Journal of Computer Science

Research Article

Enhancing Sentiment Analysis for Malayalam With mBERT: A Profoundly Transparent and Accurate Approach Using LIME

Anitha R.[1], K. S. Anil Kumar[2], Rajeev R. R.[3], Ansil Shafee[1], Manju G.[4] and Reshmi L. B.[1]

[1]Department of Futures Studies, University of Kerala, Thiruvananthapuram, India
[2]KSMDB College, Sasthamcotta, KOLLAM, India
[3]ICFOSS, Thiruvananthapuram, India
[4]Department of Computer Science, P. M. Govt. College, Chalakudy, India

Article history
Received: 08-07-2025
Revised: 21-01-2026
Accepted: 23-03-2026

Corresponding Author:
Anitha R.
Department of Futures Studies,
University of Kerala,
Thiruvananthapuram, India
Email:
anithathilak_2025@keralauniversity.ac.in

Abstract: The overlapping sentiment boundaries, intensifiers, and intricate morphological structures in Malayalam present particular difficulties for sentiment analysis, making it hard for traditional machine learning techniques to produce consistent results. We present an explainable sentiment analysis framework in this paper that refines a Multilingual Bidirectional Encoder Representations from Transformers (mBERT) model on a novel constituency-level dataset that has been manually curated and annotated into five-class (very positive, positive, neutral, negative, and very negative) and three-class (positive, neutral, and negative) categories. In contrast to previous research that focuses solely on accuracy, our method incorporates Local Interpretable Model-Agnostic Explanations (LIME) to identify linguistic cues that significantly impact sentiment prediction in Malayalam, including intensifiers, negations, and context-dependent modifiers. Despite the inherent linguistic complexity, the suggested model demonstrated consistency, achieving 61.78% precision for three-class classification and 61.47% for five-class classification. More significantly, the LIME-based interpretability analysis provides a clear and linguistically grounded standard for low-resource sentiment analysis by highlighting the impact of Malayalam-specific features on classification results. In addition to presenting one of the earliest explainable BERT-based sentiment models for Malayalam, this work lays the groundwork for further studies on interpretable deep learning in underrepresented languages. As far as we know, the current work is the first to create an explainable, transformer-based sentiment analysis framework for Malayalam that incorporates BERT with LIME and is underpinned by a constituency-level curated dataset. This contribution sets a new standard for NLP in low-resource languages in terms of performance and explainability.

Keywords: Sentiment Analysis, Malayalam, BERT, Explainable AI, LIME, NLP

Introduction

Sentiment analysis is a branch of Natural Language Processing (NLP) that looks for specific sentiments in texts, such as positive, negative, or neutral. Many applications, such as market research, social media monitoring, and customer feedback analysis, depend on this activity. Despite notable breakthroughs in sentiment analysis for widely spoken languages like English, research on low-resource languages like Malayalam is scarce. To ensure inclusive technological advancements and accommodate diverse linguistic groups, sentiment analysis for various languages must be addressed.

In the past few decades, NLP has seen a considerable evolution, moving from rule based approaches to machine learning and, more recently, deep learning techniques. The

introduction of transformer models, which have raised the bar for several NLP tasks (Devlin et al., 2019), including sentiment analysis, is one of the most significant advances in the field. The Bidirectional Encoder Representations from Transformers (BERT) is one of these transformer models that has gained popularity because of its strong performance and adaptability.

BERT is a deep learning model that looks at both the left and right sides of words to comprehend their context in a sentence in both directions (Anitha et al., 2023). Due to its bidirectional approach, BERT is particularly useful for applications such as sentiment analysis, as it can capture complex relationships and dependencies within text. BERT (Cañete et al., 2020) is suitable for low-resource languages like Malayalam, as it has been pre-trained on a substantial corpus of text and can be finetuned for specific tasks with relatively minimal additional training data.

The decision-making process of BERT is sometimes opaque despite its well-established, formidable capabilities due to its complicated design. Comprehending the rationale behind a model's predictions is crucial, particularly in scenarios where trust and accountability are significant. Local Interpretable Model-agnostic Explanations (LIME) (Chandrakala and Sindhu, 2012) are useful in this situation. LIME is a method that uses a locally approximable interpretable model to explain the predictions of any classifier. LIME improves transparency and trust by providing an explanation for each prediction, assisting in the discovery of the underlying characteristics that influence the model's judgments.

In this work, we introduce a new method of sentiment analysis for the Malayalam language, which combines the BERT model with LIME. Our method makes use of BERT's strong multilingual sentiment classification skills as well as LIME's interpretability to offer insights into the model's predictions (Ribeiro et al., 2016). This work's main contributions are:

- Creation of a sentiment analysis model using BERT for the Malayalam language
- Interpretability of the model is improved by LIME
- A comprehensive evaluation of the model's performance using a sentiment dataset selected from Malayalam

Our test results show that the suggested method works well; we were able to obtain an impressive accuracy of 57% on the test dataset. Furthermore, the LIME explanations provide insightful information about the behaviour of the model, emphasising the significance of interpretability in contemporary NLP applications.

Research Gap

Recent deep learning studies focus mostly on accuracy, neglecting interpretability. Most existing research on Malayalam sentiment analysis depends on lexicon-based or superficial machine learning methods. Additionally, most datasets lack constituency-level annotation or representation of language variety due to their origins in noisy social media content. To our knowledge, no prior research has incorporated explainability for Malayalam into transformer-based models. This study addresses the gap by providing a manually curated constituency-level dataset and proposing an explainable BERTLIME methodology tailored to the morphological richness and sentiment complexity of the language. Other studies relied on lexicon-based approaches or unexplainable transformer models. In contrast, we combine the best performing deep transformers with linguistically motivated rationales. Such engineering innovations make our system the first explainable transformer-based sentiment analyzer for Malayalam.

Related Work

Opinion mining, also known as sentiment analysis, has been thoroughly investigated across a wide range of languages and fields (Vaswani et al., 2017). In earlier studies, the sentiment was assessed using lexicon-based techniques, which employed predetermined dictionaries of positive and negative words defined in Pang et al. (2002). Later, in Manning et al. (2008), machine learning methods for sentiment classification tasks gained popularity, including support vector machines and naive Bayes classifiers. Due to their capacity to identify intricate patterns in text, deep learning models, in particular, Convolutional Neural Networks (CNNs) and recurrent neural networks (RNNs), have demonstrated notable advancements in sentiment analysis performance in more recent times explained in Kim (2014). Sentiment analysis studies in low-resource languages such as Malayalam is somewhat rare (Conneau et al., 2020). Early approaches combined handmade features with rule-based and machine-learning techniques. Ranjan et al. (2016) studied applications of deep learning models, such as CNNs and RNNs, to increase accuracy. The absence of substantial annotated datasets is still a major obstacle, though. Sentiment analysis for Malayalam is being improved by the use of transfer learning algorithms and the curation of extensive datasets, says Vasudevan et al. (2020).

