MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 3302-3307

Multilingual Customer Feedback Analyzer Using MuRIL

Aayush Mishra, *Dept. of Computing technologies School of Computing*, *SRM Institute of Science and Technology* Kattankulathur, India, Email id: am4776@srmist.edu.in
Shreyansh Parvadia, *Dept. of Computing technologies School of Computing*, *SRM Institute of Science and Technology* Kattankulathur, India, Email id: sp3559@srmist.edu.in
Dr. A Pandian*, Professor, *Dept. of Computing technologies School of Computing*, *SRM Institute of Science and Technology* Kattankulathur, India, Email id: pandiana@srmist.edu.in
*Corresponding Author: pandiana@srmist.edu.in

Abstract: Customer feedback is considered one of the richest resources for an organization. It provides value regarding knowledge on product performance, user satisfaction, and areas that need improvement. However, continuous analysis of a large volume of unstructured feedback in diverse languages is very time consuming and inefficient. This paper presents a multilingual approach to automatically classify reviews in the English, Hindi, and Gujarati languages into categories like bug reports, feature requests, positive feedback, and others. The proposed system uses MuRIL (Multilingual Representations for Indian Languages), a transformer-based model designed for Indic languages. This model captures contextual and cross-lingual meanings. This model results in strong performance when fine-tuned on a balanced multilingual dataset for classification across various language structures. This framework is scalable, automated, and cloud-based feedback analysis. It supports the organizations in making quicker, data-driven choices, improving product quality, and increasing overall customer satisfaction.

Keywords: Multilingual NLP, MuRIL, Customer Feedback Classification, Indic Languages, Transformer Models, Cross-Lingual Learning, Sentiment Analysis, AI-driven Analytics

1. INTRODUCTION

In the contemporary technology era of digital transformation, there is an increasing reliance on customer feedback to gauge how well a product works, as well as how well a system works. With the tremendous surge in different online tools such as mobile apps, e-commerce websites, questionnaires, and social media platforms, the expansion in customer feedback has been tremendous. Analysis of large volumes of such unorganized data is unmaintainable and highly problematic, which signals the need for intelligent analysis of customer feedback.Although various techniques have already been proposed that emphasize the processing of English language feedback, a large percentage of reviews in user feedback, especially in the case of countries like India in the Asia region, which is culturally diverse, use languages like Hindi and Gujarati. Traditional NLP techniques perform poorly in the processing of such multilingual or code-mixed data because of the high degree of linguistic variability, structural complexity, and the lack of such annotated resources for Indian languages.To address the challenges, this paper suggests the design of a system called Multilingual Customer Feedback Analyzer, which will automatically classify the customer feedback, which will be in English, Hindi, and Gujarati, and map it to different categories, which may be bug reports, feature requests, positive feedback, and other categories. For this purpose, the presented system will utilize the MuRIL, which stands for Multilingual Representations for Indian Languages. It is based on the transformer model and has been specifically designed to learn contextual representations for Indian languages.To ensure scalability as well as accessibility, the system is set up with a cloud platform like Google Colab for efficient training of the models. The proposed pipeline helps to save time, automate more processes, speed up the iteration of the product, as well as enable organizations to effectively leverage insights from multilingual feedback.

1.1 Motivation

1) Manual Analysis Becomes Impractical at Scale: At present, the organizations are bombarded with a number of customer reviews, which are filtered through entities like mobile applications, websites, social media, etc. Processing these reviews is not only time-consuming but also not consistent. Further, handling languages is quite complex as it is almost impossible.
2) Regional Language Feedback Often Gets Ignored: In countries that offer services in more than one language, such as India, most users would prefer to give opinions in Hindi or Gujarati rather than English. However, most of the existing feedback analysis techniques have been tailored to work on the English language only. As such, the gained insights in regional languages go unnoticed.
3) Basic Tools Fail to Capture Meaning Correctly: Generally, keyword-based systems may not interpret user feedback well, as these systems do not understand the context. For instance, the sentence used finds the word “bug” and assumes this is negative feedback, which might not always be the scenario.
4) Language Variations Create Additional Challenges: Along with the main problems the constituency faces, language variation is a secondary problem that the constituency faces and needs to overcome. There is a lot of morphological content in both Hindi and Gujarati.
5) Turning Something Scalable and Practical: As the companies grow, customer feedback goes on increasing to a very large number. Due to the different languages present among reviews, a person is unable to go through them and sort them. A system is required by business to handle the high volume without causing disruption. A speedy automated scalable solution for processing reviews will get the team to be quick in addressing issues and improving.

