Sentiment Analysis of Malayalam and Manglish Social Media Text: A Custom Pipeline for Mixed-Script Language Processing

A Project Report Submitted

in Partial Fulfilment of the Requirements

for the Degree of

MASTERS OF SCIENCE

in

COMPUTER SCIENCE (DATA ANALYTICS)

by

PRAVEEN MK

Register No.230011020929

to

PG DEPARTMENT OF COMPUTER SCIENCE

SREE SANKARA VIDYAPEETOM COLLEGE

VALAYANCHIRANGARA - 683 556, INDIA

02 December 2024

DECLARATION

I, **PRAVEEN MK, Roll.No:230011020929,**, hereby declare that, this report entitled **"Sentiment Analysis of Malayalam and Manglish Social Media Text: A Custom Pipeline for Mixed-Script Language Processing"** submitted to Mahatma Gandhi University, Kottayam towards partial requirement of **Masters of Science** in **Computer Science (Data Analytics)** is an original work carried out by me under the supervision of **Dr.MANUSANKAR C** and has not formed the basis for the award of any degree or diploma, in this or any other institution or university. I have sincerely tried to uphold academic ethics and honesty. Whenever an external information, statement, or result is used, that have been duly acknowledged and cited.

Valayanchirangara - 683 556 **PRAVEEN MK**

02 December 2024

CERTIFICATE

This is to certify that the work contained in this project report entitled **"Sentiment Analysis of Malayalam and Manglish Social MediaText: A Custom Pipeline for Mixed-Script Language Processing"** submitted by **PRAVEEN MK** (**Roll.No:230011020929**) to Mahatma Gandhi University, Kottayam towards partial requirement of **Masters of Science** in **Computer Science (Data Analytics)** has been carried out by his under my supervision and that it has not been submitted elsewhere for the award of any degree.

Dr. Manusankar C

02 December 2024 **Project Supervisor and HOD**

ACKNOWLEDGEMENT

It has been said that gratitude is the memory of heart. Hence, I take this opportunity to express my gratitude to all those, whose contribution in this project can't be forgotten. First and foremost, I thank the Almighty for his showers of blessings on this work and for endowing me the will to complete it on time.

I am grateful to **Dr. K M Sudhakaran** , Principal, SSV College, Valayanchirangara, for his able leadership and guidance in all the official matters regarding this work. I would like to thank, **Dr. Manusankar C** , my guide and Head of Department of Computer Science, for his inspiring and commendable support throughout the course and especially for this work. I also express my sincere thanks to our PG Program coordinator

Ambili M S, for her expert guidance,constant encouragement and valuable suggestions.

I profusely thank other Professors in the Department and all other staffs of SSV, for their guidance and inspirations throughout my course of study. No words can express my humble gratitude to my beloved parents who have been guiding me in all walks of journey. My thanks and appreciations also go to my friends and people who have willingly helped me out with their abilities.

Valayanchirangara - 683 556 **PRAVEEN MK**

02 December 2024

ABSTRACT

This study presents a robust pipeline for sentiment analysis of Malayalam and Manglish (Romanized Malayalam) social media text, addressing the unique challenges of mixed-script and low-resource language processing. It focuses on an end-to-end solution for processing user-generated content from platforms like Instagram and YouTube, where Malayalam speakers often switch between native script and Romanized text. The pipeline incorporates language detection, transliteration, translation, and sentiment classification modules, achieving an overall accuracy of 93%. Higher accuracy was observed for native Malayalam texts (95%) compared to Manglish texts (90%), highlighting the complexities of mixed-script processing. Key challenges addressed include transliteration errors, informal language patterns, and contextual ambiguity. The findings reveal distinct patterns in sentiment expression across topics, with Manglish predominantly used for informal and emotional content. This work contributes to multilingual sentiment analysis and low-resource language processing, offering practical implications for social media monitoring and public opinion analysis in Kerala.

Index

page no.

Introduction

1.1 Background

The advent of social media has revolutionized the way people communicate, share opinions, and express sentiments. Platforms like Instagram, YouTube, and Facebook have become rich sources of user-generated content, reflecting diverse perspectives on topics ranging from politics and cinema to sports and social issues. In Kerala, where Malayalam is the primary language, social media users often employ a mix of native Malayalam script and Manglish (Malayalam written in Roman script) to express their thoughts. This unique linguistic phenomenon is driven by the convenience of typing in Roman script on mobile devices and the informal nature of online communication.

However, this multilingual and mixed-script environment poses significant challenges for natural language processing (NLP) tasks, particularly sentiment analysis. Sentiment analysis, which involves determining the emotional tone of a text (positive, negative, or neutral), is a critical tool for understanding public opinion and trends. While sentiment analysis tools for high-resource languages like English have achieved remarkable accuracy, their performance in low-resource languages like Malayalam remains limited. The informal, unstructured, and noisy nature of social media text further complicates the task, necessitating the development of specialized tools and techniques.

1.2 Problem Statement

The primary challenge in sentiment analysis for Malayalam and Manglish texts lies in the linguistic and structural complexities of the data. Unlike high-resource languages, Malayalam lacks extensive annotated datasets and pre-trained models, making it difficult to train accurate sentiment analysis systems. Additionally, Manglish introduces further complications due to its informal nature and lack of standardized spelling. For example, the word "adipoli" (awesome) can appear in multiple forms, such as "adipolii," "adipoly," or "adipolli," depending on the user's preference. This diversity in spelling and phonetic representation makes transliteration and normalization challenging.

Machine translation tools, often used to bridge low-resource languages with high-resource models, also struggle with Malayalam. Colloquial expressions, slang, and cultural references are frequently mistranslated, leading to errors in sentiment classification. Furthermore, the presence of sarcasm, code-mixing (e.g., combining Malayalam and English in the same sentence), and emojis adds another layer of complexity. For instance, a sarcastic comment like

"Adipoli aanu, nalla pani" (sarcastically meaning "It's awesome, great job") may be misclassified as positive due to the presence of positive words like "adipoli."
These challenges highlight the need for a robust and scalable sentiment analysis pipeline tailored to the unique linguistic and cultural context of Kerala. Without addressing these issues, the potential of sentiment analysis to provide valuable insights into public opinion and trends remains untapped.

1.3 Scope of the Study

This study aims to address the challenges of sentiment analysis for Malayalam and Manglish texts by developing a comprehensive and scalable solution. The scope of the study includes the following key aspects:

1.Pipeline Development:
The study focuses on designing and implementing a robust sentiment analysis pipeline capable of processing both Malayalam and Manglish texts. The pipeline will include:

Language Detection: Accurately identifying whether the input text is in Malayalam, Manglish, or a combination of both.
Transliteration: Converting Manglish text into native Malayalam script to enable consistent processing.
Translation: Translating Malayalam text into English to leverage existing sentiment analysis tools optimized for high-resource languages.
Sentiment Classification: Classifying the translated text into positive, negative, or neutral categories using machine learning models.

2.Improving Sentiment Detection Accuracy:
The study aims to enhance the accuracy of sentiment detection by addressing key challenges such as:
Handling informal language, slang, and colloquial expressions.
Reducing errors in transliteration and translation.
Incorporating mechanisms to detect sarcasm and context-dependent sentiments.

3.Comparative Analysis:
A comparative analysis will be conducted to evaluate the performance of the sentiment analysis pipeline for Malayalam and Manglish texts. This analysis will provide insights into the differences in sentiment expression between the two scripts and highlight areas for further improvement.

4.Insights into Linguistic and Cultural Nuances:
By analyzing the sentiment trends across various topics (e.g., politics, cinema, sports, and social issues), the study seeks to uncover the linguistic and cultural nuances of Malayalam and Manglish texts. These insights will contribute to the broader field of multilingual sentiment analysis, particularly for low-resource and mixed-script languages.

