ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov954

Robust Multilingual Sentiment Classification for Low-Resource Telugu and Telugu-English Code-Mixed Text

Sama Sai Mukesh Reddy[1]

[1]Department of Computer Science and Engineering TKR College of Engineering and Technology Hyderabad, Telangana, India

Publication Date: 2026/11/27

Abstract: Sentiment analysis in low-resource languages presents challenges due to linguistic diversity, limited annotated resources, and variations in the way people express opinions. These challenges become more complex in Telugu-English code-mixed text, where words from two languages are used within the same sentence. This study investigates sentiment classification across Telugu and Telugu-English code-mixed text using two different approaches: Character-Level Term Frequency–Inverse Document Frequency (TF-IDF) with Logistic Regression and the multilingual transformer model XLM-RoBERTa. The experiments are conducted on two datasets, ACTSA and CMTET, covering three sentiment categories: negative, neutral, and positive. Both approaches are evaluated using accuracy, precision, recall, and macro F1-score. On the ACTSA test set, XLM-RoBERTa achieves an accuracy of 61.55% and a macro F1-score of 60.50%, compared with 58.33% accuracy and 57.97% macro F1-score for the TF-IDF baseline. On the CMTET test set, XLM-RoBERTa obtains 81.55% accuracy and 79.98% macro F1-score, while TF-IDF achieves 80.74% accuracy and 79.56% macro F1-score. To examine the reliability of these performance differences, McNemar's exact test and paired bootstrap confidence intervals are used. The statistical analysis does not establish a significant difference between the two approaches on either dataset. The results indicate that XLM-RoBERTa achieves slightly higher overall scores, while the character-level TF-IDF approach remains competitive, particularly on the code-mixed dataset. This study provides an empirical comparison of traditional character-based features and multilingual transformer representations for sentiment classification in Telugu and Telugu-English text.

Keywords: *Sentiment Analysis; Telugu Language; Low-Resource Languages; Code-Mixed Text; Natural Language Processing; Character-Level TF-IDF; Logistic Regression; XLM-RoBERTa.*

How to Cite: Sama Sai Mukesh Reddy (2026) Robust Multilingual Sentiment Classification for Low-Resource Telugu and Telugu-English Code-Mixed Text. *International Journal of Innovative Science and Research Technology,* 11(11), 1884-1900. https://doi.org/10.38124/ijisrt/26nov954

I. INTRODUCTION

➢ *Background*

Sentiment analysis is a natural language processing (NLP) task that identifies the opinions and emotions expressed in textual data. It is widely used to understand public opinions, customer feedback, product reviews, and discussions on social media platforms. Depending on the application, sentiment classification commonly categorises text into positive, negative, and neutral classes.

Although sentiment analysis has achieved considerable progress in widely studied languages such as English, developing reliable systems for low-resource languages remains challenging. Telugu, one of the major Dravidian languages, has comparatively fewer annotated datasets and language-specific computational resources. Variations in spelling, word forms, sentence structure, and informal writing make automated sentiment classification more difficult.

➢ *Motivation*

A considerable amount of online communication involving Telugu speakers includes both Telugu and English words within the same sentence. This practice, commonly known as code-mixing, introduces additional complexity because language switching, transliteration, and informal expressions can influence the sentiment conveyed by a sentence.

Traditional machine-learning approaches can provide useful baselines by extracting character-level patterns from text. However, they may have limitations in representing contextual relationships. Multilingual transformer models offer another approach by learning contextual representations across different languages.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov954

These challenges motivate an empirical comparison of traditional character-based features and multilingual transformer representations for Telugu and Telugu-English code-mixed sentiment classification.

➢ *Problem Statement*

Sentiment classification for Telugu and Telugu-English code-mixed text remains challenging because of limited annotated resources, spelling variations, informal expressions, and language mixing. A model that performs well on standard Telugu text may not necessarily achieve similar performance on code-mixed text.

Therefore, this study investigates how two different approaches perform across these text types and examines whether the observed performance differences are statistically meaningful.

➢ *Research Objectives*

The main objectives of this study are:

- To investigate sentiment classification in Telugu and Telugu-English code-mixed text using the ACTSA and CMTET datasets.
- To develop a character-level TF-IDF and Logistic Regression baseline for sentiment classification.
- To evaluate XLM-RoBERTa as a multilingual transformer-based approach.
- To compare both approaches using accuracy, precision, recall, and macro F1-score.
- To examine model performance through statistical significance testing and prediction error analysis.
- To identify the strengths and limitations of both approaches across the two datasets.

➢ *Research Contributions*

This study provides an empirical comparison of a character-level traditional machine-learning approach and a multilingual transformer model across two Telugu sentiment classification datasets.

- *The Main Contributions are:*

✓ A comparative evaluation of two different modelling approaches on Telugu and Telugu-English code-mixed sentiment data.
✓ An analysis using multiple classification metrics, including macro F1-score, to assess performance across sentiment classes.
✓ Statistical comparison using McNemar's exact test and paired bootstrap confidence intervals.
✓ An examination of model agreement, disagreements, and prediction errors to better understand classification behaviour.

II. RELATED WORK AND RESEARCH GAP

➢ *Sentiment Analysis in Low-Resource Languages*

Sentiment analysis is an important natural language processing task that focuses on identifying opinions, attitudes, and emotional expressions in textual data. Although substantial progress has been achieved for resource-rich languages such as English, developing reliable sentiment classification systems for low-resource languages remains challenging. These difficulties are associated with the limited availability of annotated datasets, variations in linguistic structures, differences in writing conventions, and the lack of language-specific computational resources. Such limitations can affect both model training and the reliability of performance evaluation.

Traditional machine-learning approaches, including Logistic Regression and Support Vector Machines, have been widely used for text classification. These methods commonly employ feature extraction techniques such as Term Frequency–Inverse Document Frequency (TF-IDF) to represent textual information numerically. Character-level and subword-based representations are particularly relevant when processing languages that contain spelling variations, morphological differences, and informal expressions. Bojanowski et al. [1] demonstrated the usefulness of subword information in word representations, while Zhang et al. [2] explored character-level representations for text classification. These approaches provide useful alternatives when conventional word-level features are insufficient to capture variations in written text.

More recently, multilingual transformer-based models have received considerable attention for natural language processing in resource-limited languages. Chattu et al. [3] investigated transformer-based approaches for Telugu sentiment classification using multiple domain-specific datasets. Their study evaluated IndicBERT, RoBERTa, DeBERTa, and XLM-RoBERTa, highlighting the growing application of pretrained multilingual representations to Telugu text classification. Their findings provide further evidence of the relevance of transformer-based approaches for sentiment analysis in languages with comparatively limited computational resources.

Recent research has also expanded the scope of low-resource sentiment analysis beyond conventional classification performance. Raj Kumar et al. [4] introduced TeSent, a large-scale Telugu sentiment dataset containing human-selected rationales alongside sentiment labels. Their study examined transformer-based models through classification performance, explanation quality, and social bias analysis. This work illustrates the increasing importance of annotation quality, interpretability, and fairness when developing sentiment analysis systems for underrepresented languages.

Despite these developments, the effectiveness of different text representation approaches can vary according to the linguistic characteristics and datasets being examined. Traditional character-based methods offer a comparatively simple approach to capturing textual patterns, whereas multilingual transformers provide contextual representations learned from large-scale multilingual pretraining. Consequently, evaluating these approaches on relevant low-resource language datasets remains important for

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov954

understanding their respective strengths, limitations, and practical applicability.

➢ *Telugu Sentiment Analysis*

Telugu sentiment analysis continues to face challenges because the availability of annotated language resources is relatively limited compared with that of widely studied languages. The Annotated Corpus for Telugu Sentiment Analysis (ACTSA), presented by Mukku and Mamidi [5], is an important resource in this area. It contains manually annotated Telugu sentences and provides a basis for evaluating sentiment classification across three categories.

Further studies have explored different machine-learning and deep-learning techniques for Telugu sentiment classification. Research has also examined transformer-based approaches for analysing Telugu text across multiple domains [3]. These developments show the growing interest in applying computational methods to Telugu language processing.

However, the presence of spelling differences, informal expressions and variations in sentence structure continues to make classification challenging. Although existing datasets support experimental investigation, the effectiveness of a model needs to be assessed through appropriate evaluation measures and carefully designed experiments.

➢ *Telugu-English Code-Mixed Sentiment Analysis*

Code-mixing occurs when speakers combine words or expressions from two or more languages within the same sentence. In Telugu-English communication, this often involves switching between Telugu and English, along with differences in spelling, transliteration and vocabulary. Such variations create additional difficulties for automated sentiment classification.