1.2 Proposed Methodology

1) Data Collection: Feedback of customers is obtained through various channels including surveys, support tickets, app store reviews as well as social media. Given that the users are posting their views using various languages, the dataset will comprise English, Hindi, and Gujarati feedbacks to achieve a realistic multilingual representation.
2) Data Preprocessing: The unprocessed text feedback is processed and formatted and then it is fed to the model by Tokenization, Lemmatization, Stop Word Removal and Language Processing
3) Text Embedding using MuRIL: It is used to transform the cleaned text into contextual vectors. Contrary to conventional embeddings, MuRIL is able to capture deep contextual meaning in more than one Indian language facilitating successful cross lingual interpretation.
4)Strategies used in learning: Supervised Learning Approach: MuRIL model is trained on labeled feedback data to categorize the reviews into the following categories like bug report, positive feedback, feature request and others. This enables the model to acquire meaningful patterns and linguistic differences in different languages.
5) Unsupervised Learning Approach: Clustering techniques are constantly being combined with classification methods. They deploy clustering methods like K-Means, DBSCAN etc. so that they can find better themes in new feedbacks. Finding previously unseen issues is made simpler as it can unveil emerging patterns.
6) Model Evaluation: The standard evaluation measures help in the evaluation of the performance of the classification model. Accuracy, precision, recall, and F1-score measures are designed by people. These evaluation measures are used to ensure that the model is performing equally for all the supported languages.
7) Cloud Based Deployment: The deployment of the model can be done on the cloud environment. For example, we can run it on Google Colab

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 3302-3307

and scale its training and inference on manageable hardware. This setup will make the solution feasible and accessible for public use.

1.3 Problem Statement

Nowadays, Customer Feedback helps companies improve their products and services offering in the knowledge economy. The rise of digital platforms has resulted in a burst of users discussing their opinions about your products, experiences with customer service support and the employees' performance via mobile apps, websites, support portals, emails and social networking sites. While this feedback is of great value, it is typically unstructured, and expressed in natural language.

The problem is further complicated in multilingual countries like India where users post their views not only in English but also in Hindi, Gujarati etc. Most conventional feedback analysis systems are either designed for English only or based on simple key word-based approaches. As a result, feedback that is written in the regional languages is often ignored, misclassified or not fully utilized.

Manual processing of this type of multilingual feedback is not only time consuming, but also inconsistent and subject to human error. As the amount of data grows, the manual methods are not practical and scalable. This in turn directly affects response times and delays product improvements and, possibly, results in a negative impact on overall customer satisfaction.

Existing tools for the market often are too generic or too specific to the domain they are used in and are not able to accurately understand the context, tone and intent. Many systems have trouble separating out different forms of feedback, for example, bug reports, feature requests, critical issues or general opinions. In addition, one review may have multiple sentiments/intentions, making classification even more difficult. Traditional rule-based systems are not flexible enough to cope with such complexity.

Another shortcoming of the existing methods is that they are hard to be generalized by domain-specific vocabulary and linguistic variation in multiple languages. Moreover, many of the existing approaches do not make an explicit use of either real-time processing or cloud-based scalability, thus limiting their usefulness in such an ever evolving and fast world.

Customers portray their feelings as feedback that they express to the enterprises. Customer-created feedback is an essential requisite for business decision-making. However, a variety of feedback on such items is available in different languages on ecommerce platforms. The proposed system attempts to bridge this gap by using MuRIL based context-aware embeddings and deep learning methodologies to convert raw, unstructured feedback into structured and operational insights. Instead, a system like this would enable quicker decisions and awareness of what needs attention, along with more responsive customer interaction.

1.4 Challenges

Creating a Customer Feedback Analyzer can present several challenges, including:

1) Unstructured and Noisy Data: The customer feedback can be written in informal language, mixed scripts, slangs, emojis, grammatical errors, etc., particularly in Hindi and Gujarati which makes it difficult to be processed with high accuracy.