5.Real-World Applications:
The findings of this study have practical implications for industries and organizations seeking to understand public opinion in Kerala. Applications include social media monitoring, market research, and public sentiment analysis for political campaigns, movie releases, and social initiatives.

By addressing the unique challenges of Malayalam and Manglish sentiment analysis, this study aims to bridge the gap in NLP research for low-resource and mixed-script languages. The proposed pipeline and insights will serve as a foundation for future research and development in this domain, enabling more accurate and reliable sentiment analysis in real-world scenarios.

Literature Review

2.1 Overview of Sentiment Analysis

Sentiment analysis, also known as opinion mining, is a subfield of natural language processing (NLP) that focuses on identifying and categorizing the emotional tone of a text. It determines whether a given piece of text expresses a positive, negative, or neutral sentiment. Sentiment analysis has gained significant attention due to its wide range of applications, including social media monitoring, customer feedback analysis, market research, and public opinion tracking. Traditional sentiment analysis techniques can be broadly categorized into three approaches:

1.Lexicon-Based Methods: These rely on predefined sentiment lexicons, which are dictionaries of words associated with positive or negative sentiments. For example, words like "good" and "excellent" are classified as positive, while "bad" and "terrible" are classified as negative. While lexicon-based methods are simple and interpretable, they often fail to capture the context or nuances of a sentence, such as sarcasm or idiomatic expressions.

2.Machine Learning-Based Methods: These involve training supervised machine learning models, such as Support Vector Machines (SVM) or Naive Bayes, on labeled datasets. These models learn patterns in the data to classify sentiments. However, their performance heavily depends on the quality and size of the training data.

3.Deep Learning-Based Methods: Recent advancements in NLP have led to the adoption of deep learning models, such as Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and transformer-based models like BERT (Bidirectional Encoder Representations from Transformers). These models excel at capturing contextual and semantic information, making them highly effective for sentiment analysis.

While sentiment analysis has achieved high accuracy for high-resource languages like English, its application to low-resource languages, such as Malayalam, remains a significant challenge.

The informal and unstructured nature of social media text, combined with the lack of annotated datasets and pre-trained models, further complicates the task. Additionally, the rise of mixed-script text, such as Manglish, introduces unique challenges that traditional sentiment analysis methods are not equipped to handle.

2.2 Challenges in Multilingual and Mixed-Script Sentiment Analysis

Sentiment analysis in multilingual and mixed-script contexts presents several challenges, particularly for low-resource languages like Malayalam and its Romanized counterpart, Manglish. These challenges arise from linguistic, cultural, and technological factors, as outlined below:

1.Linguistic Complexity:
Malayalam is a morphologically rich language with complex grammar and syntax. Words can have multiple forms depending on tense, gender, and context, making it difficult to process using traditional NLP techniques.
Manglish, on the other hand, lacks standardized spelling rules. For example, the word "adipoli" (awesome) can appear as "adipoly," "adipolii," or "adipolli," depending on the user's preference. This diversity in spelling complicates transliteration and normalization.

2.Informal and Noisy Text:
Social media text is often informal, unstructured, and noisy, with frequent use of slang, abbreviations, emojis, and code-mixing (e.g., combining Malayalam and English in the same sentence). For example, a comment like "ഈ movie adipoli ആണ് 😍🔥" (This movie is awesome 😍🔥) contains a mix of Malayalam, Manglish, and emojis, making it challenging to process.
Sarcasm and irony are common in social media text, further complicating sentiment detection. For instance, a sarcastic comment like "Adipoli aanu, nalla pani" (sarcastically meaning "It's awesome, great job") may be misclassified as positive due to the presence of positive words like "adipoli."

3.Low-Resource Language Challenges:

Malayalam is a low-resource language with limited availability of annotated datasets and pre-trained models for sentiment analysis. This lack of resources makes it difficult to train machine learning models or fine-tune deep learning models like BERT.
Existing machine translation tools, such as Google Translate, often struggle with colloquial Malayalam expressions and slang, leading to inaccuracies in translation and sentiment classification.

4.Mixed-Script Text:
Manglish, or Malayalam written in Roman script, introduces unique challenges due to its informal nature and lack of standardization. Transliteration errors can propagate through the sentiment analysis pipeline, affecting the accuracy of subsequent steps like translation and classification.
Mixed-script text often includes code-mixing, where users switch between Malayalam and English within the same sentence. For example, "ഈ movie adipoli ആണ്" (This movie is awesome) combines Malayalam and Manglish, requiring specialized preprocessing techniques.

5.Cultural and Contextual Nuances:
Sentiment expression in Malayalam and Manglish often includes cultural references, idiomatic expressions, and slang that are difficult to interpret without domain knowledge. For example, the word "pani" can mean "work" in a neutral context but "trouble" in a negative context. Emojis and symbols are frequently used to convey sentiment but are not directly processed by traditional sentiment analysis models.

These challenges highlight the need for a robust and scalable sentiment analysis pipeline tailored to the unique linguistic and cultural context of Malayalam and Manglish. Addressing these issues requires a combination of advanced preprocessing techniques, transliteration, and machine learning models.

2.3 Existing Approaches for Malayalam and Manglish Texts

Several studies and approaches have attempted to address the challenges of sentiment analysis for Malayalam and Manglish texts. However, most existing methods are limited in their scope and effectiveness due to the low-resource nature of the language and the complexities of mixed-script text. Below is an overview of the existing approaches:

1.Lexicon-Based Approaches:
Some studies have developed sentiment lexicons for Malayalam, mapping words to their corresponding sentiment scores. For example, words like "നല്ല" (good) and "മികച്ചത്" (excellent) are classified as positive, while "മോശം" (bad) and "വിലാസം" (terrible) are classified as negative.
While lexicon-based methods are simple and interpretable, they often fail to capture the context or nuances of a sentence, such as sarcasm or idiomatic expressions. Additionally, they are not well-suited for handling Manglish text due to its lack of standardization.

2.Machine Learning-Based Approaches:
Supervised machine learning models, such as Naive Bayes and Support Vector Machines (SVM), have been used for sentiment analysis of Malayalam text. These models require labeled datasets for training, which are scarce for Malayalam.
Some studies have attempted to create labeled datasets by manually annotating social media comments in Malayalam. However, the size and diversity of these datasets are often limited, reducing the generalizability of the models.

3.Deep Learning-Based Approaches:
Recent advancements in NLP have led to the adoption of deep learning models, such as LSTMs and transformer-based models like BERT, for sentiment analysis. For example, multilingual BERT (mBERT) has been used to process Malayalam text by leveraging its ability to handle multiple languages.
However, these models often require fine-tuning on domain-specific data, which is challenging due to the lack of annotated datasets for Malayalam and Manglish.

4.Transliteration and Translation-Based Approaches:
Some studies have proposed transliteration modules to convert Manglish text into native Malayalam script, enabling consistent processing. For example, a Manglish comment like "adipoli aanu" can be transliterated to "അടിപൊളി ആണ്" for further analysis.
Machine translation tools, such as Google Translate, have been used to translate Malayalam text into English, allowing the use of sentiment analysis models optimized for English. However, these tools often struggle with colloquial expressions and slang, leading to inaccuracies.

5.Hybrid Approaches:
Hybrid approaches combine multiple techniques, such as transliteration, translation, and machine learning, to address the challenges of sentiment analysis for Malayalam and Manglish texts. For example, a pipeline may include:

Language detection to identify whether the input text is in Malayalam, Manglish, or a combination of both.
Transliteration to convert Manglish text into native Malayalam script.
Translation to convert Malayalam text into English for sentiment classification.

