Sentimental Analysis for Code-Mixed Language Using Deep Learning

MAJOR PROJECT PHASE-1 REPORT
Submitted by

Dikshya Aryal (1BM21CS058)
Anitha K J (1BM22CS401)
B S Swaraj(1BM22CS403)
B Venkatesh(1BM22CS404)

in partial fulfillment for the award of the degree of
BACHELOR OF ENGINEERING
in
COMPUTER SCIENCE AND ENGINEERING

Under the Guidance of
Prof. Sandhya A Kulkarni
Assistant Professor, BMSCE

B. M. S. COLLEGE OF ENGINEERING
(Autonomous Institution under VTU)
BENGALURU-560019
2024-2025

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

Certified that the project entitled "**Sentimental Analysis for Code-Mixed Language Using Deep Learning**" is a bonafide work carried out by **Dikshya Aryal (1BM21CS058)**, **Anitha K J (1BM22CS401), B S Swaraj(1BM22CS403), B Venkatesh(1BM22CS404)** in partial fulfillment for the award of Bachelor of Engineering in Computer Science and Engineering of the Visvesvaraya Technological University, Belagavi during the academic year 2024-25. The project report has been approved as it satisfies the academic requirements in respect of project work prescribed for the said degree.

Guide

Prof. Sandhya A Kulkarni,
Assistant Professor,
Dept of CSE,
B.M.S. College of Engineering

Head of Department

Dr. Kavitha Sooda,
Professor and HOD,
Dept of CSE,
B.M.S. College of Engineering

Principal

Dr. Bheemsha Arya
B.M.S. College of Engineering

External Viva

Name of the Examiners	Signature with Date
1.	
2.	

Table of Contents

Abstract

Code-mixed languages are a blend of two or more languages within a single sentence or discourse. In today's multilingual digital world, code-mixed languages have become increasingly common, especially on social media platforms and informal communication channels. The task of sentimental analysis in such data is challenging because of the lack of standardized syntax, frequent language switching, and ambiguous emotional expressions. This project is concerned with sentimental analysis of code-mixed languages based on advanced Deep Learning techniques. In particular, it targets critical research gaps, like the detection of sarcasm and irony, and managing complicated emotional ambiguities that have often been neglected in the design of tradition al sentimental analysis models. The project leverages the DL architecture such as LSTM, and transformers-based models, for instance, BERT in order to extract contextualized embeddings along with rich semantic information from code-mixed sentences.

These architectures would capture short-as well as long -range dependencies inside the code - mixed sentences. This approach combines different features, like textual context and linguistic cues, to interpret sentiments effectively and identify complex patterns, including sarcasm and irony.

The proposed framework has profound implications for understanding user emotions in code-mixed languages, which are highly prevalent in multilingual communities. Applications of this research include social media monitoring, customer feedback analysis, and multilingual chatbot systems, where accurate sentiment analysis is critical. By overcoming the limitations of existing approaches and addressing the complexities of code- mixed language processing, this work contributes to advancing NLP for diverse linguistic environments.

DECLARATION

We, hereby declare that the Major Proje c t Phase - 1 work entitled " **Sentimental Analysis for Code-Mixed Language Using Deep Learning**" is a bonafide work and has been carried out by us under the guidance of **Prof. Sandhya A Kulkarni ,** Assistant Professor, Department of Computer Science and Engineering, B.M.S. College of Engineering, Bengaluru**,** in partial fulfillment of the requirements of the degree of Bachelor of Engineering in Computer Science and Engineering of Visvesvaraya Technological University, Belagavi.

I further declare that, to the best of my knowledge and belief, this project has not been submitted either in part or in full to any other university for the award of any degree.

Candidate details:

SL. NO.	Student Name	USN	Student's Signature
1	Dikshya Aryal	1BM21CS058	
2	Anitha K J	1BM22CS401	
3	B S Swaraj	1BM22CS403	
4	B Venkatesh	1BM22CS404	

Place: Bengaluru

Date: 27-01-2025

Certified that these candidates are students of Computer Science and Engineering Department of B.M.S. College of Engineering. They have carried out the project work of titled "**Sentimental Analysis for Code-Mixed Language Using Deep Learning"** as Major Project Phase-1 work. It is in partial fulfillment for completing the requirement for the award of B.E. degree by VTU. The works is original and duly certify the same.