2) Ambiguity in Feedback: A single review can involve different intentions in the same review such as positive comments with complaints that makes the category difficult to define.

3) Limited Labeled Multilingual Data: It can take much time and be extremely costly to develop labeled feedback data in several languages for fine-tuning transformer-based model.

4) Contextual Misinterpretation: An advanced model might misinterpret the context of a message. These include messages with sarcasm, implication based meaning, or regional usage.

5) Scalability Constraints: The processing of massive multilingual feedback in real time can be very resource intensive.

6) Domain Adaptability: A model trained with feedback from one domain may not generalize well to another without some fine tuning.

7) Multilingual Complexity: The processing of operational feedback becomes more complicated due to the different languages of feedback in English, Hindi and Gujarati. In addition, high use of linguistic variability, script and contextual differences make it hard to process this feedback.

8) System Integration: Combining the settings of an analyzer, business dashboards, support systems or cloud infrastructure is technically not straightforward.

2. LITERATURE REVIEW

1) Customer Feedback Analysis Overview: Customer feedback analysis has already been extensively studied using machine learning, deep learning, and natural language processing techniques. The various models of sentiment detection, complaint classification and automated feedback interpretation have been presented by scholars for effective decision making and customer relationship management.

2) Emotion-Based Feedback Analysis: There have been numerous research on focus emotion based feedback analysis which will be machine learning model based. Techniques like Computer Vision coupled with Deep Learning approaches such as Convolutional Neural Networks (CNN) identify emotional states and understand feedback. Some frameworks have mechanisms to protect users' sensitive data and a guarantee of privacy. Despite gaining fruitful results on automated feedback understanding with such systems, their disadvantages are small datasets and high computational loads. A system such as this cannot be implemented in the real world, due to either variations in data or complexity.

3) Privacy-Saving and Distributed Learning Methods: Privacy-conscious feedback analysis too has been investigated with the distributed learning methods. In these systems, training of models is done on multiple sources of data at the same time as user privacy is maintained by using secure learning systems. Even though such methodologies increase the level of data security, they present issues to do with computation overhead, data non-homogeneity, and efficiency in communication. Also, most of these systems are characterized by difficulties in real-time implementation, and the absence of support of various modalities of input.

4) Traditional Machine Learning and Deep Learning Sentiment Analysis: Traditional Customer feedback sentiment analysis approaches have received widespread use. Comparative studies of machine learning algorithms, which include Logistic Regression, Support vector machines, random forest and Naive Bayes have proved to be effective in terms of their ability to determine positive and negative sentiment. Newer neural networks like Long short term memory (LSTM) networks have exhibited better results as they can extract sequential dependencies in text better. Nevertheless, the models usually have a problem with neutral sentiment prediction, need substantial labels, and might be ineffective in terms of inter-domain generalization.

5) Transformer-Based Feedback Classification: The developments in deep learning have further enhanced feedback classification by using Transformer-based architecture. Such models give contextual descriptions of language and have shown better performance than the traditional recurrent neural networks. Transformer-based methods even though correct are usually memory-intensive and demand domain adaptation fine-tuning. In addition to that, most of the implementations concentrate more on sentiment polarity instead of detailed feedback categorization.

6) Complaint Classification with Neural Networks: Another case that has been handled through deep networks models and word embedding approaches is consumer complaint classification. Models such as LSTM, BiLSTM, CNN and lightweight transformers showed good performance for classification. Even so, the reliability of the system is hindered by an unbalanced data set, lacking or thin contextual representation in the traditional embeddings, and incapability of real time information processing.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 3302-3307

7) NLP-Based Automated Classification Systems: Research on automated complaint classification based on the general NLP preprocessing algorithms and supervised learning algorithm is being conducted. While these methods can operate at moderate levels, they often face short or ambiguous texts, overlapping meanings, and the lack of uniformity in the quality of the data involved. This means that the models need to capture better contextual meaning as well as linguistic variability.