In Xu et al. (2018), extensive research has been conducted on sentiment analysis in languages other than English. Character-level embeddings and attention mechanisms, for instance, have led to notable advances in Chinese sentiment analysis. According to Al Sallab et al. (2015), morphological richness in Arabic presents particular difficulties that can be solved with the use of deep learning models and specialized preprocessing methods. Cross-lingual transfer learning, which uses models pre-trained on English corpora, has helped studies

in Spanish and Portuguese (Radford et al., 2018). These initiatives demonstrate the variety of issues and approaches in sentiment analysis across linguistic domains. Mikolov et al. (2013) defined NLP as covering a broad range of problems, such as machine translation, text categorization, and language modeling. Deep learning has completely changed the area; previous methods mostly relied on feature engineering and statistical models. The creation of word embeddings that capture the semantic links between words, such as Word2Vec and GloVe (Pennington et al., 2014), is one notable milestone. In Vaswani et al. (2017), NLP was further developed with the addition of attention mechanisms and transformer models, which made it possible to build models that are particularly good at comprehending dependencies and context in text.

By bringing a bidirectional approach to language modeling, BERT (Bidirectional Encoder Representations from Transformers) marks a major advancement in NLP, (Devlin et al., 2019). BERT takes into account the context from both directions, in contrast to earlier models that processed text in a unidirectional manner, enabling a more thorough comprehension of the meaning of words in a phrase. BERT can be optimized for certain tasks and achieves state-of-the-art performance in several benchmarks because it has been pre-trained on large corpora. Yang et al. (2019) defined other transformer-based models, such as XL Net and RoBERTa (Liu et al., 2019), which have been developed due to their success. (Ribeiro et al., 2016) The problem of interpretability in complex machine learning models is addressed by Local Interpretable Model-agnostic Explanations (LIME). LIME produces locally faithful explanations by approximating the original model with an interpretable one centered around the relevant prediction. This method is beneficial in understanding how “black-box” models, such as deep neural networks, behave, as it reveals which features are most important for a specific prediction. LIME has been widely used in various applications and provides valuable insights into how decision-making processes are modeled.

Research for low-resource languages continues to highlight the problems of the lack of data as well as data-efficient modeling frameworks. For instance, Gokani and Mamidi (2023) developed the GSAC, comprising Twitter data, manually collected and annotated, and developed initial benchmark models like SVM, logistic regression, and ensemble methods to aid subsequent work. Also, Aliyu et al. (2024) reviewed work on low-resource sentiment analysis from 2018 to 2023 and noted the dominance of transfer learning and BERT/XLM-R-based approaches, as well as the stagnation of older methods (lexicon, classifier machine learning) due to their low applicability. Both works add to the evidence showing the lack of large-scale annotated resources for Indian languages and the consequent inability to build sentiment analysis tools.

There are also transformer-based solutions that have demonstrated strong potential in closing cross-lingual gaps. Kumar and Albuquerque (2021) showed that XLM-R predicts English sentiment and, with zero shot transfer learning, projects it onto Hindi datasets with greater than 60% accuracy, exceeding the performance of numerous baseline models. Outside of Indian languages, Raychawdhary et al. (2025) extended sentiment analysis in 12 multilingual low-resource African languages using XLM-R, AfroXLMR, AfriBERTa, mDeBERTaV3, and other fine-tuned multilingual transformer models, achieving the best results in AfriSenti SemEval 2023 shared task AfriXLMR 75.8% weighted F1 score with the rest of the competition built from fundamental techniques like CNN, Naïve Bayes and SVM. All these works demonstrate the value of multilingual transformer-based models, dataset synthesis, and their combination as critical driving forces in performing sentiment analysis in low-resource languages.

Methods

Dataset

Public opinion on some selected political and social issues relevant to the people of Kerala was sought through two polls. The study employs both qualitative and quantitative approaches since the aim of the study is to capture the data from all people in as representative a manner as possible. Therefore, both direct face-to-face and indirect digital voice responses were necessary in conjunction with the self-administered surveys. Alongside these methods, several targeted open-ended questions were crafted before the survey to facilitate recognition of the different sentiments expressed in both Malayalam and English. Below is a detailed account of the survey question and data collection procedure. The results from these opinion polls are set to conduct sentiment analysis as well, and the data set is rich (Anitha et al., 2024).

Figure 1 illustrates how the sentiment analysis of the multi-class classification approach operates on a dataset constructed from a questionnaire. For starters, the dataset is categorized into three key sentiment labels: Positive, negative, and neutral. The left half of the figure showcases some examples of sentences written in Malayalam and their corresponding labels, along with the assigned values. So, for example, the “negative” label is assigned to sentences that express dissatisfaction or unfavorable attitude, while opinions that are complimentary or express any satisfaction are written as “positive”. Neutral sentiments do not invoke either strong positive or emphatic negative emotions.