Guide Name Signature

Prof. Sandhya A Kulkarni

Date: 27-01-2025

Acknowledgment

We would like to take this opportunity to express our deepest gratitude to **B.M.S. College of Engineering**, particularly the **Department of Computer Science and Engineering**, for providing us with the platform and resources to undertake our project titled **"Sentiment Analysis for Code-Mixed Language Using Deep Learning."** The academic environment, infrastructural support, and encouragement we received from the institution played a vital role in the successful completion of this project.

Our sincere thanks and appreciation go to our esteemed project guide, **Prof. Sandhya A Kulkarni**, for her unwavering support, expert guidance, and invaluable insights throughout this journey. Her extensive knowledge, timely feedback, and constructive criticism have been instrumental in helping us overcome obstacles, refine our approach, and achieve the desired outcomes. Her encouragement to think critically and independently has not only enhanced the quality of our project but also significantly contributed to our personal and professional growth.

We are profoundly thankful to our **Head of Department** and all the faculty members of the **Department of Computer Science and Engineering** for their constant encouragement, valuable suggestions, and constructive feedback during various stages of this project. Their dedication to fostering a learning-centric environment and their willingness to help at every step has been truly inspiring.

Additionally, we would like to acknowledge the immense support and motivation provided by our family and friends. Their understanding and encouragement during the demanding phases of this project gave us the strength to persevere. Without their constant emotional and moral support, completing this project would have been far more challenging.

Finally, we extend our gratitude to everyone who directly or indirectly contributed to the successful completion of this project. Their contributions, no matter how big or small, have been integral to the realization of our objectives. It has been an enriching experience, and we are profoundly grateful for all the learning and support we have received during this endeavor.

List of Figures

List of Tables

Chapter 1

Introduction

In today's digital era, sentiment analysis plays a critical role in understanding user emotions, particularly on platforms where individuals communicate using **code-mixed languages**—a blend of two or more languages in a single sentence. This project, titled **"Sentiment Analysis for Code-Mixed Language Using Deep Learning"** aims to address these challenges using **Deep Learning (DL)** techniques. By leveraging state-of-the-art DL architectures, the project enhances sentiment detection accuracy, bridging gaps in areas such as sarcasm, irony, and ambiguous emotion identification.

1.1 Overview

This project focuses on developing a framework for **sentiment analysis** in code-mixed languages, combining advanced Deep Learning techniques. Traditional sentiment analysis models often fail to process complex and inconsistent linguistic structures in code-mixed data. Our solution employs **Long Short-Term Memory (LSTM)**, and transformer-based models to analyze contextual and semantic information in mixed-language content. By integrating features such as linguistic cues and contextual embeddings, this project addresses critical research gaps like **sarcasm detection** and ambiguous emotion recognition. The outcome will contribute to improving sentiment analysis across platforms like social media, forums, and multilingual communication tools.

1.2 Motivation

With the rise of social media and global communication, **code-mixed languages** have become increasingly prevalent, especially in multilingual communities. However, analyzing user sentiments in such languages poses a significant challenge due to the lack of standardized syntax and frequent switching between languages. Existing sentiment analysis models often overlook nuances like **sarcasm and mixed emotions**. This motivated us to create a framework that leverages Deep Learning to overcome these limitations. By addressing these gaps, we aim to improve sentiment analysis accuracy and contribute to real-world applications such as **social media monitoring**, customer reviews analysis, and multilingual AI systems.

1.3 Objectives

The primary objective of this project is to develop a robust **Deep Learning-based framework** for sentiment analysis in **code-mixed languages**. Specific objectives include:

- To Build /Collect a dataset of code-mixed Kannada-English and Hindi-English texts, annotated for sentiment, sarcasm, and idioms.
- To develop a system to identify the language of the text given.
- To build a system that integrates text and fine-tune LLMs, and deep learning models to analyze mixed languages and detect hidden emotions.
- To test and validate the model's effectiveness, measuring improvements over traditional sentiment analysis systems.