8) Shortcomings of Current Solutions: Despite the prominence of advances in automated feedback analysis prior to this, there are still a number of limitations seen in the system. Most existing frameworks also focus on the single language corpus, i.e. predominantly the English corpus and do not allow for multilingual set-ups. Thereby restricting the problem of exploiting the analysis of input systems in the areas of high linguistic diversity. Most of the work deals with the classification of mood rather than fine grained classification of inputs into actionable organizational relevant categories.

9) Multilingual Transformer Models Emerging: Recent studies have started to explore multilingual language models based on transformers that are believed to be useful to cope up with these deficiencies. Models capable of semantic induction may allow for better reasoning to classify radio varied textual inputs. In light of such developments, the present study is on way to create Multilingual Customer Feedback Classification through Context-definite representations created in a linguistically diverse context. The solution submitted here offers flexibility and feasibility for this task in the real world. Variety of languages and categorization of feedback will be possible on the basis of categories of feedback.

3. METHODOLOGY

3.1 System Architecture

The proposed MCFA uses MuRIL (Mulitlingual Representations for Indian Languages) as core transformer encoder. MCFA, or Multilingual Customer Feedback Analyzer, is a modular NLP pipeline designed to customer feedback in English, Hindi, and Gujarati.

The architecture includes the following main components.

1) Input Section: The CSA file or web form accepts unprocessed text of customer feedback. Unicode Text for Hindi (Devanagari code) and Gujarati script supported.
2) Data Preprocessing: This includes various steps that help ensure the classifier can classify the reviews in a proper manner. This includes steps like, Letter case conversion in text normalization and clean up unwanted spaces and special symbols, Unicode standardization, Detection of language for analysis optional.
3) MuRIL Tokenizer: It first employs the WordPiece tokenization, it Turns text into input IDs attention masks and token type IDs, this then helps in managing inputs with different scripts.
4) Encoder for Transformer (MuRIL): A transformer encoder with 12 layers, it makes use of self-attention mechanism for the contextualized embedding representations. It yields a sentence representation context with a CLS token embedding.
5) Classification head: This is an all-purpose fully connected dense layer, it Makes use of softmax Function activation, after which it gives probability distribution over four classes such as, Bug Report, Feature Request, Positive Feedback and other.
6) Module for Evaluation and Prediction: This module provides estimated labels, calculates key metrics, including things like the Accuracy and Precision, and finally saves outcomes for review.

3.2 Data Description

The customer feedback multilingual dataset that was used for training this model contained four predefined classes. such as:

Bug report, Feature request, Positive Feedback, Other.

1) Language Distribution: The dataset includes reviews of various languages such as, English feedback, Reviews in Hindi, Gujarati Reviews (Gujarati Lettering). Meticulous curation was undertaken of each language dataset to make sure the dataset has accuracy of the script, distinct samples (no repeats), uniform spread across classes.
2) Size of Dataset: To achieve consistency across experiments, we made sure to maintain the same samples per class within the constraints of the data.

3) Data Features: The dataset included data that covered a wide range of features such as. Brief and detailed feedback, use a friendly tone, mix of Languages, complaint patterns of real customers.

3.3 Data preprocessing

In order to achieve the best transformer performance, data preprocessing was necessary.

1) Cleaning: This step ensures the model got a clean, well- balanced dataset, this was done by making use of the following, Duplicates removed, removed null entries, removed extra spaces.
2) Unicode Management: Encoding (UTF-8), also etained scripts of Hindi and Gujarati.
3) Encoding Labels: The labels were assigned as,

Label	Encoding Value
bug_report	0
feature_request	1
positive_feedback	2
other	3

4) Train-Test Split: The model was trained on 80% of the overall dataset, and for the evaluation the remaining 20% of the dataset was used (test set), this was done to strategically divide to maintain class balance.

3.4 Model picking: MuRIL

MuRIL is a transformer-based Indian language model specifically designed for Indian languages. This is the chosen model for this work.

1) Reason for choosing MuRIL: It is trained on over 17 Indian languages; it also takes care of transliterated and code-mixed text. Indic languages perform better on multilingual BERT models than their generic counterparts. Also, as it is a transformer-based model, the embeddings are context aware and make use of self-attention.
2) Fine-tuning Strategy: For fine-tuning MuRIL to our use-case, we made use of several steps to ensure a robust model. This was done by, Loading Pretrained MuRIL weights, then a layer for classification was added, and then making use of a detailed dataset to fine-tune the model. Also Utilizing cross-entropy loss function., for the optimization AdamW optimizer with learning rate scheduling was used.