തുടക്കത്തിലുള്ള ആവേശവും ജാഗ്രതയും രോഗവ്യാപനം വർധിച്ചു നിൽക്കുന്ന ഈ സാഹചര്യത്തിൽ കാണാൻ സാധിക്കുന്നില്ല	negative
മുന്നാക്ക വിഭാഗക്കാരുടെ അവസരങ്ങൾ മുന്നാക്ക വിഭാഗത്തിൽത്തന്നെയുള്ള തെ കുട്ടം ആളുകൾ തന്നെയാണ് ഇല്ലാതാക്കുന്നത്.	negative
ഇതിലൂടെ സംവരണം എന്നതിന്റെ പ്രസക്തി തന്നെ ഇല്ലാതാകുന്നു	negative
വളരെ മികച്ചത് തന്നെയാണ്. പ്രതിസന്ധി ഘട്ടങ്ങളിലും പ്രതിപക്ഷ ഉത്തരവാദിത്തങ്ങൾ മറക്കാതെയുള്ള ശക്തവും മാന്യവുമായ പ്രവർത്തനം കാഴ്ചവെക്കാൻ പ്രതിപക്ഷത്തിന് കഴിഞ്ഞു	positive
കള്ളപ്പണത്തിനെതിരായ കത്തക തകർക്കാനും ജനങ്ങൾക്കിടയിൽ ഉണ്ടായ സാമ്പത്തികമായ ബുദ്ധിമുട്ട് മാറ്റാനും ഇത് കഴിഞ്ഞു	neutral
യഥാർത്ഥ വർത്തമാനത്തേക്കാൾ കാൾ പതിന്മടങ്ങ് കൂടുതലാണ് ആണ് ജനങ്ങളുടെ വർത്തമാനം.	neutral
വളരെ നല്ലതാണ് നമ്മുടെ അടിസ്ഥാന വികസനങ്ങൾ എല്ലാം	very positive
ആരോപണങ്ങളുടെ ഉന്നയിപ്പ കത്ത് കയ്യിൽ ഒതപാട് പേർക്ക് സപ്പം സാക്ഷാത്കരിക്കാതെ വന്നു	very negative
പൊതുജനങ്ങൾക്ക് വേണ്ട കാര്യങ്ങൾ ഏറെക്കുറെ ഇ കാര്യത്തിൽ നടത്തുന്നുണ്ട്.	positive
കേന്ദ്ര സർക്കാർ നടപ്പിലാക്കുന്ന ഈ ഒരു പദ്ധതിയുടെ എനിക്ക് ഒരിക്കലും സാധ്യമല്ല	negative

Fig. 1: Sample dataset for class 3 and class 5 Labelled

Further extending this part of the work, now the dataset is also more detailed and divided into five classes according to the sentiments expressed, such as very positive, positive, neutral, negative, and very negative. This is what is depicted in the right half of the figure. Initial examples of extreme positivity are captured in the form of very positive statements, while extreme negativity is represented by very negative ones. This additional refinement makes it easier to comprehend the contents of the text. For example, whether an opinion is used to describe something in a strongly favorable or enthusiastic focus will affect how positively or negatively it is used. This highly focused method of annotation will yield a better sentiment analysis of the dataset, more so for cases with subtle differences in sentiments. And indeed, shifting from three classes to five classes demonstrates an expansion of sentiment representation that will benefit the machine learning models' development and linguistic research in Malayalam. This ensures the dataset will satisfy both the basic and advanced requirements for sentiment analysis.

To develop thoughtful editorial materials for this study, two public opinion surveys on relevant social and political topics were carried out at the constituency level throughout the entire state of Kerala. The sample collection employed both in-person interviews and online survey instruments, and aimed at obtaining responses from all age groups, all geographic areas, and various strata of the social hierarchy. The surveys were designed to get responses to attitudinal, perceptual, and emotive questions that were posed in English and Malayalam. Lower case text, case folding, selective stop-word removal phrases for the contextually relevant, and the normalization processes of punctuation, non-Malayalam markings, websites, and other forms of extraneous space were used to preprocess the data. These automated processes preserved important linguistic elements.

The social survey responses captured from various social media sites were used to create a five-class sentiment analysis dataset comprising 22,756 Malayalam text samples. Each text sample were classified into five distinct sentiment categories to fine tune the Malayalan sentiment analysis, going beyond the traditional three tier sentiment scale. Each text sample was verified and manually classified into one of five categories: Very positive, positive, neutral, negative, or very negative. The dataset was divided to facilitate trustworthy assessment and versatile application.

The dataset was divided into three categories: 15,930 instances (70%) for training, 3,413 cases (15%) for validation, and 3,413 instances (15%) for testing. This separation maintained class balance in subsets, which minimizes overfitting while maximizing generalization. The last dataset not only provides the quantity of data necessary for deep learning-based sentiment classification, but also exhibits ample political discourse along with the linguistic diversity of Malayalam. The dataset used in this study comprises a total of 22,756 Malayalam text samples collected through constituency-level public opinion surveys and selected social media sources. Data were gathered using both faceto-face interviews and online survey instruments, including manually transcribed voice responses to ensure linguistic accuracy. Each sample was carefully verified and manually annotated into five sentiment categories: Very Positive, Positive, Neutral, Negative, and Very Negative. The class distribution includes 7,876 Negative, 7,818 Positive, 2,354 Neutral, 2,363 Very Negative, and 2,345 Very Positive instances, ensuring sufficient representation across all sentiment intensities. To maintain data integrity, a stratified split of 70% for training, 15% for validation, and 15% for testing was applied. This structured dataset enables effective learning of both coarse and fine-grained sentiment variations, supporting robust model evaluation for low-resource Malayalam sentiment analysis.

Data Preprocessing and Tokenization

All raw Malayalam text was prepared uniformly as to be ready for the model training while preserving linguistic accuracy. The following was achieved.

Normalization of the Malayalam script: Unicode normalization to eliminate differences in Malayalam characters and inconsistencies from various input channels (social media channels, surveys transcripts) from different Malayalam character encodings:

- Case Folding: Given that the Malayalam script does not have an upper and lower case distinction, the representation was uniformly done for the sake of consistency
- Stop-word Removal: A selective stop word list was made in an attempt to eliminate high frequency functional particles that do not have any sentiment polarity and sentiment polarity sensitive marker that should not be removed as it is context sensitive such as intensifier and negation
- Nuisance Removal: All punctuation marks, URLs, emojis, symbols other than Malayalam, and unnecessary spaces were eliminated. Code mix (Malayalam-English) text were normalized while sentiment bearing words and transliterated words were retained
- Tokenization: Separate pieces of text were grouped and converted to tokens by the HuggingFace Multilingual BERT WordPiece tokenizer, which is especially designed for more morphologically rich and low resource languages. Additionally, preliminary experiments had demonstrated that this particular tokenizer captures essential subword units (prefixes, suffixes, and inflections) to the Malayalam Language

This particular preprocessing pipeline made sure that sentiment bearing linguistic cues such as the markers of negation (“ഇല്ല”), context intensifiers (“വളരെ”), and context modifiers were preserved to permit classified as the sentiment cues.

Model Architecture

Figure 2 shows a detailed architecture of the sentiment analysis process in the Malayalam language from the collection, preprocessing, and the sophisticated techniques like BERT and LIME algorithms. Below you will find an overview of how everything works concerning Class 3 (positive, negative, neutral) sentiments classification and Class 5 (very positive, positive, neutral, negative, very negative). The process begins with the data collection, in this case, through surveys, which are referred to as data collection. Respondents are permitted to provide raw text data, which will be utilized later. All collected data from both surveys is compiled into a single set of information ready for analysis.