Kusampudi et al. [6] studied sentiment analysis for Telugu-English code-mixed text using unsupervised data normalization and a multilayer perceptron. Their work highlighted the relevance of text normalization when processing informal and mixed-language expressions.

Other research has explored syllable-based and word-based embeddings for sentiment classification in Telugu-English text [7]. Deep-learning methods, including RNN, GRU, BiLSTM, CNN and BERT, have also been examined for sentiment classification using the CMTET dataset [8].

In addition, a broader survey of Telugu-English code-mixed natural language processing has reviewed the tasks, resources and techniques available for this language setting [9]. Collectively, these studies show the importance of accounting for linguistic variation and mixed-language expressions when developing sentiment classification systems.

➢ *Multilingual Transformer-Based Approaches*

Transformer-based language models have made it possible to represent words and sentences using contextual information. Unlike approaches that rely mainly on manually designed or frequency-based features, these models learn relationships between tokens within a sequence. Devlin et al. [10] introduced BERT, demonstrating the usefulness of bidirectional pretraining for a range of natural language processing tasks.

Conneau et al. [11] introduced XLM-RoBERTa (XLM-R), a multilingual transformer pretrained on text from 100 languages. Their research demonstrated the potential of multilingual representations for cross-lingual understanding. The Cross-lingual Natural Language Inference (XNLI) benchmark provides another framework for assessing sentence representations across multiple languages [12].

Multilingual transformers offer an alternative to traditional feature-based classification methods. Nevertheless, results obtained on general multilingual benchmarks do not automatically indicate how well a model will perform on a particular low-resource language or code-mixed dataset. Their effectiveness therefore needs to be examined using experiments conducted on the target language data.

➢ *Research Gap*

Previous studies have explored several approaches to sentiment analysis in Telugu and Telugu-English code-mixed text. Research using the ACTSA dataset has contributed to the development and evaluation of sentiment classification methods for Telugu [5], while studies involving CMTET have investigated the challenges associated with mixed-language expressions, text normalization and different feature representations [6]–[9]. In parallel, multilingual transformer models have provided contextual representations that can be applied to a range of natural language processing tasks [10], [11]. These studies provide a useful foundation for sentiment classification in both language settings.

However, the approaches used in these studies differ in their feature representations, model architectures and evaluation procedures. As a result, comparing their reported performance directly may not provide a consistent basis for understanding the relative behaviour of traditional machine-learning methods and multilingual transformer models. In particular, character-level features and contextual multilingual representations capture textual information in different ways, and their effectiveness may vary between standard Telugu text and Telugu-English code-mixed text.

This study addresses this comparison by evaluating Character-Level TF-IDF with Logistic Regression and XLM-RoBERTa on two datasets: ACTSA for Telugu sentiment classification and CMTET for Telugu-English code-mixed sentiment classification. Both approaches are assessed using the same respective test subsets and a common set of evaluation measures, including accuracy, macro precision, macro recall and macro F1-score. This provides a consistent experimental basis for examining their performance across the two language settings.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov954

In addition to comparing classification metrics, the study examines whether the observed differences are statistically meaningful through McNemar's exact test and paired bootstrap confidence intervals. Model agreement, prediction disagreements and high-confidence classification errors are also analysed to provide further insight into the behaviour and reliability of the evaluated approaches.

The contribution of this work is therefore an empirical comparison of established classification methods across two Telugu-related sentiment datasets, supported by statistical evaluation and prediction analysis. Rather than proposing a new model architecture, the study investigates how character-level traditional features and multilingual contextual representations perform under a consistent evaluation framework, and identifies differences and limitations that can inform further research in Telugu and Telugu-English sentiment classification.

III. DATASET AND EXPLORATORY DATA ANALYSIS

➢ *Dataset Overview*

This research uses two publicly available sentiment datasets to examine classification performance in Telugu and Telugu-English code-mixed text. The selected datasets are the Annotated Corpus for Telugu Sentiment Analysis (ACTSA) and the Code-Mixed Telugu-English Text (CMTET) dataset. Both datasets support three-class sentiment classification, consisting of negative, neutral and positive categories.

ACTSA is considered for evaluating sentiment classification in Telugu text, whereas CMTET is used to study text that combines Telugu and English. The use of these two datasets provides an opportunity to assess the selected classification approaches across two different language settings.

Table 1 Overview of the Datasets Used in this Study.

Property	ACTSA	CMTET
Dataset type	Telugu sentiment	Telugu-English code-mixed sentiment
Total records	5,280	19,673
Number of classes	3	3
Training records	4,224	15,738
Validation records	528	1,967
Testing records	528	1,968
Split ratio	80:10:10	Approximately 80:10:10

➢ *ACTSA Dataset*

The Annotated Corpus for Telugu Sentiment Analysis (ACTSA) is used in this study to evaluate sentiment classification on Telugu-language text. The experimental dataset contains 5,280 records after the data preparation and filtering procedures.

The original ACTSA corpus is reported to contain 5,457 annotated records, while the downloaded raw dataset used in the experiments contained 5,410 parseable records. During preprocessing, 69 records associated with conflicting sentiment labels were removed. A further 61 duplicate records were also excluded, resulting in 5,280 records for the experiments. The difference between the reported corpus size and the downloaded raw file was present before preprocessing.

The prepared dataset was divided into training, validation and testing subsets using an 80:10:10 split. Each record belongs to one of three sentiment categories: negative, neutral or positive. The original labels, represented by -1, 0 and 1, were converted into 0, 1 and 2, respectively, for model training and evaluation.

The resulting dataset is used to examine the behaviour of character-level text features and multilingual contextual representations when applied to Telugu sentiment classification.

➢ *CMTET Dataset*

The Code-Mixed Telugu-English Text (CMTET) dataset is used to evaluate sentiment classification in text containing both Telugu and English expressions. This type of communication can include language switching, informal wording, spelling differences and transliterated expressions, which make text processing more challenging.

The experimental version of CMTET contains 19,673 records distributed across three sentiment categories: negative, neutral and positive. The original labels, NEG, NTL and POS, are converted into numerical values of 0, 1 and 2, respectively, to maintain consistency during model training and evaluation.

The dataset provides a basis for comparing the two classification approaches on Telugu-English code-mixed text. It also supports further examination of model performance across different language-composition groups.

➢ *Class Distribution*

The distribution of sentiment categories is examined for both datasets to understand how the available records are spread across negative, neutral and positive classes. This examination is relevant because differences in class frequency can influence the behaviour and evaluation of classification models.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov954

In ACTSA, the positive category contains more records than the negative and neutral categories. A similar difference can be observed in CMTET, where the neutral category has fewer records than the other two sentiment categories.

Table 2 Overall Sentiment Class Distribution Across the Experimental Splits.

Dataset	Negative	Neutral	Positive
ACTSA	1,397	1,470	2,413
CMTET	7,659	4,150	7,864

The observed class distributions indicate that the datasets are not perfectly balanced across all sentiment categories. Therefore, evaluating model performance only through accuracy may not provide a complete picture. Macro F1-score is also considered alongside accuracy, precision and recall to assess performance across the three sentiment classes.

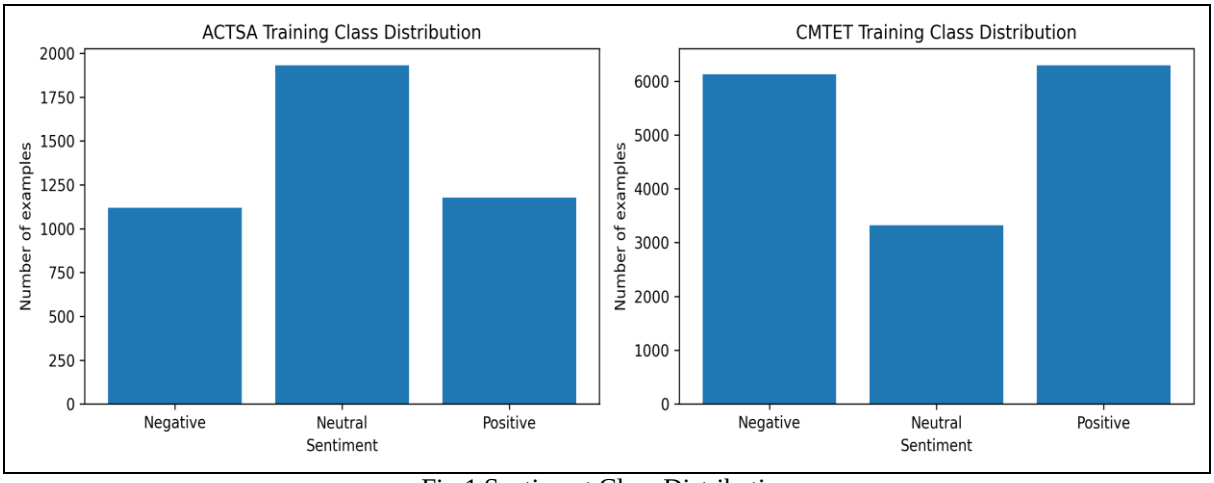

Fig 1 Sentiment Class Distribution

➢ *Dataset Splitting and Integrity*

Both datasets are divided into training, validation and testing subsets using an approximately 80:10:10 partition. The training data are used for model fitting, while the validation data support model selection and development. The test data are kept separate for the final performance evaluation.

