" + 0.004*"കോടി" + 0.0
04*"കാരണം" + 0.004*"ബാങ്ക്" + 0.004*"നിന്ന്"

Topic ID: 1
Words: 0.015*"ഇന്ത്യൻ" + 0.012*"ഓഹരി" + 0.008*"വീണ്ടും" + 0.007
*"റൺസ്" + 0.007*"ഇന്ത്യ" + 0.007*"പരമ്ബര" + 0.006*"ഏകദിനം" +
0.006*"വിപണി" + 0.005*"വില" + 0.005*"താരം"

Topic ID: 2
Words: 0.008*"ഇന്ത്യ" + 0.007*"കുറിച്ച്" + 0.007*"ഇന്ത്യക്ക്" + 0.006
*"കോടി" + 0.005*"ഇന്ത്യയ്ക്ക്" + 0.005*"സ്വന്തമാക്കി" + 0.005*"വില"
+ 0.004*"ധവാൻ" + 0.004*"മമ്മൂട്ടി" + 0.004*"നേട്ടത്തോടെ"

Topic ID: 3
```

Fig.4. Topic-word distribution given by LDA for Malayalam news headline dataset

Table.3. Comparison of performance of different machine learning models

Model Name	Accuracy			F1 Score			ROC-AUC		
	df1	df2	df3	df1	df2	df3	df1	df2	df3
SVC	0.86	0.92	0.81	0.89	0.92	0.83	0.95	0.97	0.92
Bernoulli NB	0.83	0.91	0.80	0.83	0.91	0.80	0.96	0.97	0.93
Multinomial NB	0.81	0.85	0.78	0.81	0.86	0.78	0.97	0.96	0.94
Logistic Regression	0.83	0.90	0.80	0.83	91	0.79	0.94	0.96	0.91
Random Forest	0.83	0.89	0.81	0.83	0.89	0.80	0.92	0.94	0.90
Bagging Classifier	0.79	0.88	0.73	0.79	0.88	0.73	0.92	0.94	0.90
ExtraTrees	0.85	0.91	0.80	0.85	0.91	0.81	0.95	0.98	0.94
Gradient Boosting	0.72	0.88	0.69	0.72	0.88	0.69	0.90	0.93	0.89
XGB Classifier	0.80	0.89	0.75	0.80	0.89	0.75	0.94	0.96	0.92
Voting Classifier	0.87	0.92	0.83	0.87	0.92	0.84	0.97	0.98	0.93
stacking	0.86	0.91	0.81	0.86	0.91	0.82	0.97	0.98	0.93

df1: Malayalam news headlines; **df2**: Amazon massive dataset; **df3**: Translated chat sentiment dataset

Table.4. Comparison of performance of different deep learning models

Model Name	Accuracy			F1 Score			ROC-AUC		
	for df1	for df2	for df3	for df1	for df2	for df3	for df1	for df2	for df3
CNN-LSTM	0.95	0.96	0.92	0.95	0.96	0.92	1.00	1.00	0.97
Feed Forward	0.93	0.96	0.93	0.93	0.96	0.93	0.99	0.99	0.98
LSTM network	0.93	0.96	0.93	0.93	0.96	0.93	0.99	1.00	0.97
BiLSTM network	0.93	0.96	0.92	0.93	0.96	0.92	0.99	1.00	0.98
CNN-BiLSTM	0.94	0.96	0.93	0.94	0.96	0.93	1.00	1.00	0.98

df1: Malayalam news headlines; **df2**: Amazon massive dataset; **df3**: Translated chat sentiment dataset

4.3 COMPARISON OF TEXT CLASSIFICATION MODELS

The classification performance of various text classification algorithms were compared based on accuracy, F1 score and area under receiver operating curve (ROC-AUC). Based on the results shown in Table 3, it could be concluded that ensemble models such as XGBoost, Voting classifier and Stacking classifer gave higher performance among machine learning models. Under deep learning models bidirectional lstms performed better than other models in the comparison, as clear from Table 4.

4.4 CALIBRATION CURVES

Calibration curves are used to check whether the prediction given by a model math with real outcomes [11]. For instance calibration curves of different machine learning models and deep learning models over Malayalam news headline dataset are given in Fig.5 and Fig.6 respectively. Upon analyzing calibration curves, the study reveals important insights. Comparing the machine learning models, some models i.e., logistic regression and multinomial NB have their calibration curves deviating significantly from the ideal $x=y$ line meaning badly calibrated. All other ml models have their calibration curves closely matching the ideal $x=y$ line, indicating their predicted probabilities are highly accurate, this also suggests they are well-calibrated with respect to the datasets used.