3.5 Training Configuration

The training setup made use of the following, Optimizer: AdamW, Loss Function: Cross-Entropy loss, Batch size: 16-32, Learning Rate: 2e-5 to 5e-5, Epochs: 3–5. Along with this, early stopping was implemented to prevent overfitting.

3.6 Evaluation Metrics

To assess model performance, metrics such as. Accuracy. Precision. Recall, F1-Score (Macro and Weighted), Confusion Matrix were used. The Macro-F1 was given more priority due to the multi-class classification of our project and the chance of a potential class imbalance.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 3302-3307

4. RESULT

4.1 Transformer Based Models

The Transformer architecture is being increasingly used for single and multi-sentence processing tasks. This is due to the architecture utilizing attention mechanisms to capture context. Thus, the research assesses three transformers-based model MuRIL, mBERT and Zero-Shot Roberta.

1) Muril: The MuRIL model yielded the best result among all the models tested in the paper. MuRIL is a model specific to Indian languages which is pretrained on other Indic languages on large multilingual corpora. The accuracy was achieved on the mixed language dataset for the model making it capable of understanding context on the whole. The model attained significant precision and recall for all classes. The model's ability to generalize is strong. The performance of the model for bug reports and feature requests is particularly promising as these are the most important categories.

2) Multilingual-BERT (mBERT): The mBERT model was evaluated as the baseline multilingual transformer model. The mBERT model supports multiple languages but it is not that specifically trained for Indic languages. The model achieved an overall accuracy of 81%, which is a lower figure compared to MuRIL. The model misclassified some of the categories correctly. For instance, positive feedback and other classes weren't detected properly, hence the low recall. MuRIL is more suited for Indic languages as it has a domain-specific multilingual model for Indic languages such as hindi and gujarati in the dataset.

3) Zero-Shot RoBERTa: The RoBERTa model used a zero-shot classification approach. In other words, the training data was not fine-tuned on the model. The model did not get any labeled training data. It was provided label descriptions with succinct definitions for each category. The model was forced to depend on training data, which caused the outcome had much lower performance than expected. The zero-shot RoBERTa model reached an accuracy of 35% which is the least one as compared to the other models that were tested. In addition, the model was largely confused in identifying bug reports and feature requests.

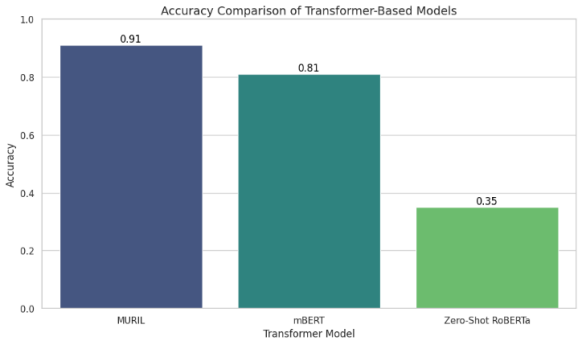

Fig. 5.1 Accuracy Comparison of Transformer Models

Model	Accuracy	Precision	Recall	F-1 Score
Muril	0.91	0.91	0.91	0.90
mBERT	0.81	0.83	0.80	0.80
RoBERTa Large (zero-shot)	0.35	0.48	0.39	0.31

Table 5.1: Comparing Muril, mBERT, RoBERTa

4.2 Deep Learning Models: Architectures for deep learning were evaluated. Their performance was evaluated for capturing semantic patterns and contextual dependencies in texts. For these architectures such as CNN, BiLSTM were used.

1) Convolutional Neural Network (CNN): CNN model hit 64% accuracy. CNNs are adept at detecting local features like n-grams and short phrases utilizing convoluted filters. CNNs have shown tremendous performance in several text classification tasks. Despite their solid performance compared to other methods, they may be impaired from capturing long-distance dependencies.

2) Bidirectional Long Short-Term Memory (BiLSTM): An accuracy of 53% was achieved by this model. Methods that were based on transformers tried to capture global contextual dependencies but this model showed weaker performance in comparison even though BiLSTM takes a bidirectional approach when trying to understand the context.