1.4 Scope

The scope of this project encompasses the analysis and interpretation of **sentiments in code-mixed languages** using Deep Learning methodologies. It includes the development and testing of models like **LSTM** and transformer-based techniques for better contextual understanding. The project focuses on identifying emotions such as **positive, negative, neutral sentiments**, as well as complex emotions like sarcasm. The outcomes are applicable in areas such as **social media analytics, multilingual chatbot development**, and feedback systems for businesses. By addressing challenges specific to code-mixed data, this project paves the way for further research and innovation in **Natural Language Processing** for multilingual settings.

1.5 Existing System

2. **Rule-Based Systems:**
   Use dictionaries and regex to identify languages. Work well for structured text but struggle with informal code-mixed languages.
3. **Traditional Machine Learning Models:**
   SVM and Naive Bayes rely on feature engineering but fail to capture complex patterns in code-mixed text.

4. **Deep Learning Models (BiLSTM, CNN):**
   Handle sequence data better but need large datasets and struggle with code-mixed sentences.
5. **Multilingual Models (mBERT, XLM-R):**
   pre-trained models handle multiple languages but aren't optimized for code-mixed texts.

1.6 Proposed System

1. BiLSTM with Attention Mechanism:

Achieves high accuracy (94.8%) and F1-score (0.91) for Kannada-English code-mixed data by focusing on word-level identification.

2. Multilingual Fine-Tuned Models (XLM-R, mBERT):

Fine-tuning pre-trained models improves sentiment analysis for code-mixed texts across multiple languages.

3. Sentiment Analysis:

Combines text and emojis to better understand sentiment in informal communication.

4. Contextual and Hybrid Models:

Use attention mechanisms and combine language identification with sentiment analysis for context-rich code-mixed data.

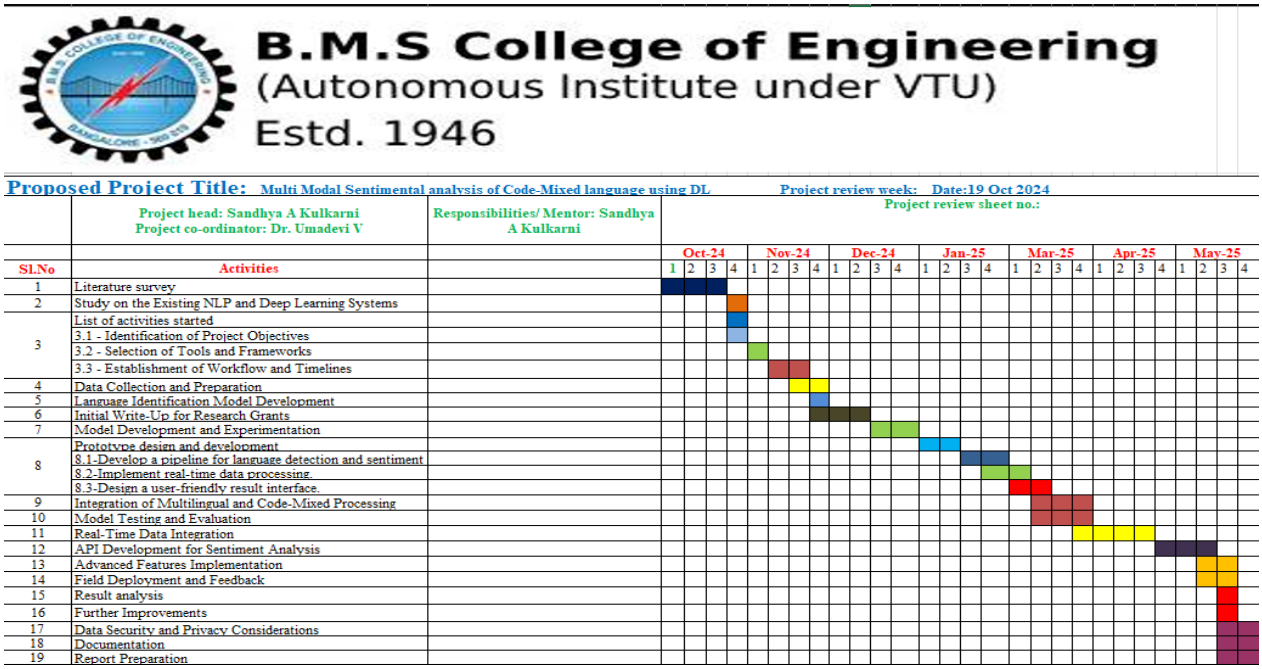
Domain-Specific Fine-Tuning:

Tailor models to specific use cases (e.g., social media or reviews) to enhance performance and generalization.