The prepared dataset files were checked for missing text, missing sentiment labels and overlapping text records across the three subsets. The verification also considered both exact and normalized text overlap within each dataset.

The checks identified no missing text records or missing labels in either dataset. No train-test, train-validation or validation-test text overlap was detected in the prepared splits.

Table 3 Dataset Integrity Verification.

Verification	ACTSA	CMTET
Missing text records	0	0
Missing labels	0	0
Train-test text overlap	0	0
Train-validation text overlap	0	0
Validation-test text overlap	0	0

These verification results indicate that the prepared subsets contain no detected text overlap among training, validation and testing data. Maintaining this separation helps reduce the possibility of evaluating a model using text examples that were already present in another subset.

A. Data Preprocessing

➢ *Overview of Data Preprocessing*

Data preprocessing is an important stage in the experimental workflow because the quality and consistency of the input data can influence the performance of the classification models. In this study, the ACTSA and CMTET datasets were prepared independently while following a consistent preprocessing procedure. The process involved organizing the dataset records, verifying the consistency of text and sentiment labels, converting sentiment categories into numerical representations, and preparing the text according to the requirements of each classification approach.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov954

➢ *Text Preparation*

The text records were obtained from the prepared CSV files for both datasets. During preparation, the linguistic characteristics of each dataset were retained. In particular, the Telugu text in ACTSA was kept distinct from the Telugu-English code-mixed content in CMTET.

The preprocessing approach avoided excessive normalization that could remove useful linguistic information or change the original meaning of the text. Unnecessary modifications to the text were also avoided, particularly those that could affect sentiment-bearing words and expressions. This helped retain the characteristics of the original text for subsequent model training and evaluation.

➢ *Sentiment Label Encoding*

Both datasets contain three sentiment categories: Negative, Neutral, and Positive. Since the classification models require numerical class labels, the original sentiment categories were converted into integer values. The same mapping was maintained throughout model training, validation, testing, and evaluation.

For the ACTSA dataset, the original labels were mapped as follows: Negative (-1) to 0, Neutral (0) to 1, and Positive (1) to 2. For CMTET, the labels NEG, NTL, and POS were mapped to 0, 1, and 2, respectively.

Table 4 Sentiment Label Mapping Used in the Experiments.

Dataset	Original label	Numerical label
ACTSA	Negative (-1)	0
ACTSA	Neutral (0)	1
ACTSA	Positive (1)	2
CMTET	NEG	0
CMTET	NTL	1
CMTET	POS	2

➢ *Feature Preparation for Character-Level TF-IDF*

For the traditional machine learning approach, the text data was converted into numerical feature vectors using character-level Term Frequency–Inverse Document Frequency (TF-IDF). Character n-grams ranging from 2 to 5 were considered, with the maximum number of features limited to 100,000. The minimum document frequency was set to 2, and sublinear term-frequency scaling was enabled.

Character-level features were used to represent textual patterns that may occur within words, including parts of words, spelling variations, and morphological characteristics. The resulting TF-IDF vectors were used as input to the Logistic Regression classifier.

➢ *Tokenization for XLM-RoBERTa*

For the transformer-based approach, the text was processed using the pretrained XLM-RoBERTa tokenizer. The tokenizer converted the input text into token IDs suitable for processing by the model. A maximum sequence length of 128 tokens was used, and the resulting inputs were organized into batches for model processing.

This tokenization procedure provided the input representation required for fine-tuning XLM-RoBERTa on the two sentiment classification datasets.

➢ *Preprocessing Consistency*

The preprocessing procedures were applied consistently across the training, validation, and testing subsets. For the TF-IDF approach, the vectorizer was fitted using only the training data. The fitted vectorizer was subsequently used to transform the validation and test data. This procedure prevented information from the evaluation subsets from influencing feature construction during training.

For XLM-RoBERTa, the same pretrained tokenizer and maximum sequence length were used across all three subsets. Maintaining consistent preprocessing conditions ensured that the models received inputs prepared according to the same procedures during training and evaluation.

IV. PROPOSED METHODOLOGY

➢ *Overall Research Framework*

This study presents a comparative framework for sentiment classification in Telugu and Telugu-English code-mixed text. Two approaches are investigated: character-level TF-IDF with Logistic Regression and the multilingual transformer model XLM-RoBERTa. The experiments are carried out independently using the ACTSA and CMTET datasets.

The research framework follows six stages. First, the ACTSA and CMTET datasets are prepared and organized into training, validation, and testing subsets. Second, the text is converted into the input representation required by each approach: character-level TF-IDF features for Logistic Regression and token representations for XLM-RoBERTa.

In the third stage, Logistic Regression is trained using the TF-IDF features, while XLM-RoBERTa is fine-tuned using the training data. Fourth, validation performance, particularly the macro F1-score, is used to select the XLM-RoBERTa checkpoint. Fifth, both approaches are evaluated using their reserved test subsets.

Finally, the classification metrics, statistical significance, and prediction errors are examined to compare the two approaches. This framework establishes a consistent experimental basis for evaluating traditional character-based features against multilingual transformer representations.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov954

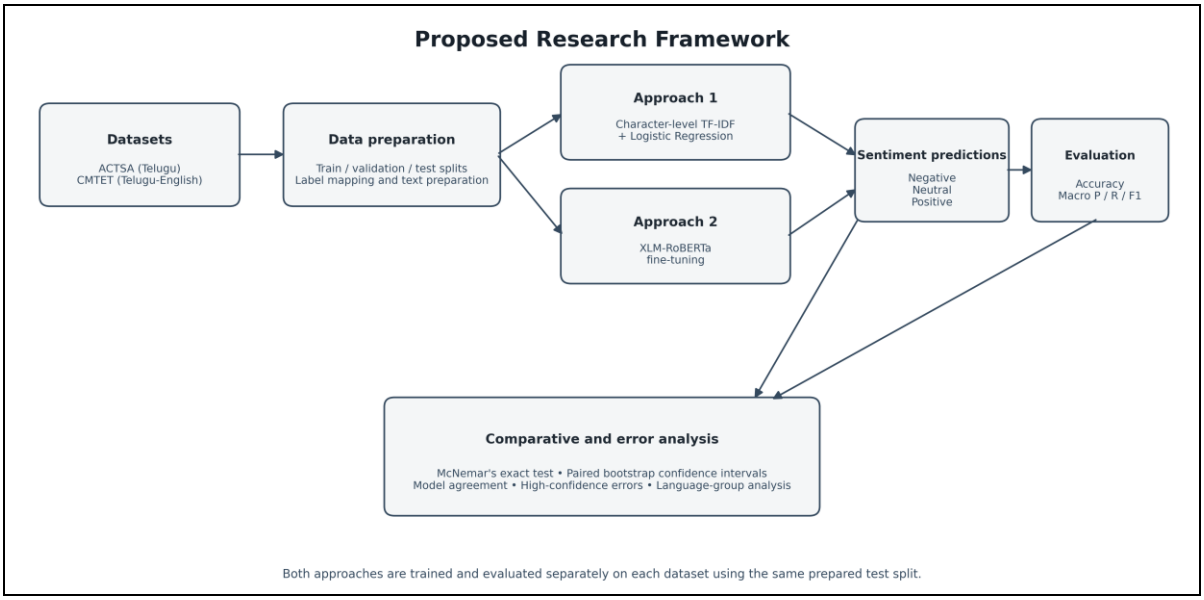

Fig 2 Proposed Research Framework

➢ *Character-Level TF-IDF with Logistic Regression*

The first classification approach combines character-level Term Frequency–Inverse Document Frequency (TF-IDF) feature extraction with Logistic Regression. This approach serves as the traditional machine learning baseline in the comparative study.

TF-IDF transforms text into numerical feature vectors by assigning weights according to the frequency and distribution of terms or character sequences. Instead of using complete words, this experiment uses character n-grams with lengths ranging from 2 to 5. The vectorizer is configured with a maximum of 100,000 features and a minimum document frequency of 2. Sublinear term-frequency scaling is also enabled.