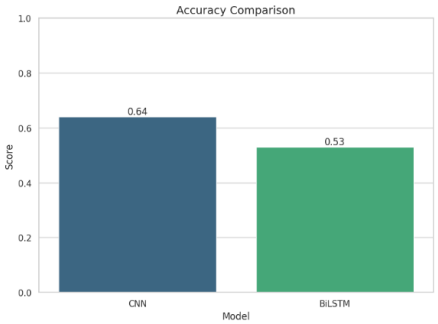

Fig. 5.2 Accuracy Comparison of Deep Learning Models

Model	*Accuracy*	*Precision*	*Recall*	*F1-Score*
CNN	*0.64*	*0.65*	*0.63*	*0.62*
BiLSTM	*0.53*	*0.65*	*0.51*	*0.48*

Table 5.2: Comparing CNN, BiLSTM

4.3 Classical Machine Learning Models

Traditional machine learning models with TF-IDF feature extraction was applied as a baseline technique using machine learning models commonly employed in the literature such as Support vector machine and Logistic Regression.

1) TF-IDF + Support Vector Machine: The TF-IDF accompanied by the SVM produced an accuracy of 69%. Support Vector Machines are extremely efficient in the case of a high dimensional feature space.

Although we achieved decent output through the model, it lacked the context that deep learning models could provide.

2) TF-IDF + Logistic Regression: The model yielded 70%, demonstrating comparable performance to the SVM model. Logistic regression is often used as a baseline in text classification problems due to its simplicity and efficiency.

An executive summary is a document that summarizes a longer report or proposal, or a group of related reports.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 3302-3307

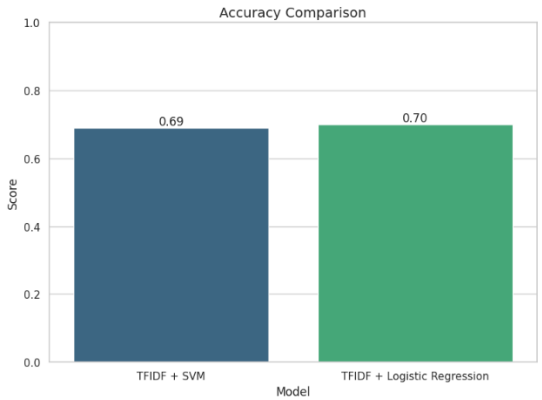

Fig. 5.3 Accuracy Comparison of classical machine learning models

Model	*Accuracy*	*Precision*	*Recall*	*F1-Score*
TFIDF + SVM	0.69	0.72	0.69	0.69
TFIDF+ Logistic Regression	0.70	0.75	0.69	0.69

Table 5.3: Comparing TFIDF SVM, Logistic Regression

4.4 Overall Model Comparison

1) Trends in Performance Throughout Model Categories: According to what we got from the experimental results, the models with complex architecture and capability can represent the context information more highly.The performance of the Transformer-based models is the best because they can capture long-range dependencies using attention. MuRIL model outperformed other models by a wide margin scoring accuracy of 91%, thereby demonstrating the power of Language specific Pretraining for Indic languages.

The increased depth of the transformer aids MuRIL in surpassing mBERT. The mBERT is the second-best performing language model with accuracy of 80.5 %. It suggests that even though it isn't made specifically for Indic languages it still can perform well.

On the other hand, there are the simpler or traditional machine learning models, i.e., TF-IDF + SVM and TF-IDF + Logistic Regression. All of these models, based on their respective, achieve an accuracy 69–70%. They use only the statistical word-frequency, which is precisely the reason why this result was achieved. CNN's performance was 64 % but BiLSTM was lower in performance with 53 % accuracy. This means that a sequential model without an attention mechanism may be unable to represent multilingual semantic patterns.

2) Effect of Contextual Representations: The capacity to generate contextual word representations considerably impacts model performance. Traditional machine learning models cannot differentiate between feedback and it's contrary when both have the same words. These models use statistical features which do not capture semantic meaning or context.Deep learning models, including CNN and BiLSTM, can learn from sequences' patterns. On the other hand, deep learning models rely on contextual details less than transformers do. Transformer-based models don't have this limitation. Through a self-attention device, they capture the contextual relationship between the words and phrases. This becomes especially beneficial when working with multi-lingual datasets where the language structures are quite different.