For deep learning models, CNN-LSTM model consistently follows the $x=y$ line across all classes across all three datasets. In contrast, other DL models used in the comparison showed varied patterns, with various models excelling in different classes. For example, in dataset 1: LSTM network is better for sports, Bi-LSTM network is better for entertainment, whereas CNN-BiLSTM is good for business. These variations highlight that while some deep learning models are well-calibrated for certain classes, their performance can vary significantly across different categories. Another important observation is that machine learning models display smoother calibration curves than deep learning models, suggesting more reliable predicted probabilities, suggesting that machine learning models might be better

calibrated overall for the datasets used. Therefore, deep learning models, although powerful, can sometimes produce less consistent probability estimates, leading to less smooth calibration curves.

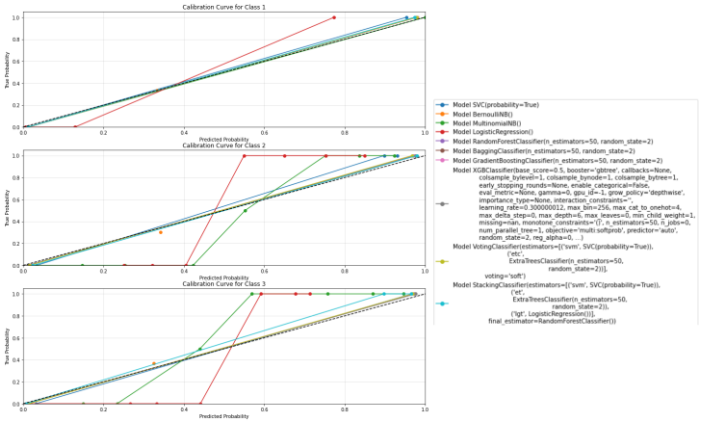

Fig.5. Calibration curve for the machine learning models for Malayalam news headline dataset

4.5 EXPLAINABILITY OF MODELS

The best models based on the performance evaluation and calibration curves were considered in the next level to compare their explainability. The comparison was done by finding the key features in the text which influences the class prediction by the model. Later these topics were passed to a method which delivers a comparative score of the text, indicating the explainability of that model. Further SHAP and LIME plots are examined in the process, to solidify the claims and to decide the explainability parameter. For instance, LIME and SHAP (waterfall plot) explainability plots for a sample text from the Malayalam news headline dataset using the Extra-tree classifier is given in the Fig.7.

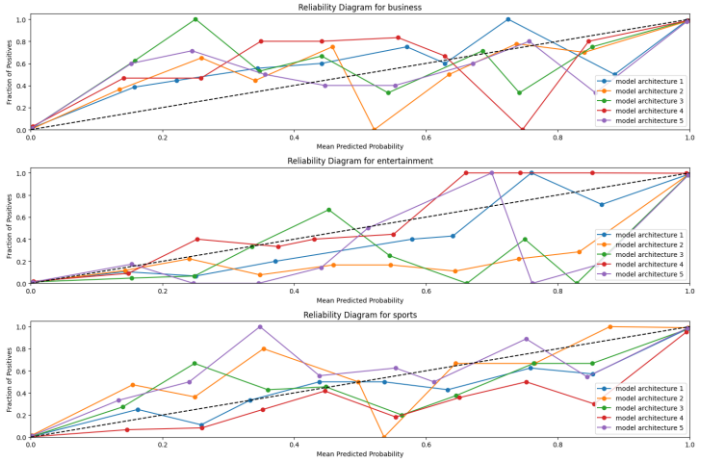

Fig.6. Calibration curve for the deep learning models for Malayalam news headline dataset

Further to compare the explainability of the models, A score calculation method was introduced in which the main topics extracted through EDA and topic modelling for each class would be compared with the main topics for the respective class each model obtained via LIME calculations. The comparative score is generated by computing the cosine similarity between word embeddings from EDA and LIME. BERT-base-multilingual-cased model was used [12] for computing the word embedding. The values of a given model for all three classes are taken and average is computed thus obtaining a score signifying the level of explainability of each of the models. The model explainability scores thus computed is given in Table 5. It can be observed from the table that the extra-tree classifier gave the highest score amongst all the models while the feed forward architecture and CNN-LSTM architecture got the least scores for explainability amongst the pack.