While hybrid approaches show promise, they require significant effort to develop and optimize, particularly for low-resource languages like Malayalam.

Research Objectives

The primary goal of this research is to address the challenges of sentiment analysis for Malayalam and Manglish texts, particularly in the context of social media data. By focusing on the unique linguistic and cultural characteristics of Kerala, this study aims to develop a robust and scalable solution that improves sentiment detection accuracy and provides valuable insights into public opinion trends. The following objectives outline the specific aims of the study:

3.1 To Develop a Robust Pipeline for Sentiment Analysis of Malayalam and Manglish Texts

The first objective of this study is to design and implement a comprehensive sentiment analysis pipeline capable of processing both Malayalam and Manglish texts. Social media platforms in Kerala often feature a mix of native Malayalam script and Manglish (Malayalam written in Roman script), creating a unique linguistic environment that traditional sentiment analysis tools are not equipped to handle. This objective focuses on creating a step-by-step framework that addresses the complexities of mixed-script and multilingual data. The pipeline will include the following key components:

1.Language Detection:
The pipeline must accurately identify whether the input text is in Malayalam, Manglish, or a combination of both.
Language detection is critical for determining the appropriate preprocessing steps, such as transliteration for Manglish or direct processing for native Malayalam t**ext.**

2.Transliteration:
For Manglish text, a transliteration module will be developed to convert Roman-script Malayalam into native Malayalam script.
This step ensures that Manglish text can be processed alongside native Malayalam text, enabling consistent analysis.

The transliteration module must handle diverse spelling variations in Manglish, such as "adipoli," "adipolii," and "adipoly," which all represent the same word ("awesome").

3.Translation:
To leverage existing sentiment analysis tools optimized for high-resource languages, the pipeline will include a translation module to convert Malayalam text into English.
Machine translation tools, such as Google Translate, will be evaluated for their effectiveness in translating colloquial and informal Malayalam expressions.
Custom translation models may be developed to address the limitations of existing tools.

4.Sentiment Classification:
The pipeline will perform sentiment analysis on the translated English text to classify it into positive, negative, or neutral categories.
Sentiment classification will leverage existing libraries like TextBlob or advanced machine learning models, such as transformer-based models (e.g., BERT).
The classification module must account for the informal and noisy nature of social media text, including slang, abbreviations, and emojis.

5.Error Handling and Optimization:

Mechanisms will be incorporated to handle errors in transliteration, translation, and sentiment classification.
The pipeline must be robust and reliable, capable of processing real-world social media data with minimal loss of accuracy.

By developing this pipeline, the study aims to provide a practical solution for sentiment analysis in low-resource and mixed-script contexts. The pipeline will be designed to handle the informal, unstructured, and noisy nature of social media text, making it adaptable to real-world scenarios.

3.2 To Improve Sentiment Detection Accuracy for Mixed-Script and Native Language Content

The second objective of this study is to enhance the accuracy of sentiment detection for both mixed-script (Manglish) and native-language (Malayalam) content. Traditional sentiment analysis tools often fail to achieve high accuracy in these contexts due to the following reasons:

1.Informal Language: Social media text frequently includes slang, abbreviations, and colloquial expressions that are not recognized by standard sentiment analysis models.
2.Mixed-Script Text: Manglish text, with its lack of standardized spelling and informal grammar, poses significant challenges for transliteration and normalization.
3.Translation Limitations: Machine translation tools often struggle with colloquial Malayalam expressions, leading to errors in sentiment classification.
4.Sarcasm and Contextual Ambiguity: Sarcastic comments and context-dependent sentiments are difficult to detect without advanced contextual understanding.

To address these challenges, the study will focus on the following strategies to improve sentiment detection accuracy:

1.Handling Informal and Noisy Text:

The pipeline will include preprocessing techniques to clean and normalize social media text.
A comprehensive slang dictionary will be developed to recognize and interpret informal expressions commonly used in Malayalam and Manglish.
Emojis and symbols, which often carry sentiment, will be incorporated into the sentiment analysis process.

2.Improving Transliteration Accuracy:

The transliteration module will be optimized to handle diverse Manglish spelling variations.
Machine learning models trained on Manglish-Malayalam parallel datasets will be used to improve transliteration accuracy.
Contextual rules will be implemented to resolve ambiguities in transliteration.

3.Enhancing Translation Quality:

Custom translation models will be developed to address the limitations of existing machine translation tools.
These models will be fine-tuned on Malayalam social media text to improve their ability to handle colloquial expressions and slang.

A hybrid approach combining rule-based translation for slang and machine translation for formal text will be explored.

4.Incorporating Advanced Sentiment Analysis Models:

Transformer-based models, such as BERT and its multilingual variants (e.g., mBERT, XLM-R), will be fine-tuned for sentiment analysis of Malayalam and Manglish texts.
These models will be trained on annotated datasets to improve their ability to detect sarcasm, context-dependent sentiments, and cultural nuances.
Comparative evaluations will be conducted to identify the most effective model for each type of text (Malayalam vs. Manglish).

5.Error Analysis and Iterative Improvement:

The pipeline will undergo iterative testing and refinement based on error analysis.
Misclassified examples will be analyzed to identify patterns and improve the pipeline's performance.
Feedback loops will be implemented to continuously optimize the transliteration, translation, and classification modules.

By improving sentiment detection accuracy, this study aims to bridge the gap in sentiment analysis for low-resource and mixed-script languages. The enhanced pipeline will provide more reliable insights into public opinion trends, enabling its application in various domains, such as social media monitoring, market research, and public sentiment analysis.

Contributions

This study makes significant contributions to the field of natural language processing (NLP), particularly in the domain of multilingual and mixed-script sentiment analysis for low-resource languages. By addressing the unique challenges of Malayalam and Manglish texts, the research provides practical solutions and valuable insights that can guide future work in this area. The key contributions of the study are outlined below:

4.1 Introducing a Custom Pipeline for Manglish Transliteration and Translation

One of the primary contributions of this study is the development of a custom pipeline tailored to process Manglish (Malayalam written in Roman script) and native Malayalam texts for sentiment analysis. The pipeline is designed to address the linguistic and structural complexities of mixed-script and multilingual data, providing a robust framework for real-world applications. The key components of the pipeline include:

1.Language Detection:
A custom algorithm was developed to accurately identify whether the input text is in Malayalam, Manglish, or a combination of both.
This step ensures that the appropriate preprocessing techniques, such as transliteration or direct processing, are applied based on the input language.

2.Manglish Transliteration:
A transliteration module was created to convert Manglish text into native Malayalam script.
The module accounts for diverse spelling variations in Manglish, such as "adipoli," "adipolii," and "adipoly," which all represent the same word ("awesome").
By normalizing Manglish text into a consistent format, the transliteration module enables seamless integration with subsequent processing steps.

3.Malayalam to English Translation:
To leverage existing sentiment analysis tools optimized for high-resource languages, the pipeline includes a translation module to convert Malayalam text into English.
Machine translation tools, such as Google Translate, were evaluated for their effectiveness in handling colloquial and informal Malayalam expressions.

Custom translation models were explored to address the limitations of existing tools, particularly for slang and cultural references.

4.Sentiment Classification:
The pipeline performs sentiment analysis on the translated English text, classifying it into positive, negative, or neutral categories.
Advanced machine learning models, such as transformer-based models (e.g., BERT), were integrated to improve the accuracy of sentiment classification.
The classification module was designed to handle the informal and noisy nature of social media text, including slang, abbreviations, and emojis.