The extracted feature vectors are supplied to a Logistic Regression classifier. Balanced class weights are used to account for differences in the number of samples across sentiment categories. The classifier is configured with a maximum of 1,500 iterations. Its regularization parameter (C) is set to 0.5 for ACTSA and 2.0 for CMTET.

➢ *XLM-RoBERTa Architecture*

The second approach uses XLM-RoBERTa, a multilingual transformer-based language model built on the RoBERTa architecture and pretrained on multilingual text.

In this study, the pretrained xlm-roberta-base model is fine-tuned independently on the ACTSA and CMTET datasets. The input text is processed using the corresponding XLM-RoBERTa tokenizer, which converts the text into token IDs and attention masks required by the model.

During fine-tuning, the model learns task-specific representations and predicts one of the three sentiment categories: negative, neutral, or positive. Unlike the TF-IDF baseline, which relies on character-level feature frequencies, XLM-RoBERTa uses contextual representations produced through its transformer layers. The two approaches therefore provide different ways of representing Telugu and Telugu-English code-mixed text.

➢ *Model Training Configuration*

Separate training configurations are used for the two classification approaches. The experimental settings for character-level TF-IDF with Logistic Regression and XLM-RoBERTa are summarized in Tables 5 and 6, respectively.

Table 5 Character-Level TF-IDF and Logistic Regression Configuration.

Parameter	**Configuration**
Feature extraction	Character-level TF-IDF
Analyzer	Character
Character n-gram range	2–5
Maximum features	100,000
Minimum document frequency	2
Sublinear TF	Enabled
Classifier	Logistic Regression
Maximum iterations	1,500
Class weights	Balanced
ACTSA regularization (C)	0.5
CMTET regularization (C)	2.0

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov954

Table 6 XLM-RoBERTa Training Configuration.

Parameter	Configuration
Pretrained model	xlm-roberta-base
Maximum sequence length	128
Batch size	8
Number of epochs	5
Learning rate	2×10^{-5}
Optimizer	AdamW
Learning-rate scheduler	Linear
Warm-up ratio	0.1
Gradient clipping	1.0
Numerical precision	FP32
Checkpoint selection	Validation macro F1
Random seed	42

The XLM-RoBERTa model is fine-tuned independently for each dataset. The checkpoint that achieves the highest validation macro F1-score is selected for the final test evaluation.

➢ *Evaluation Metrics*

The performance of both classification approaches is evaluated using accuracy, precision, recall, and macro F1-score. These measures provide an overall assessment of classification performance and help examine how effectively the models identify the individual sentiment categories.

- Accuracy: Measures the proportion of correctly classified samples among all evaluated samples.

Accuracy = Number of Correct Predictions / Total Number of Predictions

- Precision: Measures the proportion of correctly predicted instances within all instances predicted as a particular class.

Precision = True Positives / (True Positives + False Positives)

- Recall: Measures the proportion of actual instances of a class that are correctly identified.

Recall = True Positives / (True Positives + False Negatives)

- F1-score: Represents the harmonic mean of precision and recall.

F1-score = 2 × (Precision × Recall) / (Precision + Recall)

Macro F1-score is calculated by taking the unweighted average of the F1-scores for the three sentiment categories. Each category therefore receives equal importance, irrespective of the number of samples it contains.

These evaluation metrics are used to compare the two approaches on both datasets and provide the basis for the subsequent statistical analysis and examination of prediction errors.

A. Experimental Setup

➢ *Experimental Environment*

The experiments were conducted in a Windows-based Python environment. Scikit-learn was used for character-level TF-IDF feature extraction and Logistic Regression, whereas PyTorch and the Hugging Face Transformers library were used for fine-tuning and evaluating XLM-RoBERTa.

The hardware configuration consisted of a 13th Generation Intel Core i5-1335U processor, Intel Iris Xe Graphics, and 16 GB of system memory. The experiments were executed using CPU-based computing.

Python served as the primary programming language. The software environment included PyTorch, Hugging Face Transformers, Scikit-learn, NumPy, Pandas, Matplotlib, and Jupyter Notebook. The experiments used the CPU-based PyTorch installation without NVIDIA CUDA acceleration.

➢ *Training and Validation Procedure*

The training subsets were used to fit the classification models. For the TF-IDF approach, the vectorizer was fitted using the training text, and the resulting feature representations were supplied to the Logistic Regression classifier. For XLM-RoBERTa, the training records were tokenized and processed in batches. The pretrained model was fine-tuned to classify the input text into the three sentiment categories.

The validation subsets were used to monitor model performance during development. For XLM-RoBERTa, the checkpoint with the highest validation macro F1-score was selected for the final evaluation. For both datasets, the best validation checkpoint was obtained at epoch 4. The corresponding validation macro F1-scores were 65.27% for ACTSA and 80.92% for CMTET.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov954

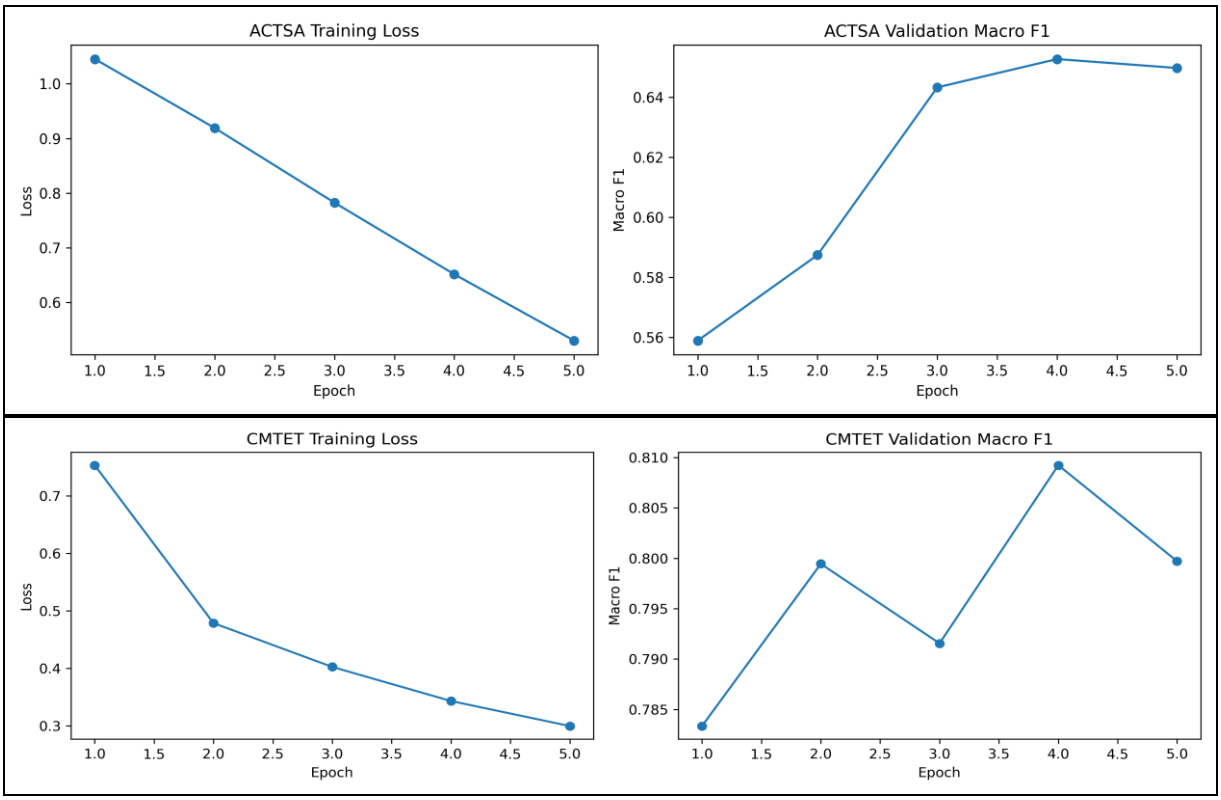

Fig 3 Model Training Curves

➢ *Final Test Evaluation*

After model selection, the selected models were evaluated on their respective test subsets. The test data remained separate from the training and validation procedures throughout the experiments.

Both classification approaches were evaluated using the same test records within each dataset. Accuracy, macro precision, macro recall, and macro F1-score were calculated to assess and compare their classification performance.

➢ *Statistical Analysis Procedure*

In addition to the standard classification metrics, statistical tests were applied to examine whether the observed performance differences between the two approaches were statistically meaningful.

McNemar's exact test was used to compare the paired classification outcomes produced by the two models on the same test samples. This test examines differences in their prediction outcomes rather than relying only on a comparison of overall accuracy [14].

Paired bootstrap confidence intervals were also calculated to estimate the uncertainty in the difference between the macro F1-scores of the two approaches [13].