Data cleaning included the cleaning of noise resulting from unimportant data, specialized characters, or stop words while delivering a quality dataset. This preprocessing step is essential for precisely classifying the sentiment in their context, as it allows the model to discover valuable linguistic patterns. Once the data has been preprocessed it is now ready for transformation into an input type that is appropriate for the sentiment analysis model. The stage of the procedure referred to as” Input Embedding” corresponds to the transformation of each word or token into a dense vector representation within the higher-dimensional space. Those embeddings are adjusted and properly trained in detail and here so that they will convey the meaning and semantics of the Malayalam language written down.

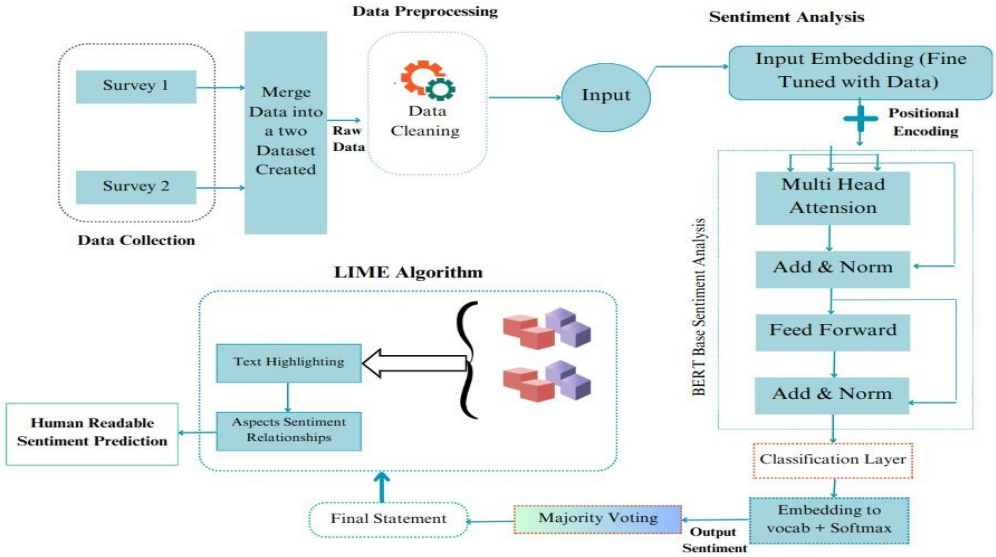

Fig. 2: Sample dataset for class 3 and class 5 Labelled

Positional Encoding is used at this point to help remember the order of the words that is needed for understanding the context of sentiment sentences. This prepares the data to be processed by the BERT-based software architecture at a later time. The heart of the framework is BERT, which stands for Bidirectional Encoder Representations from Transformers, deeply deep-embedding model that does the sentiment classification. The input embeddings are fed to several 'Multi Head Attention' layers, through which the model can observe and pay attention to different parts of the sequence at the same time. After this step, 'Add & Norm' layers are applied in order to ensure, that the 'Feed Forward' transformation network. These components are acquired several times in order to obtain a deep contextual understanding. Finally, the output goes to a 'Classification Layer', which will classify the output to the appropriate class as, a label with a specific sentiment. For Class 3 sentiment outputs, it would be either positive or negative or neutral, while for Class 5 it has additional categories of positive and very negative.

Training Procedure

Fine-Tuned BERT Model

In order to adjust the BERT model for sentiment classification, users are required to first convert the emotion labels into integers. After that, accounts can be created for specific users and the dataset can be partitioned into two groups, a training set, and a validation set, while setting aside twenty percent for validation. Each sentence in the dataset is processed by being split into tokens, and then either padded or trimmed to fit an optimal length. The output for a BERT model mentored with three sentiment classes is logits for each of the classes which corresponds to the three class types. The sparse categorical cross-entropy loss function is applied such that: During training, weights (θ) are adjusted iteratively through the use of gradient descent. Evaluation metrics such as accuracy, precision, recall, and F1 score are calculated in order to determine performance. Then, the fine-tuned model, along with a tokenized version, is saved for further use.

Model Training Details

The framework was developed with the use of the Hugging Face Transformers library and used Multilingual BERT (mBERT, base-uncased) as it's the backbone model. The dataset of Malayalam was used to fine-tune the model using the configuration as follows:

* Model variant: Multilingual BERT (mBERT, 12 layers, 768 hidden units, 12 attention heads, 110M parameters)
* Input Length: Malayalam sentences were tokenized using WordPiece and padded or truncated to a maximum of 128 tokens
* Optimizer: AdamW with weight decay
* Learning rate: 2e−5 with a linear warmup and decay schedule
* Batch Size: 32 for training and 16 for evaluation
* Epochs: 3 epochs with early stopping based on validation loss
* Dropout: 0.1 on fully connected layers
* Loss Function: Cross-entropy loss for 3-class and 5-class classifications
* Classification Head: A layer of the fully connected softmax 3 output nodes (positive, neutral, negative) for 3-class and 5 output nodes (very positive, positive, neutral, negative, very negative) for 5-class classification
* Data Split: The training-validation-test split was 70-15-15, keeping the class distribution stratified

This configuration was chosen based on prior recommendations on fine-tuning transformers (Devlin et al., 2019) and preliminary experiments designed to avoid overfitting.

Algorithm for mBERT Model

Label Encoding: (1)

$y_i = f(x_i)$, where f maps:
positive → 0, negative 1, neutral→ 2)

Data Splitting: (2)

$D = D_{train} \cup D_{val}$,
$|D_{val}| = 0.2 \times |D|$

Loss Function: (3)

$$L = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(z_{i,yi})}{\sum_{k=1}^{3} \exp(z_i k)}$$

Explainability With LIME

The LIME (Local Interpretable Model-agnostic Explanations) algorithm plays an essential role in explaining difficult machine learning algorithms, such as those used for determining the sentiment of Malayalam texts (Nazeem et al., 2024). LIME, short for Local Interpretable Model-Agnostic Explanations, works by inferring explanations for model predictions using simple models that are locally interpretable around the prediction. In the case of BERT-based sentiment classification in Malayalam text, LIME helps determine the phrases and/or words that matter the most in the sentiment classification. Below is an exposition of LIME algorithms within the context of sentiment analysis.