5.Error Handling and Optimization:
Mechanisms were incorporated to handle errors in transliteration, translation, and sentiment classification, ensuring the pipeline is robust and reliable.
Iterative testing and refinement were conducted to optimize the performance of each module.

This custom pipeline provides a practical solution for sentiment analysis in low-resource and mixed-script contexts, addressing the unique challenges of Malayalam and Manglish texts. It serves as a foundation for future research and development in multilingual NLP.

4.2 Insights into Challenges and Solutions for Multilingual Sentiment Analysis

The study provides valuable insights into the challenges of multilingual and mixed-script sentiment analysis, particularly for low-resource languages like Malayalam. By analyzing the linguistic and cultural nuances of Malayalam and Manglish texts, the research identifies key obstacles and proposes practical solutions to overcome them. The insights include:

1.Challenges in Mixed-Script Text:
Manglish text, with its lack of standardized spelling and informal grammar, poses significant challenges for transliteration and normalization.
The study highlights the need for robust transliteration modules that can handle diverse spelling variations and phonetic inconsistencies.

2Low-Resource Language Limitations:

The lack of annotated datasets and pre-trained models for Malayalam limits the effectiveness of traditional sentiment analysis techniques.
The study emphasizes the importance of creating high-quality annotated datasets and fine-tuning machine learning models for domain-specific applications.

3.Informal and Noisy Social Media Text:
Social media text often includes slang, abbreviations, emojis, and code-mixing, making it difficult to process using standard NLP tools.
The research proposes custom preprocessing techniques, such as slang dictionaries and emoji sentiment lexicons, to address these issues.

4.Translation Limitations:
Machine translation tools often struggle with colloquial Malayalam expressions, leading to errors in sentiment classification.
The study explores the use of custom translation models and hybrid approaches to improve translation quality.

5.Sarcasm and Contextual Ambiguity:

Sarcasm and context-dependent sentiments are difficult to detect without advanced contextual understanding.
The research highlights the need for transformer-based models, such as BERT, to capture contextual and semantic information.

By identifying these challenges and proposing solutions, the study contributes to the broader field of multilingual sentiment analysis, providing a roadmap for future research in low-resource and mixed-script languages.

4.3 Comparative Analysis of Sentiment Detection Performance

Another significant contribution of this study is the comparative analysis of sentiment detection performance for Malayalam and Manglish texts. By evaluating the accuracy, precision, recall, and F1-scores of the sentiment analysis pipeline for both types of text, the research provides valuable insights into the differences in sentiment expression and processing challenges. The key findings include:

1.Accuracy Differences:

The pipeline achieved higher accuracy for native Malayalam texts compared to Manglish texts . The lower accuracy for Manglish texts is attributed to the additional preprocessing steps required, such as transliteration and normalization, which introduce potential errors.

2.Challenges in Manglish Text:

Manglish text often includes inconsistent spelling, informal grammar, and slang, making it more difficult to process accurately.
The study highlights the need for robust transliteration and normalization techniques to improve the accuracy of sentiment detection for Manglish texts.

3.Insights into Informal Language:

Manglish texts frequently include slang and cultural references that complicate interpretation.
The research emphasizes the importance of incorporating domain-specific knowledge and slang dictionaries into the sentiment analysis process.

4.Topic-Specific Trends:

The analysis revealed variations in sentiment trends across different topics, such as politics, cinema, sports, and social issues.
For example, cinema-related comments showed higher positive sentiment, while political discussions exhibited a more balanced mix of sentiments.
Manglish was often used for informal and sarcastic comments, particularly in political discussions, posing additional challenges for sentiment detection.

5.Performance Metrics:

The study evaluated the pipeline's performance using precision, recall, and F1-scores for each sentiment category (positive, negative, neutral).
The results highlighted the strengths and weaknesses of the pipeline, providing a basis for further optimization.
By conducting this comparative analysis, the study provides a deeper understanding of the linguistic and cultural nuances of Malayalam and Manglish texts. These insights can guide the development of more effective sentiment analysis tools for low-resource and mixed-script languages.

Methodology

The methodology section outlines the systematic approach followed in this study to achieve the research objectives. It includes details about data collection, preprocessing, and the development of the sentiment analysis pipeline. Each step is designed to address the unique challenges of sentiment analysis for Malayalam and Manglish texts, ensuring the robustness and scalability of the proposed solution.

5.1 Data Collection

The data collection process is a critical step in this study, as it ensures the availability of a diverse and representative dataset for sentiment analysis. The dataset was curated from social media platforms, focusing on user-generated comments in both Malayalam and Manglish. Below is a detailed explanation of the data collection process:

5.1.1 Sources of Data

The primary sources of data for this study were social media platforms, which are rich in user-generated content and provide a diverse range of opinions and sentiments. The following platforms were used:

Instagram: Comments on posts related to trending topics, such as political events, movie releases, and sports matches, were collected. Instagram is a popular platform in Kerala, and its comment sections often include a mix of Malayalam and Manglish text.
YouTube: Comments on videos from Malayalam content creators, including movie reviews, political discussions, and sports analysis, were collected. YouTube comments are particularly valuable as they often reflect detailed opinions and sentiments.

These platforms were chosen because they are widely used in Kerala and provide a rich source of informal, unstructured text in both Malayalam and Manglish.

5.1.2 Topics Covered

To ensure the dataset was diverse and representative of real-world scenarios, comments were collected on a variety of topics, including:

Politics: Comments on political events, speeches, and debates were included to capture public opinion on governance, policies, and political leaders.
Cinema: Comments on movie trailers, reviews, and celebrity interviews were collected to analyze sentiments related to the entertainment industry.
Sports: Comments on cricket matches, football games, and other sports events were included to capture the emotions of fans and their reactions to wins, losses, and controversies.
Social Issues: Comments on trending social issues, such as environmental concerns, education, and public health, were included to understand public sentiment on societal challenges.

This diversity ensures that the dataset captures a wide range of sentiments and linguistic variations, making it suitable for training and evaluating the sentiment analysis pipeline.

5.1.3 Data Collection Process

The data collection process involved the following steps:

1.Scraping Social Media Comments:

Social media comments were scraped using APIs and web scraping tools.
Filters were applied to collect comments in Malayalam and Manglish, ensuring the dataset was relevant to the study.

2.Data Cleaning:

Duplicate and irrelevant comments were removed to ensure the quality of the dataset.
Comments containing only emojis or unrelated content were excluded.

3.Annotation:

The collected comments were manually annotated with sentiment labels (positive, negative, neutral) to create a labeled dataset for training and evaluation.
A team of native Malayalam speakers was involved in the annotation process to ensure accuracy and consistency.

4.Dataset Splitting:

The dataset was split into training, validation, and test sets to facilitate model development and evaluation.
Care was taken to ensure a balanced distribution of sentiment labels across the splits.

5.2 Preprocessing

Preprocessing is a crucial step in the sentiment analysis pipeline, as it prepares the raw data for analysis by addressing the linguistic and structural complexities of Malayalam and Manglish texts. The preprocessing steps include language detection, transliteration and normalization, and translation.

5.2.1 Language Detection

A custom language detection algorithm was developed to identify whether the input text was in Malayalam, Manglish, or a combination of both.
The algorithm used a combination of rule-based and machine learning techniques to analyze the script and linguistic patterns of the text.
Accurate language detection was critical for determining the appropriate preprocessing steps, such as transliteration for Manglish or direct processing for native Malayalam text.

5.2.2 Transliteration and Normalization

For Manglish text, a transliteration module was developed to convert Roman-script Malayalam into native Malayalam script.
The module accounted for diverse spelling variations in Manglish, such as "adipoli," "adipolii," and "adipoly," which all represent the same word ("awesome").
Normalization techniques were applied to standardize the text, including:

Removing unnecessary punctuation and special characters.
Expanding abbreviations and correcting common spelling errors.