V. RESULTS AND PERFORMANCE EVALUATION

➢ *ACTSA Dataset Results*

The ACTSA dataset was used to assess the performance of both classification approaches on Telugu text. Its test subset consists of 528 records covering the three sentiment categories. The evaluation results for character-level TF-IDF with Logistic Regression and XLM-RoBERTa are presented in Table 7.

The results show that XLM-RoBERTa achieves an accuracy of 61.55%, compared with 58.33% for the TF-IDF baseline. Its macro F1-score is 60.50%, while the baseline obtains 57.97%. XLM-RoBERTa also records higher macro precision, whereas the TF-IDF approach achieves a slightly higher macro recall.

These results indicate higher numerical accuracy and macro F1-score for XLM-RoBERTa on the ACTSA test subset. However, the statistical significance of the observed differences is examined separately in the statistical analysis section.

Table 7 Model Performance on the ACTSA Test Dataset.

Evaluation Metric	TF-IDF + Logistic Regression	XLM-RoBERTa
Accuracy	58.33%	61.55%
Macro Precision	57.87%	61.69%
Macro Recall	58.82%	59.88%
Macro F1-score	57.97%	60.50%

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov954

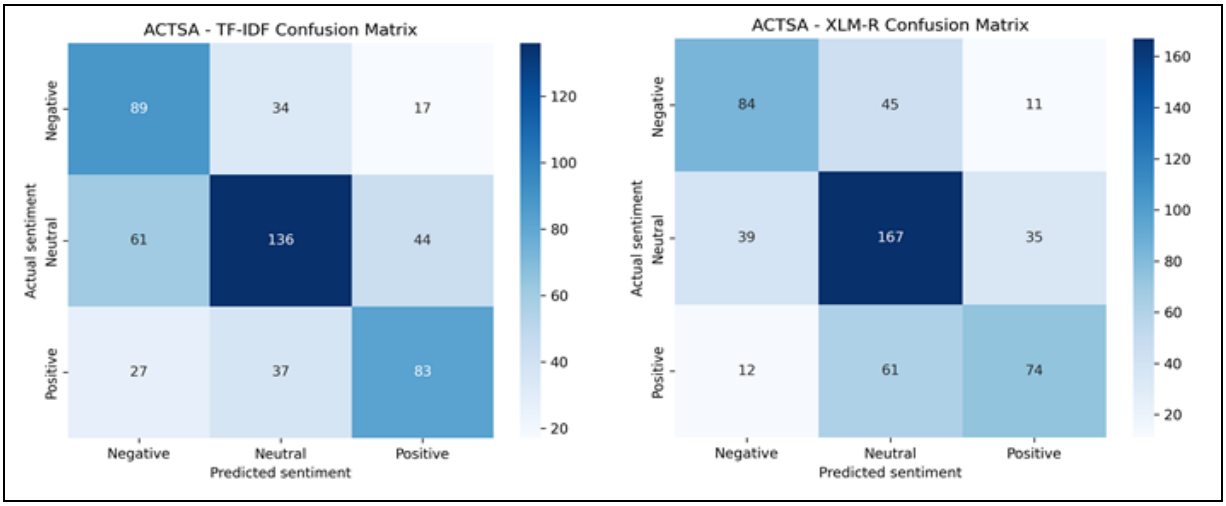

Fig 4 ACTSA Confusion Matrices

➢ *CMTET Dataset Results*

The CMTET dataset was used to evaluate sentiment classification performance on Telugu-English code-mixed text. The test subset contains 1,968 records. The results obtained by both approaches are summarized in Table 8.

XLM-RoBERTa obtains an accuracy of 81.55% and a macro F1-score of 79.98%. The TF-IDF baseline records 80.74% accuracy and a macro F1-score of 79.56%. The difference between the two approaches is relatively small.

A comparison of the individual metrics shows that XLM-RoBERTa achieves slightly higher accuracy, macro precision, and macro F1-score. In contrast, TF-IDF obtains a slightly higher macro recall. These findings indicate that both approaches provide competitive results on the CMTET test subset.

Table 8 Model Performance on the CMTET Test Dataset.

Evaluation Metric	TF-IDF + Logistic Regression	XLM-RoBERTa
Accuracy	80.74%	81.55%
Macro Precision	79.35%	79.83%
Macro Recall	80.90%	80.65%
Macro F1-score	79.56%	79.98%

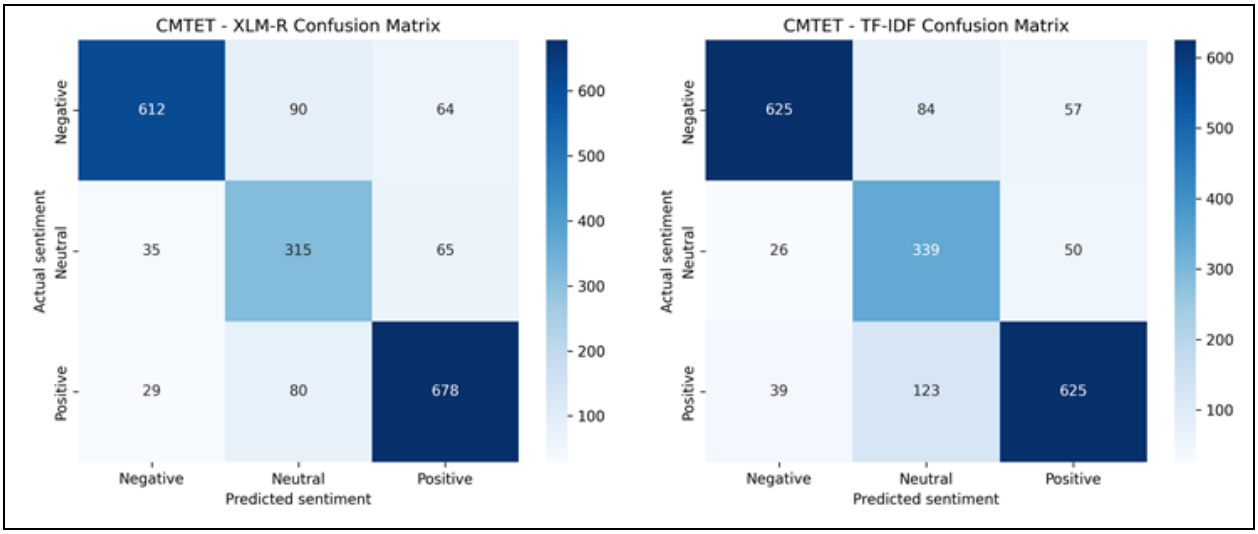

Fig 5 CMTET Confusion Matrices

➢ *Comparative Model Performance*

The results obtained from the two datasets provide an opportunity to compare the approaches under different language conditions. The performance differences between XLM-RoBERTa and the TF-IDF baseline are summarized in Table 9.

For ACTSA, XLM-RoBERTa records an accuracy improvement of 3.22 percentage points and a macro F1-score improvement of 2.53 percentage points over the baseline. For CMTET, the corresponding differences are 0.81 and 0.42 percentage points, respectively.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov954

Although XLM-RoBERTa obtains higher accuracy and macro F1-score on both datasets, the magnitude of the differences varies. The difference is more noticeable on ACTSA, whereas the two approaches produce closer results on CMTET. These numerical comparisons alone do not establish that either approach is consistently or statistically superior.

Table 9 Difference in Performance between XLM-RoBERTa and TF-IDF.

Dataset	Accuracy Difference	Macro F1 Difference
ACTSA	+3.22 percentage points	+2.53 percentage points
CMTET	+0.81 percentage points	+0.42 percentage points

➢ *Class-Wise Performance Analysis*

Overall evaluation metrics provide a general summary of classification performance, but they do not fully describe how each sentiment category is identified. To examine this aspect, class-wise precision, recall, and F1-score are considered using the classification reports generated from the final test predictions.

The analysis is presented separately for ACTSA and CMTET. It examines the classification performance for negative, neutral, and positive sentiment categories, allowing differences between the two approaches to be observed at the individual class level.

Table 1 Class-Wise Performance of Both Approaches on ACTSA.

Sentiment Class	TF-IDF Precision	TF-IDF Recall	TF-IDF F1-score	XLM-R Precision	XLM-R Recall	XLM-R F1-score	Support
Negative	50.28%	63.57%	56.15%	62.22%	60.00%	61.09%	140
Neutral	65.70%	56.43%	60.71%	61.17%	69.29%	64.98%	241
Positive	57.64%	56.46%	57.04%	61.67%	50.34%	55.43%	147
Macro Average	57.87%	58.82%	57.97%	61.69%	59.88%	60.50%	528
Weighted Average	59.37%	58.33%	58.48%	61.59%	61.55%	61.29%	528

Table 2 Class-Wise Performance of Both Approaches on CMTET.