1.7 Work Plan

Table 1.1: Work Plan Table

1. Research and Data Collection (Weeks 1-4): **Weeks 1-2:** Gather code-mixed Kannada and Hindi data from social media platforms (e.g., YouTube, Twitter). Collect data **Weeks 3-4:** Organize collected data and pre-process (e.g., removing noise, translations, etc.)	**2. Data Annotation (Weeks 5-8):** Weeks 5-6: Annotate the dataset for sentiment, sarcasm expressions. Use both manual and automated annotation tools. Weeks 7-8: Validate and finalize the annotation for accuracy.
3. Model Design and Development (Weeks 9-12): Weeks 9-10: Design the sentiment analysis architecture. Fine-tune LLMs (e.g., mBERT, GPT) and integrate text and audio data. Weeks 11-12: Begin training models with both text and audio data.	**4. Model Training (Weeks 13-16):** Train the model on the annotated dataset. Evaluate performance on sentiment, sarcasm, and idiom detection using precision, recall, and F1-score.
5. Testing and Validation (Weeks 17-20): Test the model on unseen data. Compare results against existing sentiment analysis models. Perform result analysis and identify areas for improvement.	**6. Result Analysis (Weeks 21-24):** Analyze model performance in detail, including across different modalities (text, audio, video). Implement improvements (e.g., enhance sarcasm detection). Fine-tune based on insights from testing.
7. Final Implementation (Weeks 25-26): Finalize the system with all improvements incorporated. Ensure the model is ready for real-world application.	**8. Documentation (Weeks 27-28):** Prepare detailed project documentation (data, models, evaluation). Submit the final report and present findings and insights to stakeholders or research bodies.



B.M.S College of Engineering
(Autonomous Institute under VTU)
Estd. 1946

Proposed Project Title: Multi Modal Sentimental analysis of Code-Mixed language using DL

Project review week: Date:19 Oct 2024

Project review sheet no.:

Project head: Sandhya A Kulkarni
Project co-ordinator: Dr. Umadevi V

Responsibilities/ Mentor: Sandhya A Kulkarni

Sl.No	Activities	Oct-24				Nov-24				Dec-24				Jan-25				Mar-25				Apr-25				May-25			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	Literature survey																												
2	Study on the Existing NLP and Deep Learning Systems																												
3	List of activities started																												
	3.1 - Identification of Project Objectives																												
	3.2 - Selection of Tools and Frameworks																												
	3.3 - Establishment of Workflow and Timelines																												
4	Data Collection and Preparation																												
5	Language Identification Model Development																												
6	Initial Write-Up for Research Grants																												
7	Model Development and Experimentation																												
8	Prototype design and development																												
	8.1-Develop a pipeline for language detection and sentiment																												
	8.2-Implement real-time data processing.																												
	8.3-Design a user-friendly result interface.																												
9	Integration of Multilingual and Code-Mixed Processing																												
10	Model Testing and Evaluation																												
11	Real-Time Data Integration																												
12	API Development for Sentiment Analysis																												
13	Advanced Features Implementation																												
14	Field Deployment and Feedback																												
15	Result analysis																												
16	Further Improvements																												
17	Data Security and Privacy Considerations																												
18	Documentation																												
19	Report Preparation																												

Fig 1.1: Gantt Chart

Chapter 2

Literature Survey

Paper [1]