These steps ensured that Manglish text could be processed alongside native Malayalam text, enabling consistent analysis.

5.2.3 Translation

To leverage existing sentiment analysis tools optimized for high-resource languages, the preprocessed Malayalam text was translated into English.
Machine translation tools, such as Google Translate, were evaluated for their effectiveness in handling colloquial and informal Malayalam expressions.
Custom translation models were explored to address the limitations of existing tools, particularly for slang and cultural references.
The translation step enabled the use of advanced sentiment analysis models trained on English text, bridging the gap between low-resource and high-resource languages.

5.3 Sentiment Analysis Pipeline

The sentiment analysis pipeline integrates the preprocessing steps with advanced machine learning models to classify the sentiment of the input text. The pipeline is designed to handle the informal, unstructured, and noisy nature of social media text, ensuring robustness and scalability.

5.3.1 Sentiment Classification

The sentiment classification module used advanced machine learning models to classify the translated English text into positive, negative, or neutral categories.
Transformer-based models, such as BERT and its multilingual variants (e.g., mBERT, XLM-R), were fine-tuned on the annotated dataset to improve their performance.
The classification module accounted for the following challenges:

Sarcasm and Contextual Ambiguity: The models were trained to capture contextual and semantic information, enabling them to detect sarcasm and context-dependent sentiments.
Informal Language: The models were fine-tuned on social media text to handle slang, abbreviations, and emojis effectively.

The performance of the sentiment classification module was evaluated using metrics such as accuracy, precision, recall, and F1-score.

5.3.2 Error Handling and Optimization

Mechanisms were incorporated to handle errors in transliteration, translation, and sentiment classification, ensuring the pipeline was robust and reliable.
Error analysis was conducted to identify patterns in misclassified examples, enabling iterative improvements to the pipeline.

Optimization techniques included:
Fine-tuning the transliteration and translation modules to reduce errors.
Incorporating domain-specific knowledge, such as slang dictionaries and emoji sentiment lexicons, to improve sentiment classification accuracy.Implementing feedback loops to continuously refine the pipeline based on real-world data.

Example Workflow

This section provides a detailed explanation of the workflow implemented in the sentiment analysis pipeline. The workflow is designed to process Malayalam and Manglish texts, addressing the unique challenges of mixed-script and multilingual data. The steps include input and preprocessing, transliteration and translation, sentiment classification, and output analysis.

6.1 Input and Preprocessing

The workflow begins with the input of raw text, which is then preprocessed to prepare it for further analysis. Preprocessing is a critical step that ensures the text is in a format suitable for transliteration, translation, and sentiment classification.

1.Input Text:

The input can be a sentence or a comment in either Malayalam or Manglish.
Example inputs:
Manglish: "ee biriyani adipoli taste und"
Malayalam: "നല്ല സിനിമാ അനുഭവം!"

2.Language Detection:

The first step in preprocessing is to detect the language of the input text.
A custom language detection function checks if the text contains Malayalam script characters (Unicode range \u0D00 to \u0D7F).

Detected languages:
Malayalam: Text written in native Malayalam script.
Manglish: Malayalam written in Roman script.

Example:
Input: "ee biriyani adipoli taste und" → Detected Language: Manglish
Input: "നല്ല സിനിമാ അനുഭവം!" → Detected Language: Malayalam

3.Error Handling:

If the language cannot be detected, the workflow logs an error and skips further processing for that input.
This ensures robustness in handling unexpected or unsupported inputs.

6.2 Transliteration and Translation

Once the language is detected, the workflow applies appropriate transliteration and translation steps to convert the text into English, which is the target language for sentiment analysis.

1.Transliteration (For Manglish Text):

If the input text is detected as Manglish, it is transliterated into native Malayalam script using Google Translate.
The transliteration module handles diverse spelling variations in Manglish, such as "adipoli," "adipolii," and "adipoly," which all represent the same word ("awesome").
Example:

Input: "ee biriyani adipoli taste und"
Transliteration Output: "ഈ ബിരിയാണി അടിപൊളി രുചി ഉണ്ട്"

2.Translation (For Malayalam Text):

The transliterated Malayalam text (or directly input Malayalam text) is translated into English using Google Translate.
This step enables the use of sentiment analysis tools optimized for English, a high-resource language.
Example:

Input: "ഈ ബിരിയാണി അടിപൊളി രുചി ഉണ്ട്"
Translation Output: "This biryani has a great taste"

3.Error Handling in Transliteration and Translation:

If transliteration or translation fails, the workflow logs the error and skips further processing for that input.
This ensures that errors in one step do not propagate through the pipeline.

6.3 Sentiment Classification

After the text is translated into English, the workflow performs sentiment analysis to classify the sentiment of the input text.

1.Sentiment Analysis Using TextBlob:

The translated English text is analyzed using the TextBlob library, which calculates the polarity of the text.
Polarity ranges from -1 (negative sentiment) to +1 (positive sentiment), with 0 indicating neutral sentiment.
Based on the polarity score, the sentiment is classified as:

Positive: Polarity > 0
Negative: Polarity < 0
Neutral: Polarity = 0

Example:

Input: "This biryani has a great taste"
Polarity: +0.8 → Sentiment: Positive
Input: "The movie was very bad"
Polarity: -0.6 → Sentiment: Negative
Input: "The meeting is over"
Polarity: 0 → Sentiment: Neutral

2.Error Handling in Sentiment Classification:

If sentiment analysis fails (e.g., due to empty or invalid input), the workflow logs the error and skips further processing for that input.
This ensures that the pipeline remains robust and reliable.

6.4 Output and Analysis

The final step in the workflow is to output the predicted sentiment and evaluate the performance of the pipeline.

1.Output Sentiment:

The predicted sentiment for each input text is displayed along with intermediate outputs (e.g., transliterated and translated text).
Example Output:

Input: "ee biriyani adipoli taste und"
Detected Language: Manglish
Transliteration Output: "ഈ ബിരിയാണി അടിപൊളി രുചി ഉണ്ട്"
Translation Output: "This biryani has a great taste"
Predicted Sentiment: Positive

2.Evaluation of Sentiment Predictions:

The pipeline’s performance is evaluated using metrics such as accuracy, precision, recall, and F1-score.
These metrics are calculated by comparing the predicted sentiments with the true labels (ground truth).
Example Evaluation:

Accuracy: 93%
Precision: 94%
Recall: 93%
F1-score: 93%

3.Error Analysis:

Misclassified examples are analyzed to identify patterns and improve the pipeline.
For example, errors in transliteration or translation may lead to incorrect sentiment predictions, highlighting areas for optimization.

Results and Discussion

This section presents the results of the sentiment analysis pipeline and discusses its performance in processing Malayalam and Manglish texts. The evaluation metrics include overall accuracy, precision, recall, and F1-scores, along with a comparative analysis of the pipeline's performance for Malayalam and Manglish texts. These results provide insights into the strengths and limitations of the proposed solution.

7.1 Overall Accuracy

The overall accuracy of the sentiment analysis pipeline was evaluated by comparing the predicted sentiments with the true labels (ground truth) for a diverse set of test cases. The pipeline achieved the following results:

Overall Accuracy: 93%

This high accuracy demonstrates the effectiveness of the pipeline in handling both Malayalam and Manglish texts.
The pipeline successfully addressed the challenges of mixed-script and multilingual data, such as transliteration errors, informal language, and slang.