Sentiment Class	TF-IDF Precision	TF-IDF Recall	TF-IDF F1-score	XLM-R Precision	XLM-R Recall	XLM-R F1-score	Support
Negative	90.58%	81.59%	85.85%	90.53%	79.90%	84.88%	766
Neutral	62.09%	81.69%	70.55%	64.95%	75.90%	70.00%	415
Positive	85.38%	79.42%	82.29%	84.01%	86.15%	85.07%	787
Macro Average	79.35%	80.90%	79.56%	79.83%	80.65%	79.98%	1,968
Weighted Average	82.49%	80.74%	81.20%	82.53%	81.55%	81.82%	1,968

A. Statistical Significance and Error Analysis

➢ *McNemar's Exact Test*

McNemar's exact test was applied to compare the classification outcomes of TF-IDF with Logistic Regression and XLM-RoBERTa using the same test samples. Rather than considering only the overall accuracy values, this test examines cases in which the two approaches produce different prediction outcomes. It was conducted independently on the ACTSA and CMTET test subsets [14].

The resulting p-values are presented in Table 12. For ACTSA, the p-value is 0.1957, while CMTET produces a p-value of 0.3447. Both values are greater than the commonly used significance threshold of 0.05.

Therefore, the statistical analysis does not provide sufficient evidence to conclude that either approach has a statistically significant advantage over the other. Although XLM-RoBERTa achieves higher overall accuracy on both datasets, these differences should not be interpreted as statistically established improvements.

Table 11 McNemar's Exact Test Results.

Dataset	McNemar's Exact p-value	Statistical Significance
ACTSA	0.1957	Not significant
CMTET	0.3447	Not significant

➢ *Paired Bootstrap Confidence Intervals*

Paired bootstrap analysis was conducted to estimate the uncertainty associated with the difference in macro F1-scores between the two approaches. The analysis uses paired test predictions and calculates a 95% confidence interval for the macro F1-score difference, defined as XLM-RoBERTa minus TF-IDF.

For ACTSA, the macro F1-score difference is +2.53 percentage points, with a 95% confidence interval ranging from -2.02 to +7.06 percentage points. For CMTET, the

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov954

difference is +0.42 percentage points, with an interval ranging from -1.18 to +2.08 percentage points.

Both confidence intervals include zero. This indicates that the observed differences do not establish a statistically significant advantage for either approach at the 95% confidence level. These findings are consistent with the results of McNemar's exact test.

Table 12 Paired Bootstrap Confidence Intervals for Macro F1-Score Differences.

Dataset	Macro F1 Difference	95% Confidence Interval
ACTSA	+2.53 percentage points	-2.02 to +7.06 percentage points
CMTET	+0.42 percentage points	-1.18 to +2.08 percentage points

➢ *Model Agreement and Disagreement Analysis*

Model agreement analysis was performed to examine how frequently the two approaches produced matching correct or incorrect outcomes on the test samples. The predictions were grouped into four categories: both models correct, both models incorrect, only XLM-RoBERTa correct, and only TF-IDF correct.

The results are presented in Table 14. For ACTSA, both models correctly classify 240 samples and incorrectly classify 135 samples. XLM-RoBERTa alone correctly classifies 85 samples, whereas TF-IDF alone correctly classifies 68 samples. The overall prediction agreement is 63.83%.

For CMTET, both models correctly classify 1,471 samples and incorrectly classify 245 samples. XLM-RoBERTa alone correctly classifies 134 samples, while TF-IDF alone correctly classifies 118 samples. The prediction agreement reaches 84.81%.

The results indicate a higher level of agreement on CMTET than on ACTSA. However, the models also produce different prediction outcomes on a portion of the test records. This shows that similar overall performance does not necessarily mean that both approaches make identical classification decisions.

Table 13 Agreement and Disagreement between the Two Models.

Prediction Category	ACTSA	CMTET
Both models correct	240	1,471
Both models incorrect	135	245
Only XLM-RoBERTa correct	85	134
Only TF-IDF correct	68	118
Total test samples	528	1,968
Prediction agreement	63.83%	84.81%

➢ *High-Confidence Prediction Errors*

An additional analysis was performed to examine incorrect XLM-RoBERTa predictions associated with confidence scores of at least 0.80. The purpose was to identify cases in which the model assigned a high confidence score despite predicting the wrong sentiment category.

The analysis identified 85 such incorrect predictions in ACTSA and 219 in CMTET. These findings demonstrate that a high confidence score does not necessarily guarantee a correct classification. Incorrect predictions may still receive strong confidence scores, particularly when the input contains expressions or linguistic patterns that are difficult to classify.

The identified records provide material for examining prediction reliability and understanding the behaviour of the model.

Table 14 High-Confidence XLM-RoBERTa Prediction Errors.

Dataset	Incorrect Predictions with Confidence ≥ 0.80
ACTSA	85
CMTET	219

➢ *Qualitative Error Analysis*

Prediction error files were generated to support the examination of incorrectly classified samples. These files include instances where both models produced incorrect predictions, along with examples of high-confidence errors made by XLM-RoBERTa. Selected records were also prepared separately for manual inspection.

The selected examples can be examined to identify possible sources of classification difficulty, such as ambiguous sentiment, informal expressions, spelling variation, and language mixing. However, the preparation of these records alone does not establish the reason behind any individual prediction error.

A detailed interpretation requires manual examination of the selected text samples and their corresponding

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov954

reference labels. Accordingly, this analysis provides a basis for studying differences in model predictions without making unsupported claims about the causes of misclassification.

VI. DISCUSSIONS

➢ *Overall Model Performance*

The experimental findings show that both character-level TF-IDF with Logistic Regression and XLM-RoBERTa can be used for sentiment classification in Telugu and Telugu-English code-mixed text.

On ACTSA, XLM-RoBERTa achieves 61.55% accuracy and a macro F1-score of 60.50%, compared with 58.33% accuracy and 57.97% macro F1-score for the TF-IDF baseline. This corresponds to improvements of 3.22 and 2.53 percentage points, respectively.

On CMTET, the difference between the approaches is smaller. XLM-RoBERTa obtains 81.55% accuracy and 79.98% macro F1-score, while TF-IDF records 80.74% accuracy and 79.56% macro F1-score.

These findings indicate that the performance gap depends on the dataset. Although XLM-RoBERTa records higher overall accuracy and macro F1-score in both experiments, its improvement is considerably smaller on CMTET.

The observed results can be considered alongside previous research on Telugu and Telugu-English sentiment classification. Chattu et al. [3] investigated transformer-based approaches for Telugu sentiment classification using multiple domain-specific datasets, demonstrating the relevance of pretrained language models for this task. Similarly, previous studies on Telugu-English code-mixed text have explored normalization techniques, embedding-based representations, and deep-learning approaches [6]–[8]. The present study extends this line of investigation by evaluating a character-level TF-IDF approach and XLM-RoBERTa under a common experimental framework across ACTSA and CMTET. Unlike comparisons based on results reported under different experimental settings, this evaluation uses the same respective test subsets and evaluation measures for both approaches. The findings show that the performance difference between the models varies across datasets, emphasizing the importance of evaluating approaches on the specific language setting rather than assuming that one representation method will perform consistently better in every case.

➢ *Performance on Telugu Text*

The ACTSA results show a numerical improvement for XLM-RoBERTa compared with the character-level TF-IDF baseline. One possible interpretation is that contextual representations provide useful information when identifying sentiment in Telugu sentences.

However, the performance of both approaches remains relatively limited on this dataset. The results also indicate that using a multilingual transformer does not automatically resolve the difficulties associated with sentiment classification in a low-resource language. Further investigation is needed to identify the linguistic characteristics associated with individual classification errors.

The difference between the two approaches may be related to how they represent textual information. Character-level TF-IDF captures patterns based on character sequences and can represent variations in spelling and word forms. However, it does not explicitly model the contextual relationships between words. In contrast, XLM-RoBERTa generates contextual representations that take surrounding tokens into account. This capability may be useful when sentiment depends on the context in which an expression appears. Nevertheless, the observed improvement on ACTSA should be interpreted cautiously because McNemar's test did not establish a statistically significant difference between the two approaches ($p = 0.1957$), and the paired bootstrap confidence interval for the macro F1-score difference included zero.

➢ *Performance on Telugu-English Code-Mixed Text*

On CMTET, both approaches achieve comparatively higher evaluation scores than on ACTSA. Their accuracy and macro F1-score values are close, indicating competitive performance on this dataset.

The character-level TF-IDF approach remains competitive despite not using contextual transformer representations. The difference between its macro F1-score and that of XLM-RoBERTa is only 0.42 percentage points.