3) Efficiency of Indic Script Language Models: The most remarkable finding from this experiment is that MuRIL outperforms the other transformer models adopted for benchmarking. MuRIL was specifically pretrained on Hindi and other languages from multilingual data sets. This is why an Indian language model pretrained was used to fine tune for our problem statement. it also accounts for Indian linguistic phenomena better than other multilingual ones that are not specifically trained on a large volume of Indian languages.

4) Limitations of Zero-Shot Learning: Of all the designs, the Zero-shot RoBERTa exhibited the lowest performance as it achieved 35%. The previous models were trained for our specific task. But the zero-shot model isn't. It examines the input text and label descriptions, and the model predicts based on the similarity of those two. Although zero-shot learning can be used as a suitable alternative for situations with no labeled data, our results indicate it does not work well at specialized classification tasks like customer feedback classification. The model fails to correctly identify the bug reports in particular due to which the precision and recall scores are extremely low.

Fig. 5.4 Accuracy Comparison of All Evaluated Models

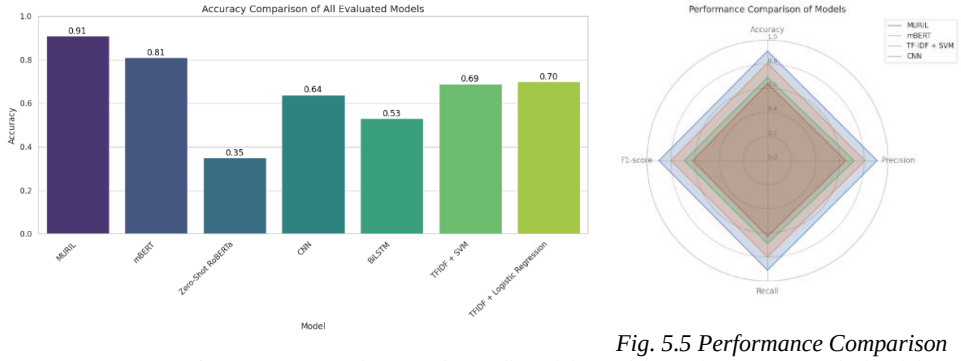

Fig. 5.5 Performance Comparison

4.5 Class-Wise Analysis of MuRIL: To get a better understanding of the model's performance, we calculated the precision, recall, f1-score and the support of both the models, that is the fine-tuned Muril. The below table shows the results that were achieved by the Fine-Tuned Muril model. The testing dataset consisted of reviews of various languages such as English, Hindi, Gujarati which were not seen by the model during the training phase to ensure a fair evaluation.

Label	Precision	Recall	F1-Score	Support
bug_report	0.89	0.98	0.93	50
feature_request	0.92	0.94	0.93	51
positive_feedback	0.97	0.80	0.88	46
other	0.85	0.89	0.87	38

Table 5.4: Muril (Fine-Tuned) Class-wise results

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1s, 2026, Pages: 3302-3307

Key observations from the classification reports:

1) Bug Report: MuRIL model in fine-tuning fashion resonates very well (F1 = 0.93, Recall = 0.98) capturing almost all bug related feedback.
2) Feature Request: MuRIL gave a strong and balanced performance. (F1 = 0.93, Recall = 0.94).
3) Positive Feedback: MuRIL showed that it understood the context well (F1 = 0.88, Recall = 0.80).
4) Other: MuRIL did well (F1 = 0.87), but its predictions were more evenly spread across categories, which shows that it had better discrimination ability.