LIME Algorithm

The following outlines the mathematical steps of the LIME (Local Interpretable Model-agnostic Explanations) algorithm for sentiment analysis in Malayalam text classification:

Problem Setup

Let:

- x: Original Malayalam text input (e.g., a sentence or document)
- $f(x)$: Prediction function of the complex model (e.g., BERT) that outputs a sentiment probability distribution
- z: Perturbed versions of x, represented as simplified inputs (e.g., binary presence/absence of words)
- $g(z)$: Simple interpretable model (e.g., linear regression) that approximates $f(x)$ locally

Generate Perturbations

Create perturbed versions of x by randomly removing or altering words:

- Represent the text x as a vector x' of size d, where d is the number of words in x
- $x'_i = 1$ if the i-th word is present in the perturbed version, and $x'_i = 0$ if the word is removed

For example, if x = "വളെര നല്ല ഭരണമാണ് കാ െവ ന്തത് ", a perturbed version x' could be:

$$x' = [1,0,1,1,1] \quad (6)$$

This means the second word ("ഭരണമാണ് ") is removed.

Obtain Model Predictions

For each perturbed version z, compute the prediction of the complex model $f(z)$.

$f(z) \rightarrow$ Sentiment probabilities for each class:

(e.g., positive, negative, neutral). (7)

Define Locality

Measure the similarity between the original input x and each perturbed version z using a similarity kernel $\pi_x(z)$. A common choice is an exponential kernel:

$$\pi_x(z) = \exp\left(-\frac{dist(x,z)^2}{\sigma^2}\right) \quad (8)$$

Where:

- dist(x,z): Distance between x and z (e.g., cosine similarity or Euclidean distance)
- σ: Bandwidth parameter controlling the locality

Train the Simple Model

Fit a simple interpretable model $g(z)$ (e.g., linear regression) to approximate $f(z)$ locally:

$$\min_g \sum_{z \in Z} \pi_x(z).(f(z) - g(z))^2 + \Omega(g) \quad (9)$$

Where:

- Z: Set of perturbed samples
- $\Omega(g)$: Regularization term to ensure g remains interpretable (e.g., sparsity constraint to limit the number of features)

Extract Feature Importance

The coefficients w_i of the simple model $g(z)$ represent the importance of each word i in x for the sentiment prediction:

$$g(z) = w_1 z_2 + ... + w_d z_d + b \quad (10)$$

Where:

- w_i: Weight of the i-th word
- b: Bias term

Output Explanation

- Words with higher w_i values (positive or negative) are identified as the most influential for the sentiment prediction
- For example, if $w = 0.75$, the word "ഭരണമാണ് " is a strong indicator of positive sentiment in the sentence, "വളെര നല്ല ഭരണമാണ് കാ െവ ന്തത് " explain in above figure 3 and 4

LIME-Based Prediction Probabilities for Five-Class Sentiment Analysis

VP- Very Positive; P- Positive; Nl- Neutral; Ne-Negative; VN- Very Negative.

Sentence 1:

വളെര മോശം. എല്ലാവരും അഴിമതിക്കാരാണ്. അപ്പോൾ മറ്റുള്ളവരെ വിമർശിക്കുന്നതിന് പരിധിയുണ്ട്.

Table 1 details the prediction probabilities based on LIME for the same sentence within the context of the five-class classification and three-class classification frameworks. In this case, the model assigns a 95% probability to Very Negative and correctly identifies the overwhelming negativity in the sentence "വളെര മോശം. എല്ലാവരും അഴിമതിക്കാരാണ്. അപ്പോൾ, മറ്റുള്ളവരെ വിമർശിക്കാൻ കുറിപ്പ്... " ("Very bad. Everyone is a corrupt. So, there is a note to criticize the rest)" in Malayalam. In the other case, in the three class arrangement the model ascribes probabilities to Negative (93%), Positive (4%), and Neutral (3%) and, again, predicts the correct class but in a more coarse manner. The example points out that in a three class system, the extreme sentiments are bundled, whereas in the five class, the extreme sentiments, particularly the extreme negative ones, are still present and differentiated.

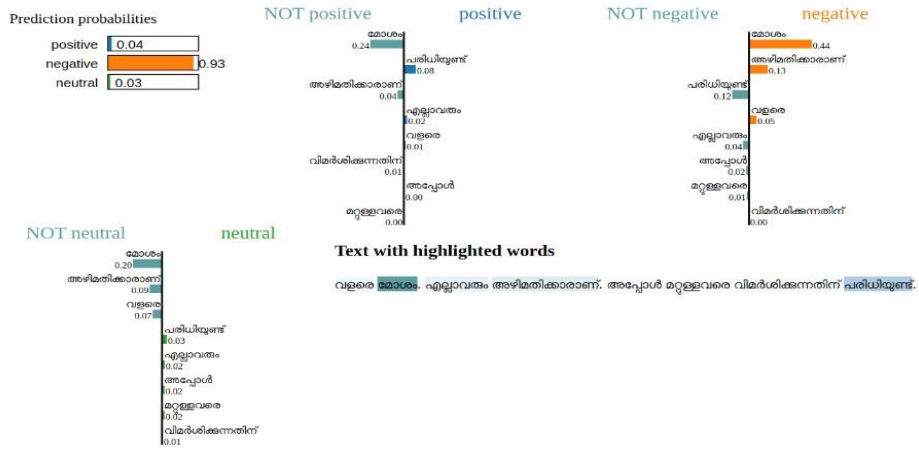

Fig. 3: LIME explanation for positive, Negative, and neutral Sentiment

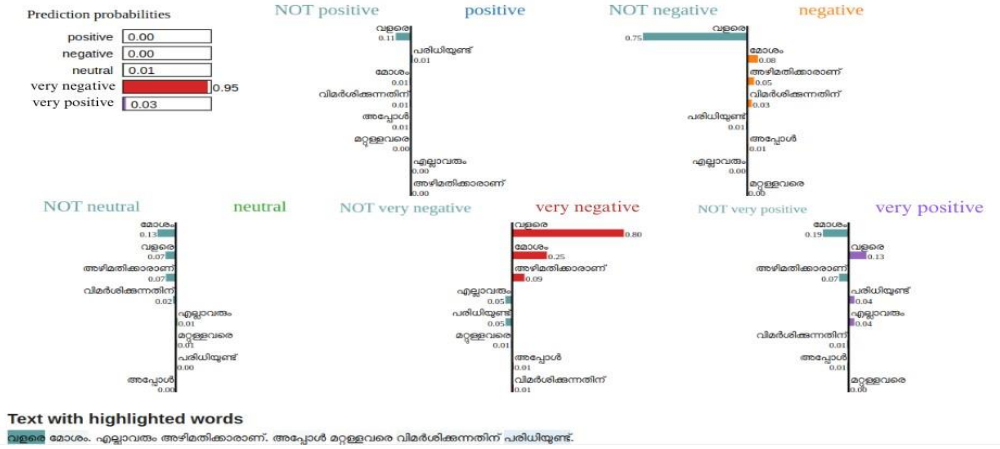

Fig. 4: LIME explanation for positive, Negative, neutral, very negative, and very positive Sentiment