The language-group analysis also reveals variations across different text-composition groups. XLM-RoBERTa achieves higher accuracy on the English-dominant subset, whereas TF-IDF performs better on the Telugu-dominant and code-mixed subsets. The difference between the approaches on the code-mixed subset is very small.

These subgroup findings should be interpreted cautiously because the number of test samples differs considerably across the language groups.

These findings suggest that the effectiveness of a text representation method may depend on the linguistic composition of the input. In Telugu-English code-mixed text, language switching, variations in spelling, and the combination of Telugu and English expressions can create challenges for consistent feature representation. Character-level TF-IDF can capture some of these variations through character sequences, which may help explain its competitive performance on the Telugu-dominant and code-mixed subsets. XLM-RoBERTa, on the other hand, can use contextual information across tokens and appears to perform better on the English-dominant subset in this experiment. However, these observations are specific to the evaluated CMTET subsets and should not be interpreted as evidence that either approach is generally better for a particular language composition.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov954

➢ *Interpretation of Statistical Results*

Although XLM-RoBERTa achieves higher overall scores on both datasets, the statistical analyses do not establish a significant difference between the two approaches.

McNemar's exact test produces p-values of 0.1957 for ACTSA and 0.3447 for CMTET. Furthermore, the paired bootstrap confidence intervals for the macro F1-score differences include zero for both datasets.

Consequently, the experimental evidence does not support the claim that XLM-RoBERTa is statistically superior to the TF-IDF baseline. A higher numerical score alone is insufficient to establish a reliable performance advantage. The statistical findings therefore support a cautious interpretation of the observed differences.

The agreement between the significance tests and the bootstrap confidence intervals provides a consistent basis for interpreting the results. Although XLM-RoBERTa obtains numerically higher accuracy and macro F1-scores, the available statistical evidence does not demonstrate that these improvements are reliable rather than potentially arising from variation in the evaluated test samples. This distinction is important when comparing models on low-resource datasets, where the number and distribution of test examples can influence measured performance. Therefore, the results should be understood as showing a numerical advantage for XLM-RoBERTa in the evaluated experiments, rather than establishing its general superiority over character-level TF-IDF with Logistic Regression.

➢ *Practical Implications*

The findings of this study have practical implications for selecting sentiment classification approaches for Telugu and Telugu-English code-mixed text. The results indicate that traditional machine-learning methods should not be overlooked when developing systems for low-resource language settings. Although XLM-RoBERTa achieves slightly higher overall accuracy and macro F1-scores in both experiments, the differences are not statistically significant. Therefore, model selection should consider the requirements of the intended application rather than relying exclusively on the highest evaluation score.

Character-level TF-IDF with Logistic Regression may be a suitable option for applications where a relatively simple feature-based classification approach is preferred. It achieves competitive performance on the CMTET dataset without requiring the use of a pretrained transformer. XLM-RoBERTa provides contextual representations and achieves higher numerical scores in the evaluated experiments, making it another relevant option when contextual language information is important. However, the present study does not include a direct benchmarking of training time, inference latency, memory consumption, or deployment cost. Consequently, no definitive conclusion can be drawn about which approach is more computationally efficient.

For practical applications such as Telugu social media analysis, customer feedback classification, and opinion mining, the choice of approach should depend on the characteristics of the input text, available computational resources, and the reliability requirements of the application. The prediction errors and high-confidence incorrect classifications observed in this study also indicate the importance of evaluating model reliability before using automated predictions in decision-making processes. Overall, both approaches offer useful alternatives for Telugu and Telugu-English sentiment classification, but further evaluation under real-world conditions is necessary before recommending either approach for a specific deployment scenario.

A. Limitations and Threats to Validity

➢ *Dataset Limitations*

The evaluation in this study is based on two datasets, ACTSA and CMTET, representing Telugu and Telugu-English code-mixed sentiment classification. Although these datasets provide a basis for comparing the selected approaches, they cannot be assumed to cover the full diversity of Telugu communication.

Both datasets consist of specific collections of annotated text. Their characteristics may differ from those of other communication sources, including social media posts, online discussions, and customer reviews. As a result, the reported performance values should be understood in relation to the datasets examined in this work rather than treated as representative of every Telugu-language setting.

➢ *Language and Annotation Challenges*

Telugu text may include spelling variations, informal expressions, transliterated words, and differences in sentence structure. These variations create additional challenges for sentiment classification. In Telugu-English code-mixed text, the use of words from both languages within a single sentence adds another layer of linguistic complexity.

Such variations can affect both character-level machine learning methods and multilingual transformer-based models. Furthermore, the experiments use the sentiment labels supplied with the respective datasets. Any limitations or inconsistencies in the original annotation process may therefore influence the evaluation outcomes.

➢ *Model and Experimental Limitations*

The experimental comparison is restricted to two approaches: Character-Level TF-IDF with Logistic Regression and XLM-RoBERTa. Other traditional machine learning techniques and multilingual language models were not included in the present evaluation.

The XLM-RoBERTa experiments were conducted using a fixed configuration, with a maximum sequence length of 128, a batch size of 8, a learning rate of 2×10^{-5}, and a maximum of five training epochs. Alternative configurations could produce different outcomes. Similarly,

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov954

the TF-IDF baseline was evaluated using fixed character n-gram and regularization settings. Therefore, the reported findings describe the performance of the configurations tested in this study and do not represent every possible configuration of the two approaches.

➢ *Statistical and Generalization Limitations*

The statistical tests did not identify a significant performance difference between the two approaches on either dataset. However, the absence of statistical significance should not be interpreted as proof that both approaches perform identically. It indicates that the available test-set evidence was insufficient to establish a statistically significant difference.

The language-group analysis performed on CMTET also requires cautious interpretation because the number of samples varies across the groups. In particular, the Other category contains only two records, which is too small a basis for drawing general conclusions about that group.

Another limitation is that the experiments use one prepared train-validation-test partition for each dataset. Evaluating the models across additional independent splits or external datasets would provide further evidence about the consistency of the observed results and their ability to generalize beyond the current experimental data.

➢ *Scope of the Findings*

The conclusions of this research are confined to three-class sentiment classification using the selected ACTSA and CMTET datasets and the model configurations described in the methodology. The findings should not be directly extended to other Telugu natural language processing tasks, different languages, unrelated domains, or real-time applications without further evaluation.

Within these boundaries, the study offers a controlled empirical comparison of character-level feature-based classification and multilingual transformer-based classification for Telugu and Telugu-English code-mixed text. The limitations identified above provide important context for interpreting the experimental results and indicate areas where further investigation would be valuable.

B. Future Work

The present research compares Character-Level TF-IDF with Logistic Regression and XLM-RoBERTa for sentiment classification involving Telugu and Telugu-English code-mixed text. The experimental findings also identify several directions through which this work can be extended.

➢ *Expansion of Telugu Sentiment Datasets*

Future studies could use larger and more varied Telugu sentiment datasets gathered from multiple sources, such as social media, product reviews, and online discussions. A broader collection of annotated text would make it possible to examine model performance across a wider range of writing styles and communication contexts.

Future dataset development could also focus on improving the diversity and quality of sentiment annotations. In particular, collecting examples from different domains, age groups, communication styles, and regional variations of Telugu could help capture a wider range of linguistic expressions. Multiple annotators could be involved in the labelling process, with disagreements reviewed to improve annotation consistency. In addition, maintaining balanced representation across negative, neutral, and positive sentiment categories would support more reliable evaluation. Such improvements could help determine whether classification approaches remain effective when applied to Telugu text from different sources and contexts.

➢ *Improved Handling of Code-Mixed Text*

Telugu-English code-mixed text involves variations in language switching, transliteration, spelling, and informal expressions. Further research could examine preprocessing and text-normalization methods designed to address these variations while retaining information relevant to sentiment classification.

Future research could investigate different text-normalization strategies for Telugu-English code-mixed data, including transliteration standardization, spelling correction, and the handling of frequently used mixed-language expressions. These techniques could be evaluated independently and in combination to determine their influence on classification performance. Another useful direction would be to compare models trained on normalized text with models trained on the original text. Such experiments could help establish whether normalization improves sentiment classification or removes useful linguistic information. Particular attention should be given to preserving sentiment-bearing expressions while reducing unnecessary textual variation.

➢ *Evaluation of Additional Multilingual Representations*

The comparison in this study includes XLM-RoBERTa as the multilingual transformer approach. Future investigations could include other multilingual language models and evaluate their performance alongside XLM-RoBERTa. Such comparisons may provide a clearer understanding of how different pretrained representations handle low-resource Telugu and code-mixed text.