The article under review analyses existing sentiments in the Indian code-mixed text as recorded in various social media. In so doing, this study has observed that traditional models based on machine learning, for instance, SVM, Naïve Bayes, and Random Forest are widely used although SVM has been convenient for analyzing multiple datasets. Deep learning models, including, inter-alia, LSTM, BiLSTM and CNN appear to be more effective in treating code mixing as they provide better accuracy and f1 scores in situations where the data being analyzed is unstructured and noisy. In addition, these studies focus on the use of ensemble techniques or a combination of surface features and neural networks. It is also apparent that the reviewed papers seem to emphasize the usage of Hindi-English and Bengali-English pairs, with little attention to the use of English and other Indian languages like Telugu, Marathi and Kannada. Some of the gaps that were identified include the unavailability of annotated corpora as well as sufficient pre-processing methods and tools for languages and parts of speech tagging in code-switched documents.

Paper [2]

The text “Sentiment Analysis of Mixed Code for Transliterated Hindi and Marathi Texts” (2018) deals with the issues pertaining with the analysis of sentiments in code mixed scripts. Social media texts that are transliterated in Hindi and Marathi is its main focus. The proposed approach combines supervised learning methods like KNN, Naïve Bayes, and SVM with ontology-based methods in one model. The authors report that the Marathi work achieved as high as 90 percent accuracy, while the work on Hindi reached between 80 and 90 percent. Their contributions pave the way for more advanced identification of languages, building a Marathi SentiWordNet, and overcoming grammatical problems in mixed code text. This work lays the groundwork for future research on multilingual transliterated texts in social contexts.

Paper [3]

The paper tackles the issue of understanding Indonesian-English code-mixed text which with the title Code-Mixed Sentiment Analysis using Transformer for Twitter Social Media Data (2023), on the language resources part of their work. To this end, the authors created a new code-mixed dataset using Twitter data adopting BERT-based pretrained models such as IndoBERTweet and Multilingual BERT along with some pre-experimental methodologies such as emoji and translation.

Paper [4]

The authors of the paper Enhancing Multilingual Hate Speech Detection: From Language-Specific Insights to Cross-Linguistic Integration argue that the issue of hate speech detection in low-resource languages crosses the the border “sternly against” the identification of violent speech in the broad sense. The authors develop methods i.e. classical ML, deep learning-based, and even transformer-based frameworks such as mBERT and XLNet. By relevance tuned preprocessing and hyperparameter configuration methods, the authors achieve high efficiency of the models deployed in the task. This paper achieved an appreciable performance upper bound, with F 1 scores 0.93 for German language and 0.95 for Roman Urdu. This is up to owe an increase of 10% as compared with corresponding baseline models in an under-resourced language model. The methodological emphasis will be placed on improving some of the preprocessing strategies, exploring prompt-based tuning and using generalised multilingual models to alleviate problem associated with under-resourced languages in this particular context. The research serves a standard of sorts when it comes to the measurement of hate speech across so many languages.

Paper[5]

The Dou et al. paper, entitled Sailor: Open Language Models for South-East Asia, addressed a challenging topic: the poor performance of language models on South-East Asian languages in being trained and biased towards an English-centric system. They brought forward the "Sailor" family of models (0.5B–7B parameters) being trained on continual pre-training on

datasets of SEA languages. Techniques like BPE dropout, code-switching, and aggressive data cleaning boosted model robustness. The models improved the language benchmarks of SEA by 10–20% in F1 scores. They surpassed Qwen1.5 and other baselines. The authors proposed the following directions for future work: extended language coverage, refined cross-lingual instruction, and better handling natural code-switching scenarios.

Paper [6]

The paper "SemEval 2024 – Task 10: Emotion Discovery and Reasoning its Flip in Conversation (EDiReF)" (2024) tackles challenges of emotion identification and reasoning about their shifts in dialogues in English and Hindi-English code-mixed. The authors launched two tasks for Emotion Recognition in Conversation (ERC) and Emotion Flip Reasoning (EFR), which are aided by annotated datasets. Tested are advanced techniques, namely, BERT and rule-based models for their F1-scores 0.70, 0.79, and 0.76, for three subtasks. But the implicit triggers of its challenges and limited cross-lingual adaptability exist, in any case. \

Paper [7]

This paper evaluates four popular language models—ChatGPT 3.5, ChatGPT 4, Gemini Pro, and LLaMA2—on multilingual sentiment analysis, focusing on ambiguous scenarios and their interpretation in 10 languages. This study reveals significant inconsistencies with the models' ability to handle nuances like sarcasm and irony. LLama2 has a positive bias, while results of Gemini and ChatGPT models are mixed over languages, showing problems with model safety filters and censorship. Future work will include correcting these biases and generally making these models more generalizable and interpretable for even better sentiment analysis in all sorts of scenarios.