Table 1: LIME-based prediction probabilities for five-class sentiment analysis

Malayalam	V P (%)	P (%)	Nl (%)	Ne (%)	V N (%)	Predicted Class
Sentence 1 (5 Class)	0	0	1	0	95	Very Negative
Sentence 1 (3 Class)	0	4	3	93	0	Negative

Results and Discussion

This research applies a modified multilingual BERT model to classify sentiments in Malayalam, a richly grammatical but resource-poor language. The model was both trained and tested on sentiment analyses in three classes (positive, neutral, negative) and five classes (very positive, positive, neutral, negative, very negative) to enable a benchmark on sentiment classification performance in various granularities. A notable innovation in this implementation is the application of a dual-mode classification system, which highlights the complexity introduced by overlapping emotional boundaries where dense sentiment spans regions of overlapping feelings. The base BERT model achieved 61.78% accuracy in a three-class scenario and performed reasonably in the five-class setting with 61.47% accuracy and 65.76% F1 score. Interpretability at the token level was done with the LIME framework, which, in the context of strongly negative sentiments, stressed the context-sensitive importance of terms like ‘theft’. These observations suggest that although lacking explicit structural awareness, BERT can provide useful sentiment information. The LIME analyses did show that neutral words occurring in highly emotional contexts can lead to erroneous classifications, particularly in the five class predictions. The training and validation curves confirmed the model’s convergence with reporting loss, accuracy, recall, and F1 score steadily improving. Ultimately, the results show that BERT, a model that can serve as a strong baseline for sentiment analysis in regional languages, benefits from fine-tuning and can be interpretable enough for model behavior scrutiny in low-resource environments.

Classifying Malayalam text is not an easy feat, and as shown in the three-class sentiment analysis 3, the model does a decent job at Positive, Negative, and Neutral classification. Prediction probabilities are indicated on the bar chart, where the Negative class seems to have a preponderance of 93% confidence, compared to the Predictive 4% and Neutral 3% scores. The LIME visualization explains some of the reasons behind the predictions made by the model, for instance, the word 'theft' is strongly negatively indicative, which increases the overall negative score.

Figure 5 and 6 illustrate the model's ability to assign values to words that are negatively weighted, albeit being strained by neutral words used in the context that result in occasional misclassification. For the five-class sentiment analysis, the model attempts to reflect the inclusion of Very Positive and Very Negative classes.

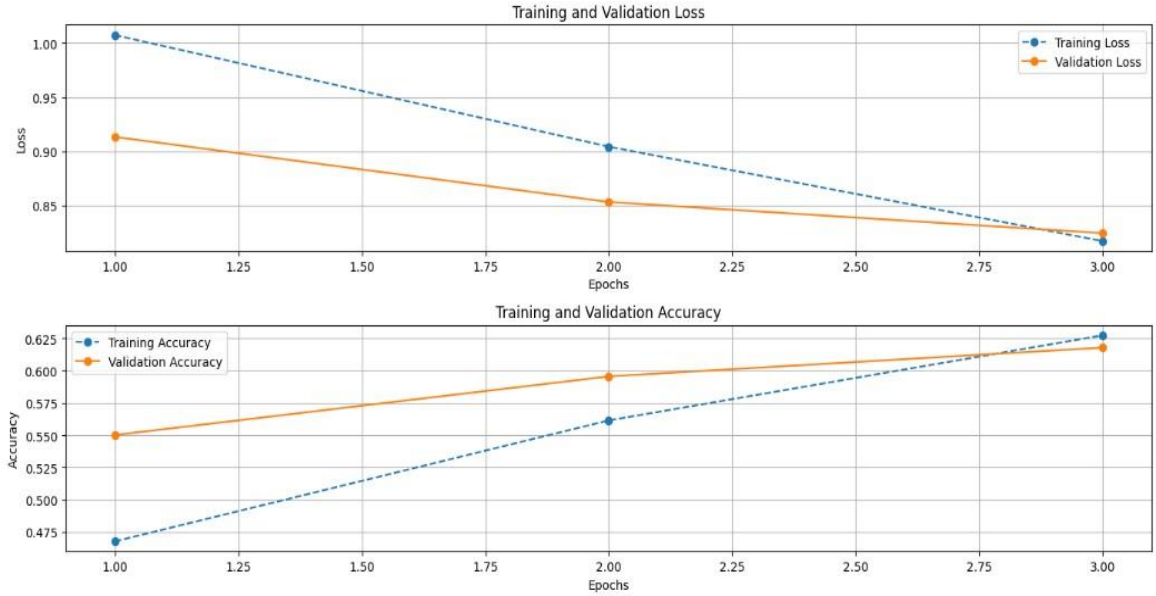

Fig. 5: Evaluations of Performance for 3-class sentiment in Malayalam throughout Training Epochs