Fig.7. LIME plot and SHAP waterfall plot explaining the choice of the class for a sample data item in Malayalam news headline dataset

Table.5. Model Explainability Scores

Model	Dataset1	Dataset2	Dataset3	Explainability
Extra Trees Classifier	0.91	0.95	0.87	0.9164
CNN-LSTM	0.89	0.93	0.87	0.8921
Feed Forward	0.90	0.92	0.87	0.8952
LSTM	0.91	0.94	0.86	0.9067
BiLSTM	0.91	0.93	0.88	0.9072
CNN-BiLSTM	0.91	0.94	0.86	0.9048

df1: Malayalam news headlines; **df2**: Amazon massive dataset; **df3**: Translated chat sentiment dataset

5. CONCLUSION

This study explored the concept of explainability in machine learning and deep learning models for Malayalam text classification across three different use cases: intent classification, sentiment analysis, and news genre classification. Utilizing methods such as LIME and SHAP, the inner workings of model decisions were explored. The results show that while deep learning models exhibit better performance in terms of class

predictions, their interpretability is lower. Conversely, some advanced machine learning models, although they may not perform as good as some of the deep learning architectures in terms of accuracy, area under the ROC curve and other performance matrices, marginally outperformed them in terms of explainability. This suggests that a trade-off exists between model performance and interpretability across various use cases in Malayalam text classification.

REFERENCES

[1] M.T. Ribeiro, S. Singh and C. Guestrin, "why Should I Trust You?" Explaining the Predictions of any Classifier", *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135-1144, 2016.

[2] S. M. Lundberg and S.I. Lee, "A Unified Approach to Interpreting Model Predictions", *Advances in Neural Information Processing Systems*, Vol. 30, pp. 1-12, 2017.

[3] F. Doshi Velez and B. Kim, "Towards a Rigorous Science of Interpretable Machine Learning", *Proceedings of IEEE International Conference on Machine and Deep Learning*, pp. 1-6, 2017.

[4] M.T. Ribeiro, S. Singh and C. Guestrin, "Anchors: High-Precision Model-Agnostic Explanations", *Proceedings of the AAAI conference on artificial intelli- gence*, vol. 32, no. 1, 2018.

[5] L. Kurasinski and R.C. Mihailescu, "Towards Machine Learning Explainability in Text Classification for Fake News Detection", *Proceedings of IEEE International Conference on Machine Learning and Applications*, pp. 775-781, 2020.

[6] C.J. Mahoney, J. Zhang, N. Huber-Fliflet, P. Gronvall and H. Zhao, "A Framework for Explainable Text Classification in Legal Document Review", *Proceedings of IEEE International Conference on Big Data*, pp. 1858-1867, 2021.

[7] W. Zhao, T. Joshi, V.N. Nair and A. Sudjianto, "Shap Values for Explaining CNN-Based Text Classification Models", *Proceedings of IEEE International Conference on Machine Learning and Applications*, pp. 344-354, 2020.

[8] A. Nayebi, S. Tipirneni, B. Foreman, C.K. Reddy and V. Subbian, "An Empirical Comparison of Explain able Artificial Intelligence Methods for Clinical Data: A Case Study on Traumatic Brain Injury", *Proceedings of IEEE International Conference on American Medical Informatics*, pp. 815-820, 2022.

[9] J. Baby, "Morphgen: Python Package for Morphological Analysis of Malayalam Text", Available at https://pypi.org/project/morph-gen/, Accessed in 2023.

[10] C. Li, S. Chen, J. Xing, A. Sun and Z. Ma, "Seed-Guided Topic Model for Document Filtering and Classification", *ACM Transactions on Information Systems*, Vol. 37, No. 1, pp. 1-37, 2018.

[11] A. Martino, E. De Santis, L. Baldini and A. Rizzi, "Calibration Techniques for Binary Classification Problems: A Comparative Analysis", *Proceedings of International Joint Conference on Computational Intelligence*, pp. 487-485, 2019.

[12] J. Devlin, M. Chang, K. Lee and K. Toutanova, "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding", *Proceedings of International Joint Conference on Computational Language*, pp. 1-12, 2018.