Paper [8]

This paper discusses the various methodologies for sentiment analysis on Twitter's unstructured and heterogeneous data. It reviews machine learning techniques such as Naive Bayes, SVM, and Maximum Entropy in comparison with lexicon-based approaches. The study

highlights the benefits of preprocessing and distant supervision via emoticons for annotating datasets. Among the compared techniques, SVM with unigrams performed the best. Further corpus quality improvement and dynamic technique adaptation to changing Twitter contexts will be the future scope of this work.

Paper [9]

This paper presents the RoBERTa-BiLSTM hybrid model that will help in overcoming long dependencies, lexical diversity, and dataset imbalances while analyzing sentiment. The model makes use of RoBERTa for word embedding generation and BiLSTM for contextual semantics extraction from text. Experimental results show superior performance than the baseline models, which obtained an accuracy of 92.36% on the IMDb dataset and showed strong results on the Twitter datasets. Future work includes diverse datasets and fine-tuning hyperparameters to obtain enhanced performance.

Paper[10]

The paper, Sentiment Analysis of Customer Reviews of Food Delivery Services Using Deep Learning and Explainable Artificial Intelligence: Systematic Review, reviews the use of machine learning (ML), deep learning (DL), and explainable artificial intelligence (XAI) for analyzing customer feedback in the food delivery services (FDS) domain. It highlights the shift from traditional ML and lexicon-based methods to DL techniques like CNNs and LSTMs, which offer higher accuracy but face challenges due to their black-box nature. The study underscores the need for XAI tools such as LIME and SHAP to enhance model transparency and trust, addressing the interpretability gap in DL models.

Paper [11]

This paper presents the ArSa-Tweets model for sarcasm detection in Arabic tweets, considering challenges such as idiomatic expressions and sparse datasets. The research is conducted using state-of-the-art preprocessing and DL models that include LSTM, CNN- LSTM-GRU, and variants of AraBERT. ArSa-Tweets was determined to be best performed

with the AraBERT-V02 version, surpassing other methods over sarcasm datasets. For future work, the study will extend coverage of the datasets and further refine multilingual sarcasm detection capabilities .

Paper [12]

The paper, BiERU: Bidirectional Emotional Recurrent Unit for Conversational Sentiment Analysis, introduces a compact, parameter-efficient framework for sentiment analysis in conversations, addressing challenges in contextual encoding and sentiment classification. Unlike traditional party-dependent models, BiERU adopts a party-ignorant approach using the Emotional Recurrent Unit (ERU), composed of a Generalized Neural Tensor Block (GNTB) and a Two-Channel Feature Extractor (TFE). GNTB performs context compositionality by integrating contextual information into utterance representations, while TFE extracts emotional features using LSTM and CNN branches.

Paper [13]

The paper looked into the analysis of sentiments and hate speech in a code-mixed text containing more than one language. Results showed better performance for the fine-tuned bilingual model as opposed to multilingual and monolingual models; this was based on evaluation and development of bilingual LLMs for English-Hindi and English-Slovene. The future work focuses on the testing of new language combinations and model specializations towards code-mixed contexts.

Paper [14]

This review examines the development of sentiment analysis, focusing on deep learning approaches such as BERT and LSTMs. It identifies some of the challenges in sarcasm detection and adapting to various applications from business insights to disaster management. The authors suggest future research on multimodal data integration and ethical applications in sentiment analysis.

Paper [15]

The paper proposes an emotionally-neutral sentiment analysis framework, bypassing physiological emotional cues to rely solely on textual data. Integrated with LLMs, the model enhances sentiment detection accuracy and supports nuanced emotional interactions in AI applications. Future directions include expanding AI memory systems and context awareness for deeper human-AI engagement.