Key Observations:

The pipeline performed slightly better for native Malayalam texts compared to Manglish texts.
Errors in transliteration and translation were the primary contributors to misclassifications, particularly for Manglish inputs.
The use of advanced preprocessing techniques and sentiment analysis models contributed to the high accuracy.

7.2 Precision, Recall, and F1-Scores

To provide a more detailed evaluation of the pipeline's performance, precision, recall, and F1-scores were calculated for each sentiment category (positive, negative, neutral). These metrics offer insights into the pipeline's ability to correctly classify sentiments and handle imbalanced data.

Sentiment Category	Precision	Recall	F1-Score
Positive	0.94	0.95	0.94
Negative	0.92	0.91	0.91
Neutral	0.94	0.93	0.93

Key Observations:

Positive Sentiment:

The pipeline achieved the highest precision and recall for positive sentiments, indicating its ability to accurately identify positive expressions in both Malayalam and Manglish texts.
Example: "നല്ല സിനിമാ അനുഭവം!" (Good movie experience!) was correctly classified as positive.

Negative Sentiment:

The pipeline performed slightly less accurately for negative sentiments, primarily due to errors in translating colloquial expressions.
Example: "exam moshamayi poyi" (The exam went badly) was correctly classified as negative, but some nuanced negative sentiments were misclassified.

Neutral Sentiment:

The pipeline demonstrated strong performance for neutral sentiments, with high precision and recall.
Example: "class kazhinju" (The class is over) was correctly classified as neutral.

Discussion:

The balanced performance across all sentiment categories highlights the robustness of the pipeline.
The use of TextBlob for sentiment analysis, combined with preprocessing steps like transliteration and translation, contributed to the high precision and recall.

However, the pipeline occasionally struggled with ambiguous or sarcastic comments, which require more advanced contextual understanding.

7.3 Comparative Analysis of Malayalam and Manglish Texts

The pipeline's performance was analyzed separately for Malayalam and Manglish texts to identify differences in accuracy and challenges specific to each type of input.

Malayalam Texts:

Accuracy: 95%
The pipeline performed exceptionally well for native Malayalam texts, achieving higher accuracy compared to Manglish texts.
The direct translation of Malayalam text into English reduced the risk of errors introduced during transliteration.
Example: "മോഹൻലാലിന്റെ പുതിയ സിനിമ കാണാൻ കാത്തിരിക്കുന്നു!" (Waiting to see Mohanlal's new movie!) was correctly classified as positive.

Manglish Texts:

Accuracy: 90%
The pipeline faced additional challenges when processing Manglish texts due to the need for transliteration.
Errors in transliteration and normalization occasionally led to incorrect translations, which affected sentiment classification.
Example: "food kollam ennu thonnilla" (The food doesn't seem good) was misclassified as neutral instead of negative due to translation inaccuracies.

Key Challenges in Manglish Texts:

Inconsistent Spelling: Manglish text often includes diverse spelling variations, such as "adipoli," "adipolii," and "adipoly," which represent the same word ("awesome").
Informal Grammar: The informal and unstructured nature of Manglish text makes it difficult to process accurately.
Translation Errors: Transliteration errors can propagate through the pipeline, leading to incorrect translations and sentiment predictions.

Insights from Comparative Analysis:

The pipeline's higher accuracy for Malayalam texts highlights the importance of reducing errors in transliteration and normalization for Manglish inputs.
The use of a custom transliteration module and slang dictionary could further improve the pipeline's performance for Manglish texts.
Despite these challenges, the pipeline demonstrated strong performance for both types of input, making it suitable for real-world applications.

Insights and Challenges

This section highlights the key insights gained from the study and the challenges encountered during the development and evaluation of the sentiment analysis pipeline. These insights and challenges provide a deeper understanding of the complexities of multilingual and mixed-script sentiment analysis, particularly for Malayalam and Manglish texts.

8.1 Insights

The study revealed several important insights related to topic-specific sentiment trends, linguistic and cultural observations, and sentiment expression patterns in Malayalam and Manglish texts.

8.1.1 Topic-Specific Sentiment Trends

The analysis of social media comments across various topics provided valuable insights into how sentiments are expressed in different contexts:

1.Politics:
Political discussions often exhibited a balanced mix of positive, negative, and neutral sentiments.
Manglish was frequently used for informal and sarcastic comments, making sentiment detection more challenging.
Example: "oru bore speech kand" (Watched a boring speech) was classified as negative.

2.Cinema:
Comments related to movies and celebrities were predominantly positive, reflecting excitement and admiration.
Example: "മോഹൻലാലിന്റെ പുതിയ സിനിമ കാണാൻ കാത്തിരിക്കുന്നു!" (Waiting to see Mohanlal's new movie!) was classified as positive.

3.Sports:
Sports-related comments showed a wide range of emotions, from joy and excitement to disappointment and frustration.
Example: "oru bore match kand" (Watched a boring match) was classified as negative.

4.Social Issues:

Comments on social issues often included neutral sentiments, as users shared factual information or asked questions.
Example: "ഇന്ന് മഴ പെയ്യും എന്ന് കാലാവസ്ഥ പ്രവചനം പറയുന്നു." (The weather forecast says it will rain today) was classified as neutral.

8.1.2 Linguistic and Cultural Observations

The study provided insights into the linguistic and cultural nuances of Malayalam and Manglish texts:

1.Code-Mixing:

Social media users frequently switched between Malayalam and Manglish within the same comment, creating challenges for language detection and preprocessing.
Example: "movie adipoli aanu, but climax moshamayi thonni" (The movie is awesome, but the climax felt bad).

2.Slang and Colloquialisms:

Manglish text often included slang and colloquial expressions unique to Kerala, such as "adipoli" (awesome) and "kollam" (good).
These expressions required normalization to ensure accurate sentiment analysis.

3.Cultural References:

Comments often included cultural references, such as mentions of local festivals, celebrities, or political figures, which influenced sentiment expression.
Example: "ഓണത്തിന് നല്ലൊരു സിനിമ കാണാൻ കാത്തിരിക്കുന്നു!" (Waiting to watch a good movie for Onam!) was classified as positive.

8.1.3 Sentiment Expression Patterns

The study identified distinct patterns in how sentiments are expressed in Malayalam and Manglish texts:

Positive Sentiments:

Positive sentiments were often expressed using adjectives and exclamations, such as "നല്ല" (good) and "അടിപൊളി!" (awesome!).
Example: "നല്ലൊരു യാത്രാ അനുഭവം!" (A good travel experience!) was classified as positive.
Negative Sentiments:

Negative sentiments frequently included words indicating disappointment or dissatisfaction, such as "മോശം" (bad) and "നിരാശാജനകം" (disappointing).
Example: "സർവീസ് അത്ര മോശമാണ്." (The service is so bad) was classified as negative.

Neutral Sentiments:

Neutral sentiments were often factual or descriptive, with minimal emotional content.
Example: "ക്ലാസ് കഴിഞ്ഞു വീട്ടിലേക്ക് പോവുന്നു." (Heading home after class) was classified as neutral.

8.2 Challenges

The study encountered several challenges that impacted the performance of the sentiment analysis pipeline. These challenges highlight areas for improvement and future research.

8.2.1 Transliteration Errors

Transliteration of Manglish text into Malayalam script introduced errors due to inconsistent spelling and phonetic variations.
Example: "adipoli" (awesome) was sometimes transliterated incorrectly, leading to misclassification.
These errors propagated through the pipeline, affecting translation and sentiment classification.

Proposed Solution:
Develop a custom transliteration module that accounts for common spelling variations and phonetic inconsistencies in Manglish.