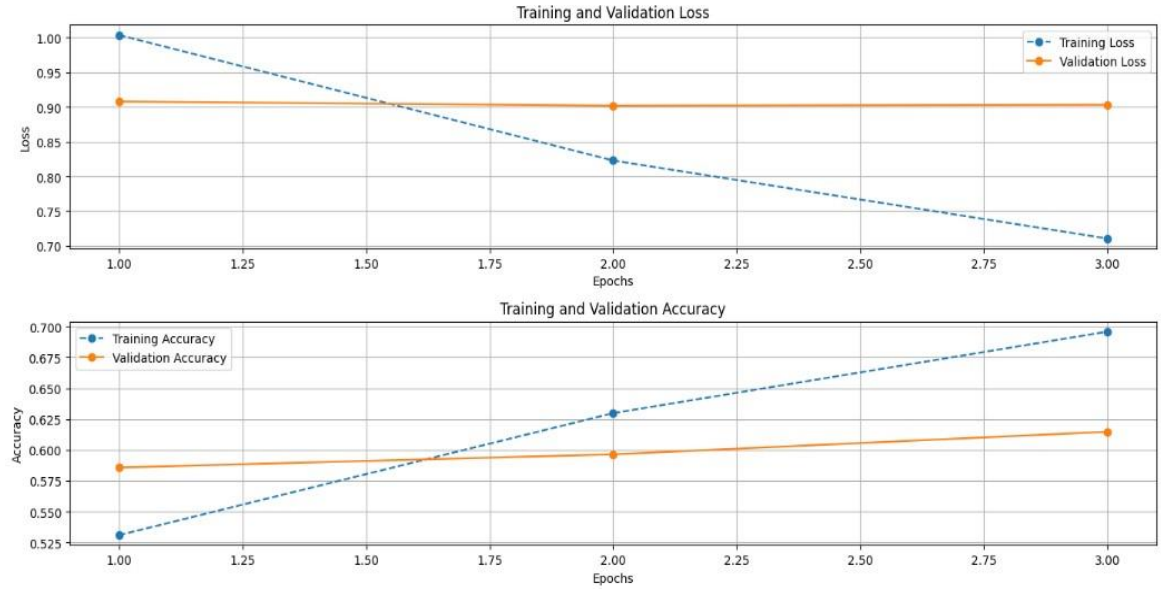

Fig. 6: Evaluations of Performance for 5-class sentiment in Malayalam throughout Training Epochs

The figure shows the highest probabilities of predictions with a strong base of 95% for the Very Negative class, with other classes such as Positive, Neutral, and Very Positive having scores are 0%, 1%, and

3%, respectively. The LIME analysis also supports these words such as "theft" which strongly assist in the identification of the Very Negative sentiment. This figure depicts how the model is able to make extreme sentiment predictions; however, it also shows the danger of adding five classes in a single prediction, where soft boundaries lead to overlapping probabilities. Figures 5 and 6 illustrate how sentiment analysis models are evaluated for a 3-class and 5-class classification problem, respectively, and how training and validation perform over several epochs. The first set of figures illustrates a possible convergence, as the loss decreases for the training data and stabilizes for the validation data. Similarly, the accuracy graphs show that training accuracy is rising but validation accuracy is much slower to rise.

This work makes a big difference by adding explainability to Malayalam sentiment analysis, which has primarily focused on accuracy without looking at how models make conclusions. The use of LIME shows that linguistic cues like negation markers and intensifiers (e.g., "വളരെ" (very), "അതിപ്രധാനമായ" (very important) can have a big effect on sentiment classification and often change predictions between classes that are next to each other. These results demonstrate that our approach is among the initial explainable sentiment analysis methods for Malayalam, since it not only classifies attitudes but also elucidates the rationale underlying predictions. This interpretability factor is critical for languages with limited resources, since trust in NLP systems relies on both transparency and correctness.

Hyperparameter Tuning

Hyperparameter tuning was performed using grid search on the validation set. Multiple configurations of batch size, learning rate, dropout, and optimizer were evaluated. The optimal model configuration was selected based on validation accuracy and F1-score. The final tuned parameters are reported in Table 2.

Past studies have examined mBERT's effectiveness on Malayalam sentiment analysis tasks (Table 3). For the Dravidian-MTL benchmark on monolingual sentiment analysis, the reported weighted F1-score was roughly 59%. In contrast, analyses on code-mixed Malayalam–English datasets have performed better, attaining around 80% weighted F1-scores (Chakravarthi et al., 2022). It is important to note, however, that these results cannot be directly compared to the monolingual configurations because the datasets have varying characteristics. When tested on a monolingual Malayalam sentiment analysis dataset, the BERT base model achieved 61.47% accuracy. This was an improvement on the typical mBERT baseline that had been reported on monolingual sentiment analysis for Malayalam, thus indicating the usefulness of the approach to improving sentiment analysis in lower-resourced languages. This improvement highlights how little potential there is to advance sentiment analysis in lower-resourced languages.

Table 2: Hyperparameter Tuning Configuration for mBERT Model

Hyperparameter	Values Tested	Final Selected Value
Model Variant	mBERT, XLM-RoBERTa	mBERT (base)
Number of Layers	12, 24	12
Hidden Units	768, 1024	768
Attention Heads	8, 12	12
Maximum Sequence Length	64, 128, 256	128
Batch Size	16, 32, 64	32
Learning Rate	1e-5, 2e-5, 3e-5	2e-5
Optimizer	Adam, AdamW, SGD	AdamW
Dropout Rate	0.1, 0.2, 0.3	0.1
Epochs	3, 5, 10	3
Weight Decay	0.01, 0.001	0.01
Warm-up Steps	0, 500, 1000	500
Loss Function	Cross-entropy, Focal loss	Cross-entropy
Classification Head	Softmax, Sigmoid	Softmax
Early Stopping Patience	2, 3, 5	3

Table 3: Performance of BERT on Malayalam Sentiment Analysis

Model / Setup	Dataset Type	Reported Score
mBERT baseline (Dravidian-MTL, 2022)	Monolingual Malay	~59.80% F% F11
mBERT (Code-Mixed Malayalam-English, 2021)	Code-mixed (DravidianCodeMix)	~61.47% Accuracy
mBERT base (2025, our work)	Monolingual Malay	

Table 4: Training and Validation Performance for 3-Class and 5-Class Classification

Class Type	Epoch	Train Loss	Val Loss	Train Acc (%)	Val Acc (%)
3-Class Sentiment	1	1.00	0.91	53.00	58.00
	2	0.82	0.90	63.00	59.00
	3	0.71	0.91	70.00	61.78
5-Class Sentiment	1	1.01	0.92	47.00	55.00
	2	0.90	0.85	56.00	59.00
	3	0.82	0.83	62.00	61.47

Table 4 presents the training and validation performance of the proposed mBERT model for both 3-class and 5-class sentiment classification tasks. For the 3-class setup, the model shows a steady decrease in training loss from 1.00 to 0.71 and an improvement in validation accuracy from 58.00% to 61.78% across epochs, indicating effective learning and convergence. Similarly, in the 5-class scenario, the training loss reduces from 1.01 to 0.82, while the validation accuracy increases from 55.00% to 61.47%, demonstrating the model's ability to handle fine-grained sentiment categories. The close alignment between training and validation performance suggests minimal overfitting and good generalization capability. Overall, the results confirm that the model performs consistently across both classification settings, with slightly higher complexity observed in the 5-class task due to finer sentiment distinctions.