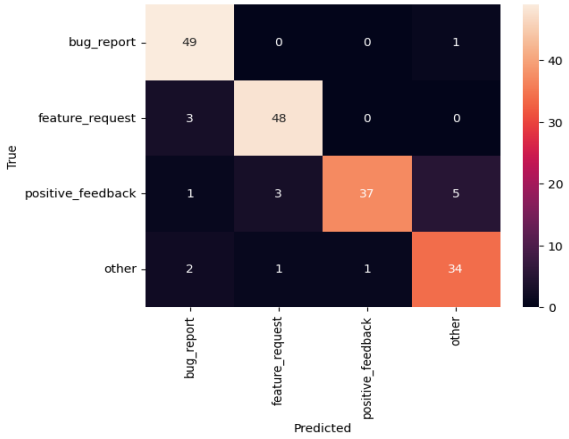

Fig. 5.6 Confusion Matrix of MuRIL

5. CONCLUSION

This study presents a detailed analysis of the various machine learning and deep learning techniques for multi-lingual customer feedback classification in this paper. The objective of this research was the automatic classification of the feedback (in the customer message) given in multi-languages into one of the predefined classes such as bug report, positive feedback, feature request etc. We evaluated the performance of a wide variety of model architectures including traditional machine learning models and the leading-edge deep learning and transformer-based language models.The experimental outcome shows that transformer-based architectures are better than traditional machine learning models for multilingual text classification. MuRIL yielded the best performance across all models, with a score of 91% accuracy. This model has been trained on Indic languages and hence it can manage and capture good contextual information on Hindi, Gujarati and other regional languages due to the pre-training. The multilingual BERT gave a very strong performance with nearly 80.5% accuracy.On the other hand, the model based on TF-IDF with different linear classifiers achieves quite moderate success. The accuracy of logistic regression classifier is 69.6% and support vector classifier is 70.0%. While these models do offer statistical representations of text that relate to context, they do not actually have any true semantic meaning. In a similar vein, BiLSTM and CNN networks also record moderate performance levels of detection. These models can address the limitations of TF-IDF based models but cannot match the contextual understanding of transformer-based models like MuRIL, mBERT. If we use the RoBERTa model in the zero-shot setting (for the same set of classes.) The results do make a case for contextual word representations being useful for multilingual NLP tasks. In particular, a transformer model trained on multilingual data gives a significant boost to classification accuracy and robustness when compared to previous models. Future work could extend this research to use bigger datasets to explore the potential of other multilingual transformer models.

6. ACKNOWLEDGEMENT

We express our heartfelt thanks to our honorable Vice Chancellor Dr. C. MUTHAMIZHCHELVAN, for being the beacon in all our endeavors. We would like to express my warmth of gratitude to our Registrar Dr. S. Ponnusamy, for his encouragement. We express our profound gratitude to our Dean (College of Engineering and Technology) Dr. T. V.Gopal, for bringing out novelty in all executions. We would like to express my heartfelt thanks to Chairperson, School of Computing Dr. Revathi Venkataraman, for imparting confidence to complete my course project. We are highly thankful to our Course project Faculty **Dr.A. Pandian**, Professor, Department of Computing Technologies, for his assistance, timely suggestion and guidance throughout the duration of this course project. We extend my gratitude to our HoD Dr. M Pushpalatha, Professor and Head, Department of Computing Technologies and my Departmental colleagues for their Support. Finally, we thank our parents and friends near and dear ones who directly and indirectly contributed to the successful completion of our project. Above all, I thank the almighty for showering his blessings on me to complete my Course project.

References:

1. S. Khanuja, D. Khandelwal, A. Singh, N. Kunchukuttan, and P. Goyal, "MuRIL: Multilingual Representations for Indian Languages," *arXiv preprint arXiv:2103.10730*, 2021.
2. Y. Liu, M. Ott, N. Goyal, et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv preprint arXiv:1907.11692*, 2019.
3. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. NAACL-HLT*, Minneapolis, MN, USA, 2019, pp. 4171–4186.
4. A. Vaswani et al., "Attention Is All You Need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.
5. T. Wolf et al., "Transformers: State-of-the-Art Natural Language Processing," in *Proc. EMNLP: System Demonstrations*, 2020, pp. 38–45.
6. A. Conneau et al., "Unsupervised Cross-lingual Representation Learning at Scale," in *Proc. ACL*, 2020.
7. R. Pires, E. Schlinger, and D. Garrette, "How Multilingual is Multilingual BERT?" in *Proc. ACL*, 2019, pp. 4996–5001.
8. T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *arXiv preprint arXiv:1301.3781*, 2013.
9. S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
10. F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
11. Y. Kim, "Convolutional Neural Networks for Sentence Classification," in *Proc. EMNLP*, 2014, pp. 1746–1751.