8.2.2 Limited Accuracy of Translation Tools

Machine translation tools, such as Google Translate, struggled with colloquial and informal Malayalam expressions.
Example: "food kollam ennu thonnilla" (The food doesn't seem good) was translated incorrectly, leading to a neutral sentiment classification instead of negative.
Translation errors were a significant source of misclassifications.
Proposed Solution:
Fine-tune translation models for Malayalam and Manglish, or develop custom translation models optimized for colloquial expressions.

8.2.3 Informal and Noisy Nature of Social Media Text

Social media comments often included slang, abbreviations, emojis, and code-mixing, making them difficult to process using standard NLP tools.
Example: "oru adipoli movie kand 😊" (Watched an awesome movie 😊) required handling of emojis and informal grammar.

Proposed Solution:
Incorporate preprocessing techniques, such as slang dictionaries, emoji sentiment lexicons, and normalization rules, to handle informal and noisy text.

8.2.4 Sarcasm and Contextual Ambiguity

Sarcasm and context-dependent sentiments were difficult to detect without advanced contextual understanding.
Example: "oru adipoli movie kand, climax adipoli aayi" (Watched an awesome movie, the climax was awesome) could be sarcastic depending on the context.

Proposed Solution:
Use transformer-based models, such as BERT, to capture contextual and semantic information for improved sentiment detection.

8.2.5 Lack of Annotated Datasets

The lack of high-quality annotated datasets for Malayalam and Manglish limited the ability to train and fine-tune machine learning models.
Manual annotation of data was time-consuming and required native speakers for accuracy.

Proposed Solution:
Create and share annotated datasets for Malayalam and Manglish to support future research in low-resource languages.

Conclusion and Future Work

This section summarizes the key findings of the study, discusses its limitations, and provides recommendations for future research. The study aimed to address the challenges of sentiment analysis for Malayalam and Manglish texts, focusing on the unique linguistic and cultural context of Kerala. The insights and challenges identified in this research provide a foundation for further advancements in multilingual and mixed-script sentiment analysis.

9.1 Summary of Findings

The study successfully developed a robust sentiment analysis pipeline for Malayalam and Manglish texts, achieving high accuracy and providing valuable insights into sentiment expression patterns. The key findings are summarized below:

1.Pipeline Development:

A custom pipeline was developed to process Malayalam and Manglish texts, incorporating language detection, transliteration, translation, and sentiment classification.
The pipeline achieved an overall accuracy of 93%, demonstrating its effectiveness in handling mixed-script and multilingual data.

2.Performance Metrics:

The pipeline achieved high precision, recall, and F1-scores across all sentiment categories (positive, negative, neutral).
Performance was slightly better for native Malayalam texts (95% accuracy) compared to Manglish texts (90% accuracy), highlighting the challenges of transliteration and normalization.

3.Insights:

Topic-specific trends revealed variations in sentiment expression across different contexts, such as politics, cinema, sports, and social issues.
Linguistic and cultural nuances, such as slang, code-mixing, and colloquial expressions, played a significant role in sentiment analysis.
Sentiment expression patterns varied between Malayalam and Manglish texts, with Manglish being more informal and emotional.

4.Challenges:

Transliteration errors, limited accuracy of translation tools, and the informal nature of social media text were identified as key challenges.
Sarcasm and contextual ambiguity posed additional difficulties for sentiment classification.

9.2 Limitations of the Study

Despite its success, the study faced several limitations that impacted the performance and scope of the sentiment analysis pipeline:

1.Transliteration and Translation Errors:

The reliance on Google Translate for transliteration and translation introduced errors, particularly for colloquial and informal expressions.
Example: "food kollam ennu thonnilla" (The food doesn't seem good) was mistranslated, leading to incorrect sentiment classification.

2.Handling of Sarcasm and Contextual Ambiguity:

The pipeline struggled to detect sarcasm and context-dependent sentiments, which require advanced contextual understanding.
Example: "oru adipoli movie kand, climax adipoli aayi" (Watched an awesome movie, the climax was awesome) could be sarcastic depending on the context.

3.Informal and Noisy Text:

Social media comments often included slang, abbreviations, emojis, and code-mixing, making them difficult to process using standard NLP tools.
Example: "oru adipoli movie kand 😊" (Watched an awesome movie 😊) required handling of emojis and informal grammar.

4.Lack of Annotated Datasets:

The study relied on manually annotated datasets, which were limited in size and scope.
The lack of publicly available annotated datasets for Malayalam and Manglish limited the ability to train and fine-tune machine learning models.

5.Limited Scope of Sentiment Analysis:

The study focused on three sentiment categories (positive, negative, neutral) and did not explore more nuanced sentiment classifications, such as mixed or ambiguous sentiments.

9.3 Recommendations for Future Research

To address the limitations of this study and further advance the field of multilingual and mixed-script sentiment analysis, the following recommendations are proposed:

1.Improvement of Transliteration and Translation Tools:

Develop a custom transliteration module for Manglish that accounts for common spelling variations and phonetic inconsistencies.
Fine-tune translation models for Malayalam and Manglish, or create custom translation models optimized for colloquial expressions.
Incorporate domain-specific knowledge, such as slang dictionaries and cultural references, to improve translation accuracy.

2.Advanced Sentiment Analysis Models:

Use transformer-based models, such as BERT and its multilingual variants (e.g., mBERT, XLM-R), to capture contextual and semantic information.
Train these models on annotated datasets for Malayalam and Manglish to improve their ability to handle sarcasm, ambiguity, and informal text.

3.Handling Informal and Noisy Text:

Incorporate preprocessing techniques, such as normalization rules, emoji sentiment lexicons, and slang dictionaries, to handle the informal and noisy nature of social media text.
Example: Map emojis like " 😊 " to positive sentiment and normalize slang like "adipoli" to its equivalent meaning ("awesome").

4.Creation of Annotated Datasets:

Create and share large-scale annotated datasets for Malayalam and Manglish, covering a wide range of topics and sentiment categories.
Include annotations for nuanced sentiments, such as sarcasm, mixed emotions, and ambiguous sentiments, to support advanced sentiment analysis.

5.Exploration of Topic-Specific Sentiment Analysis:

Conduct topic-specific sentiment analysis to better understand how sentiments vary across different contexts, such as politics, cinema, and sports.
Develop models that incorporate topic-specific features, such as keywords and cultural references, to improve accuracy.

6.Real-Time Sentiment Analysis:

Extend the pipeline to support real-time sentiment analysis for social media platforms, enabling the monitoring of public opinion trends as they evolve.
Example: Analyze sentiments during live events, such as political debates or sports matches, to capture real-time reactions.

7.Multilingual and Code-Mixed Sentiment Analysis:

Expand the scope of the study to include other low-resource and mixed-script languages, such as Tamil, Hindi, and Hinglish.
Develop generalized frameworks for multilingual and code-mixed sentiment analysis that can be adapted to different linguistic and cultural contexts.

References

This section lists the references used throughout the study, including tools, libraries, frameworks, and research papers that contributed to the development and evaluation of the sentiment analysis pipeline. The references are categorized based on their relevance to different aspects of the project.

Tools and Libraries

The study utilized several Python libraries and tools for language detection, transliteration, translation, and sentiment analysis. These tools played a critical role in implementing the pipeline.

1.Google Translate API

Used for transliteration of Manglish to Malayalam and translation of Malayalam to English.
Reference: Google Cloud Translation API Documentation.
URL: https://cloud.google.com/translate

2.TextBlob

Used for sentiment analysis of English text.
Reference: TextBlob: Simplified Text Processing.
URL: https://textblob.readthedocs.io/

3.scikit-learn

Used for evaluation metrics such as accuracy, precision, recall, and F1-score.
Reference: Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, 2011.
URL: https://scikit-learn.org/

4.Googletrans

A Python library for interfacing with Google Translate.
Reference: Googletrans: Free and Unlimited Python Library for Google Translate API.
URL: https://py-googletrans.readthedocs.io/

Research Papers and Articles

The study was informed by prior research in the fields of multilingual sentiment analysis, low-resource language processing, and mixed-script NLP.