Conclusion

This work addresses the challenges of interpretability and resource scarcity by presenting one of the earliest explainable transformer-based techniques for sentiment analysis in Malayalam. By improving BERT on a manually chosen constituency-level dataset and extending classification from three to five sentiment categories, we provide a reliable benchmark for analysing subtle sentiment changes in a morphologically rich language. Although the model achieves moderate accuracy (61.78% for the three-class classification and 61.47% for the five-class classification), the real innovation of the model lies in its integration of LIME explanations, which show how Malayalam-specific linguistic features such as intensifiers, negation markers, and context-sensitive modifiers influence sentiment predictions. This interpretability factor is particularly crucial for building trust in NLP systems for underrepresented languages, where black-box models often fail to meet expectations. In addition to reporting performance data, this paper demonstrates how explainable AI can enhance the transparency of deep learning models in sentiment analysis, especially when applied to languages with overlapping sentiment boundaries. The research adds three contributions:

(i) The building of a constituency-level, manually annotated dataset for sentiment analysis in the Malayalam language
(ii) The application of BERT with LIME for achieving both performance and explainability
(iii) The development of the first explainable sentiment analysis model of transformers' architecture for Malayalam, which enables transparent and linguistically responsible NLP in under resourced languages systems

Future options for enhancing classification performance include expanding the dataset to encompass more domains, incorporating additional interpretability techniques such as SHAP, and building hybrid models that explicitly model intensifiers and negations. Providing a transparent, linguistically informed, and explicable benchmark for sentiment analysis in Malayalam, this study paves the way for more reliable and trustworthy NLP applications in low-resource languages.

Acknowledgment

Thank you to the publisher for their support in the publication of this research article. We are grateful for the resources and platform provided by the publisher, which have enabled us to share our findings with a wider audience. We appreciate the efforts of the editorial team in reviewing and editing our work, and we are thankful for the opportunity to contribute to the field of research through this publication.

Funding Information

The authors have not received any financial support or funding to report.

Authors Contributions

All authors contributed equally to this study.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Al Sallab, A., Hajj, H., Badaro, G., Baly, R., El Hajj, W., & Bashir Shaban, K. (2015). Deep Learning Models for Sentiment Analysis in Arabic. *Proceedings of the Second Workshop on Arabic Natural Language Processing*, 9–17. https://doi.org/10.18653/v1/w15-3202

Aliyu, Y., Sarlan, A., Usman Danyaro, K., Rahman, A. S. B. A., & Abdullahi, M. (2024). Sentiment Analysis in Low-Resource Settings: A Comprehensive Review of Approaches, Languages, and Data Sources. *IEEE Access*, *12*, 66883–66909. https://doi.org/10.1109/access.2024.3398635

Anitha, R., Rajeev, R. R., Anil Kumar, K. S., & Nazeem, M. (2023). Comparative exploration of political opinion mining using machine learning techniques. *Journal of Information and Computational Science*, *13*(11), 74–90.

Anitha, R., Rajeev, R. R., Nazeem, M., Navaneeth, S., & Kumar, A. (2024). Comprehensive Approach to Dataset Creation for Sentiment Analysis in Malayalam. *2024 International Conference on Modeling, Simulation & Intelligent Computing (MoSICom)*, 127–132. https://doi.org/10.1109/mosicom63082.2024.10881451

Cañete, J. L. G., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., & Pérez, J. (2020). Spanish Pre-trained BERT Model and Evaluation Data. *ArXiv Preprint ArXiv:2308.02976*, 1–10. https://doi.org/10.48550/arXiv.2308.02976

Chakravarthi, B. R., Priyadharshini, R., Muralidaran, V., Jose, N., Suryawanshi, S., Sherly, E., & McCrae, J. P. (2022). DravidianCodeMix: sentiment analysis and offensive language identification dataset for Dravidian languages in code-mixed text. *Language Resources and Evaluation*, *56*(3), 765–806. https://doi.org/10.1007/s10579-022-09583-7

Chandrakala, S., & Sindhu, C. S. (2012). Opinion mining and sentiment classification: A survey. *Proceedings of the International Conference on Advances in Computing, Communications and Informatics*, 114–119.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451.https://doi.org/10.18653/v1/2020.acl-main.747

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. N. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Computer Science > Computation and Language*, *1*, 4171–4186. https://doi.org/10.48550/arXiv.1810.04805

Gokani, M., & Mamidi, R. (2023). GSAC: A Gujarati Sentiment Analysis Corpus from Twitter. In Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis (pp. 129–137). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wassa-1.12

Kim, Y. (2014). Convolutional neural networks for sentence classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1746–1751. https://doi.org/10.48550/arXiv.1810.04805

Kumar, A., & Albuquerque, V. H. C. (2021). Sentiment Analysis Using XLM-R Transformer and Zero-shot Transfer Learning on Resource-poor Indian Language. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *20*(5), 1–13. https://doi.org/10.1145/3461764

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., & Stoyanov, O. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *ArXiv Preprint ArXiv:1907.11692*, 1–20. https://doi.org/10.48550/arXiv.1907.11692

Manning, M. C., Prabhakar, R., & Hinrich, S. (2008). *Introduction to Information Retrieval*.

Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. A. (2013). Efficient estimation of word representations in vector space. *Preprint ArXiv:1301.3781.*, 1–12. https://doi.org/arXiv preprint arXiv:1301.3781

Nazeem, M., Anitha, R., & Navaneeth, S. (2024). Enhancing trust and interpretability in Malayalam sentiment analysis with explainable AI. *21st International Conference on Natural Language Processing (ICON 2024),* 1143–1150. Reference 14:

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - EMNLP '02*, 79–86. https://doi.org/10.3115/1118693.1118704

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/d14-1162

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*.

Ranjan, P., Shreeram, K., & Choudhary, P. (2016). Sentiment analysis for Indian languages (SAIL)–Code mixed tools contest: An overview. *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 93–101.

Raychawdhary, N., Bhattacharya, S., Seals, C., & Dozier, G. V. (2025). Empowering Sentiment Analysis in African Low-Resource Languages through Transformer Models and Strategic Language Selection. *IEEE Access*, *13*, 147859–147873. https://doi.org/10.1109/access.2025.3599480

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Vasudevan, A., Prabhu, S., & Damani. (2020). Leveraging transfer learning for sentiment analysis in Malayalam. *Working Notes of FIRE 2020: Forum for Information Retrieval Evaluation*, *2826*, 527–533.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 5998–6008.

Xu, H., Liu, B., Shu, L., & Yu, P. S. (2018). Double Embeddings and CNN-based Sequence Labeling for Aspect Extraction. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 592–598. https://doi.org/10.18653/v1/p18-2094

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Ruslan, S., & Le, Q. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, *32*, 5753–5763.