Further investigations could compare multilingual transformer models with language-specific pretrained models to examine how their representations influence sentiment classification in Telugu. The comparison could include models with different pretraining objectives, vocabulary coverage, and language-resource availability. To ensure a meaningful evaluation, the models should be trained and tested using consistent dataset partitions, evaluation metrics, and experimental conditions. In addition to classification performance, future studies could examine training time, inference time, memory requirements, and model size. This would provide a broader understanding of the trade-offs between predictive performance and computational requirements when selecting representations for low-resource language applications.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov954

➢ *Cross-Dataset Generalization*

The current experiments evaluate ACTSA and CMTET independently. An additional direction would be to conduct cross-dataset experiments, where a model trained on one dataset is tested on another compatible dataset. This would help examine whether the evaluated approaches maintain their performance across different data sources.

Future experiments could investigate cross-dataset generalization using additional Telugu sentiment datasets with compatible sentiment categories and annotation schemes. A model trained on one dataset could be evaluated on an independent dataset without further training to examine how well it handles differences in data sources, writing styles, and domain characteristics. Additional experiments could also investigate domain adaptation, in which a model is adapted using a limited amount of labelled data from the target domain. For ACTSA and CMTET, direct cross-dataset evaluation would require particular care because they represent different language settings. Their differences in language composition and text characteristics should be considered when designing transfer experiments. Such investigations could provide a clearer understanding of the generalization capabilities and limitations of the evaluated approaches.

➢ *Extended Statistical Evaluation*

Further experiments could use multiple independent data splits and repeated training runs to assess the consistency of the observed performance. Additional statistical analyses could also be conducted to obtain stronger evidence when comparing different sentiment classification approaches.

Future studies could strengthen statistical evaluation by repeating experiments across multiple random seeds and independent train-validation-test partitions. This would help determine whether the observed performance differences remain consistent under different data partitions and training conditions. In addition to reporting mean performance, researchers could report standard deviations and confidence intervals for the evaluation metrics. Paired statistical comparisons could also be applied to the predictions obtained from corresponding test samples, with appropriate consideration of multiple comparisons when several models or datasets are evaluated. These procedures would provide a more comprehensive assessment of performance variability and help distinguish consistent improvements from differences that may arise due to experimental variation.

➢ *Practical Deployment*

The approaches investigated in this work could be explored further for practical sentiment classification tasks, including Telugu social media monitoring, customer feedback analysis, and opinion mining. Before such applications are considered, additional evaluation would be needed to examine computational efficiency, prediction reliability, and performance on real-world text.

Before deploying sentiment classification systems in practical applications, future work should evaluate their performance under real-world conditions, including noisy user-generated text, spelling variations, unseen expressions, and changing language patterns. Deployment-oriented studies should also measure inference latency, memory consumption, model size, and computational cost on the intended hardware. Human review mechanisms could be incorporated for uncertain or high-confidence incorrect predictions, particularly in applications where sentiment classification may influence important decisions. Such evaluations would help determine the practical suitability and reliability of the compared approaches beyond benchmark datasets.

Overall, these research directions offer opportunities to extend the present study through broader dataset coverage, improved treatment of language-specific challenges, further model comparisons, and more extensive evaluation of generalization and practical applicability for Telugu and Telugu-English sentiment classification.

VII. CONCLUSION

This study examined sentiment classification for low-resource Telugu and Telugu-English code-mixed text using two approaches: Character-Level TF-IDF with Logistic Regression and the multilingual transformer model XLM-RoBERTa. The experiments were conducted using the ACTSA and CMTET datasets, covering three sentiment categories: negative, neutral, and positive.

The experimental results showed that XLM-RoBERTa obtained an accuracy of 61.55% and a macro F1-score of 60.50% on the ACTSA test set. The Character-Level TF-IDF baseline achieved an accuracy of 58.33% and a macro F1-score of 57.97%. On the CMTET dataset, XLM-RoBERTa recorded 81.55% accuracy and 79.98% macro F1-score, compared with 80.74% accuracy and 79.56% macro F1-score for the TF-IDF approach.

Although XLM-RoBERTa produced higher overall accuracy and macro F1-scores on both datasets, the statistical evaluation did not establish a significant difference between the two approaches. The results from McNemar's exact test and paired bootstrap confidence intervals indicate that the observed numerical improvements should not be interpreted as statistically established superiority. This distinction is important when assessing the performance of classification models, particularly when the differences between their evaluation scores are relatively small.

The prediction error analysis also showed that the two approaches produced different predictions for a number of test samples. In addition, the language-group analysis on CMTET indicated variations in classification performance across different text-composition groups. These observations provide further context for understanding the evaluation results beyond the overall performance metrics.

Overall, this research presents an empirical comparison of character-level feature extraction and multilingual transformer-based representations for sentiment

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov954

classification in Telugu and Telugu-English code-mixed text. The findings emphasize the value of evaluating multiple performance measures, examining prediction errors, and applying statistical tests before drawing conclusions about model performance. Further investigation using larger and more diverse datasets, cross-dataset evaluation, and additional analysis of linguistic variations could extend this work and provide a broader understanding of sentiment classification in low-resource and code-mixed language settings.

DOCUMENT SCOPE AND EVIDENCE POLICY

This report presents a comparative investigation of sentiment classification for low-resource Telugu and Telugu-English code-mixed text. The study examines two approaches: Character-Level TF-IDF with Logistic Regression and the multilingual transformer model XLM-RoBERTa, using the ACTSA and CMTET datasets.

The report covers dataset preparation, preprocessing, model development, experimental evaluation, statistical significance testing, and prediction error analysis. Model performance is assessed using accuracy, precision, recall, and macro F1-score.

All findings and numerical results are based on the saved experimental outputs. The report distinguishes measured results from interpretation and avoids unsupported claims about model superiority. Statistical comparisons are reported objectively, and the limitations of the datasets and experimental setup are acknowledged when interpreting the findings.

REFERENCES

[1.] Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5, 135–146. https://doi.org/10.1162/tacl_a_00051

[2.] Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. Advances in Neural Information Processing Systems, 28. https://papers.nips.cc/paper/2015/hash/250cf8b51c773f3f8dc8b4be867a9a02-Abstract.html

[3.] Chattu, K., Reddy, K. A. N., Veesam, S. B., Chirumamilla, P. S., Babu, V. D., Prakash, K., Bansal, S., Faruque, M. R. I., & Al-Mugren, K. S. (2025). Sentiment classification for Telugu using transformed based approaches on a multi-domain dataset. Scientific Reports, 15, Article 22185. https://doi.org/10.1038/s41598-025-05703-9

[4.] Raj Kumar, V., Ashwin, S., Manna, S., Sett, N., Hema Harshitha, C. V. S. N. M., Harshitha, K., Sharma, A. K., Deepakraj, B., Sarkar, T., Krishna, B. N., & Shakeer, S. (2026). Human-centered supervision for sentiment analysis in Telugu: A systematic inquiry beyond accuracy. Findings of the Association for Computational Linguistics: ACL 2026, 29519–29537.

[5.] Mukku, S. S., & Mamidi, R. (2017). ACTSA: Annotated corpus for Telugu sentiment analysis. Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems, 54–58. https://doi.org/10.18653/v1/W17-5408

[6.] Kusampudi, S. S. V., Sathineni, P., & Mamidi, R. (2021). Sentiment analysis in code-mixed Telugu-English text with unsupervised data normalization. Proceedings of RANLP, 753–760. https://doi.org/10.26615/978-954-452-072-4_086

[7.] Rayala, U. R., Seshadri, K., & Sristy, N. B. (2023). Sentiment analysis of code-mixed Telugu-English data leveraging syllable and word embeddings. ACM Transactions on Asian and Low-Resource Language Information Processing, 22(10). https://doi.org/10.1145/3620670

[8.] Suraj, N. S., Manda, A. K., & Raj, V. (2024). Sentiment analysis in code-mixed Telugu-English text (CMTET) using deep learning techniques. Proceedings of the 2024 3rd Edition of IEEE Delhi Section Flagship Conference (DELCON). https://doi.org/10.1109/DELCON64804.2024.10866784

[9.] Maddu, S., & Sanapala, V. R. (2024). A survey on NLP tasks, resources and techniques for low-resource Telugu-English code-mixed text. ACM Transactions on Asian and Low-Resource Language Information Processing. https://doi.org/10.1145/3695766

[10.] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[11.] Conneau, A., et al. (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

[12.] Conneau, A., et al. (2018). XNLI: Evaluating cross-lingual sentence representations. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2475–2485. https://aclanthology.org/D18-1269/

[13.] Dror, R., Baumer, G., Shlomov, S., & Reichart, R. (2018). The hitchhiker's guide to testing statistical significance in natural language processing. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 1383–1392. https://doi.org/10.18653/v1/P18-1128

[14.] McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, 12, 153–157.