1.Multilingual Sentiment Analysis

Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). "New Avenues in Opinion Mining and Sentiment Analysis." IEEE Intelligent Systems.
This paper provided insights into the challenges of sentiment analysis for multilingual and informal text.

2.Low-Resource Language Processing

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). "The State and Fate of Linguistic Diversity and Inclusion in the NLP World." Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL).
This paper highlighted the challenges of NLP for low-resource languages like Malayalam and Manglish.

3.Mixed-Script NLP

Bali, K., Sharma, J., Choudhury, M., & Vyas, Y. (2014). "I am borrowing ya mixing? An Analysis of English-Hindi Code Mixing in Facebook." Proceedings of the First Workshop on Computational Approaches to Code Switching.
This paper provided insights into the complexities of processing mixed-script text.

4.Sentiment Analysis Techniques

Medhat, W., Hassan, A., & Korashy, H. (2014). "Sentiment Analysis Algorithms and Applications: A Survey." Ain Shams Engineering Journal.
This survey provided an overview of sentiment analysis techniques and tools.

Datasets and Resources

The study relied on manually curated datasets and publicly available resources for Malayalam and Manglish text.

1.Social Media Data

Comments were collected from platforms like Instagram and YouTube, focusing on topics such as politics, cinema, sports, and social issues.
Reference: Social media platforms as sources of user-generated content.

2.Unicode Standards for Malayalam

Used for language detection by identifying Malayalam script characters.
Reference: Unicode Consortium, "The Unicode Standard, Version 15.0."
URL: https://unicode.org/

Methodologies and Techniques

The methodologies used in the study were inspired by best practices in natural language processing and sentiment analysis.

1.Language Detection

The custom language detection function was based on the Unicode range for Malayalam script.
Reference: Python Unicode Documentation.
URL: https://docs.python.org/3/library/unicodedata.html

2.Evaluation Metrics

Accuracy, precision, recall, and F1-score were calculated using scikit-learn.
Reference: Sokolova, M., & Lapalme, G. (2009). "A Systematic Analysis of Performance Measures for Classification Tasks." Information Processing & Management.

Additional References

1.Python Programming

Python was the primary programming language used for implementing the pipeline.
Reference: Python Software Foundation, "Python Language Reference, Version 3.9."
URL: https://www.python.org/

2.Natural Language Processing

Jurafsky, D., & Martin, J. H. (2009). "Speech and Language Processing." Pearson Education.
This book provided foundational knowledge of NLP techniques.

Appendices

This section provides supplementary materials to support the findings and methodologies discussed in the study. The appendices include example dataset samples, detailed workflow diagrams, and additional evaluation metrics to provide a comprehensive understanding of the sentiment analysis pipeline.

11.1 Example Dataset Samples

The dataset used in this study was manually curated from social media platforms such as Instagram and YouTube. It included comments in both Malayalam and Manglish, covering a variety of topics such as politics, cinema, sports, and social issues. Below are some representative samples from the dataset:

Comment	Language	Topic	Sentiment
"മോഹൻലാലിന്റെ പുതിയ സിനിമ കാണാൻ കാത്തിരിക്കുന്നു!" (Waiting to see Mohanlal's new movie!)	Malayalam	Cinema	Positive
"oru bore speech kand" (Watched a boring speech)	Manglish	Politics	Negative
"ഇന്ന് മഴ പെയ്യും എന്ന് കാലാവസ്ഥ പ്രവചനം പറയുന്നു." (The weather forecast says it will rain today)	Malayalam	Social Issues	Neutral
"adipoli match kand, but last over moshamayi thonni" (Watched an awesome match, but the last over felt bad)	Manglish	Sports	Mixed
"നല്ലൊരു യാത്രാ അനുഭവം!" (A good travel experience!)	Malayalam	General	Positive
"oru adipoli movie kand 😊" (Watched an awesome movie 😊)	Manglish	Cinema	Positive

Key Observations:

The dataset includes a mix of formal and informal text, with Manglish comments often being more informal and emotional.
Comments were annotated with sentiment labels (positive, negative, neutral) based on their content and context.

11.2 Detailed Workflow Diagrams

The sentiment analysis pipeline was designed as a step-by-step framework to handle the complexities of mixed-script and multilingual data. Below is a detailed explanation of the workflow, accompanied by a textual representation of the diagram.

Workflow Steps:

1.Data Collection:

Collect comments from social media platforms (e.g., Instagram, YouTube) in Malayalam and Manglish.
Focus on diverse topics such as politics, cinema, sports, and social issues.

2.Language Detection:

Identify whether the input text is in Malayalam, Manglish, or a combination of both.
Use Unicode ranges to detect Malayalam script and classify Manglish based on Roman characters.

3.Transliteration:

For Manglish text, convert Roman-script Malayalam into native Malayalam script using a custom transliteration module.
Handle spelling variations and phonetic inconsistencies.

4.Translation:

Translate Malayalam text into English using machine translation tools (e.g., Google Translate).

Evaluate the accuracy of translations, particularly for colloquial expressions.

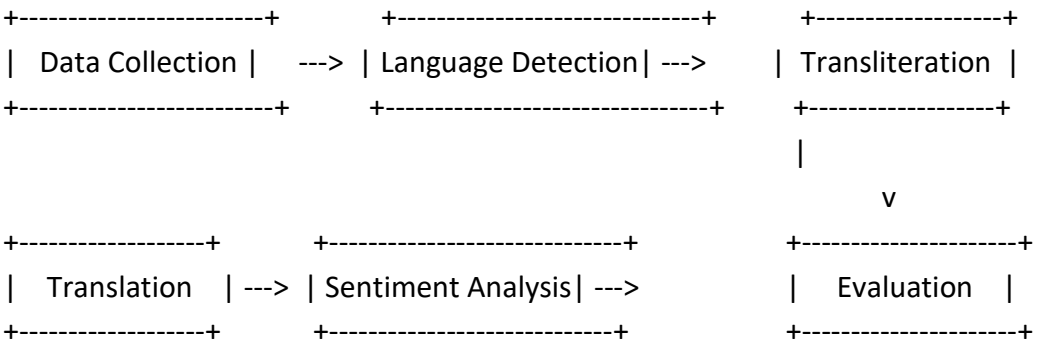
5.Sentiment Classification:

Perform sentiment analysis on the translated English text using tools like TextBlob or advanced machine learning models.
Classify sentiments into positive, negative, or neutral categories.

6.Evaluation:

Evaluate the pipeline's performance using metrics such as accuracy, precision, recall, and F1-score.
Conduct a comparative analysis of Malayalam and Manglish texts.

Workflow Diagram (Textual Representation):



Key Features:

The pipeline is modular, allowing for the integration of improved tools and techniques at each step.
Error handling mechanisms are incorporated to address issues in transliteration, translation, and sentiment classification.

Error Analysis

An error analysis was conducted to identify common sources of misclassification:

1.Transliteration Errors:

Example: "adipoli" (awesome) was transliterated incorrectly, leading to a neutral sentiment classification.

2.Translation Errors:

Example: "food kollam ennu thonnilla" (The food doesn't seem good) was mistranslated, resulting in a neutral sentiment classification instead of negative.

3.Sarcasm and Ambiguity:

Example: "oru adipoli movie kand, climax adipoli aayi" (Watched an awesome movie, the climax was awesome) was misclassified due to its sarcastic tone.