Paper [16]

This paper sheds light on the advancement of sentiment analysis powered by large language models (LLMs). Here, it discusses traditional methods, such as rule-based and lexicon methods, versus the current LLMs like BERT and GPT, which greatly advance deep learning-based contextual and semantic understanding. These include fine-tuning pre-trained models, transfer learning, and domain adaptation that support high accuracy and adaptability for various domains, such as e-commerce and healthcare. The challenges include biases, scalability, and privacy concerns, while future research emphasizes robustness, multimodal approaches, and ethical frameworks. The work highlights the transformative potential of LLMs in applications across industries.

Paper [17]

This paper is sought to study the issues that surround the sentiments of mixed linguistic texts which are as a result of language ⅼⵦ shifting, non – formal context. In this light, the authors evaluate the approaches, databases, as well as the models utilized in the analysis of sentiments with a focus on the issues of data shortages, the accuracy of models utilized (for example, F1 scores which are unsatisfactory), and the comprehension of the context in relation to code – mixing situations. They proposed future work including issues such as building adequate databases, everything is coupled with favorative multimodal methods (for example emojis or images), and addressing idioms and sarcasm. This paper can also be considered as an introductory paper for those interested in the topic of code-mixed text analysis.

Paper [18]

In the paper titled “Word Level Language Identification in Code-Mixed Kannada-English Texts using Deep Learning Approach” (2020) the authors seek to address the issue of language identification at the word level in heavily code-mixed Kannada English texts. To this aim, they suggest a Bi-LSTM model with attention which allows word tagging in many languages.

Paper [19]

The paper titled "A Sentiment Analysis Dataset for Code-Mixed Malayalam-English" (2020) addresses the lack of datasets for sentiment analysis in Malayalam-English code-mixed text, which is critical for natural language processing (NLP) in multilingual contexts. The authors created a gold-standard dataset of 6,739 posts, annotated with sentiment labels such as positive, negative, mixed feelings, and neutral. The dataset was derived from YouTube comments with a high inter-annotator agreement (Krippendorff's alpha > 0.8). Several machine learning models were tested (e.g., SVM, Logistic Regression), as well as deep learning models such as BERT. The best weighted F1-score was obtained for BERT at 0.75. The directions for further work are covering the dataset more widely and improving the annotations while further using the advanced NLP techniques. The dataset is made public along with its code.

Paper [20]

The paper "Sentimental analysis from imbalanced code-mixed data using machine learning approaches" by R. Srinivasan and C. N. Subalalitha in 2023 is focused on sentiment analysis for Tamil-English code-mixed datasets with considerable class imbalance. The methodology suggested makes use of preprocessing techniques like spelling normalization with the help of Levenshtein distance, recurrent character/word removal, and resampling techniques like SMOTE and ADASYN for handling imbalanced data. A set of machine learning classifiers such as Random Forest, Logistic Regression, and SVM are used for sentiment classification. It shows the improved macro F1-scores especially in the case of Logistic Regression which outperforms the other classifiers.

Paper [21]

The paper titled Multi-class Sentiment Classification on Bengali Social Media Comments Using Machine Learning attempted to solve the issue of carrying out multi-class sentiment classification of Bengali language, since that field lacks focus owing to the unavailability of data and resources. The authors introduce a classifier based on CNN and LSTM models, CLSTM, for social media comments classification into four different sentiments – acceptable, religious, political, and sexual. The study applies post before processing, TF-IDF and word embedding for feature extraction, and performance evaluation of initial ML models versus DL architecture. CLSTM performed best with 85.8% accuracy and 0.86 F1- score on the dataset containing 42,036 comments. The authors recommend the widening of the dataset combined with the implementation of advanced DL architectures in the goal of improving the sentiment analysis of the Bengali language.

Paper [22]

The document "A systematic review of hate speech automatic detection using natural language processing" (2023), critiques the ICT challenges of hate speech automated detection. A systematic review of published literature using PRISMA guidelines explains the progress that has been made in the field of natural language processing (NLP) and deep learning technology. This includes an analysis of a wide range of datasets as well as machine learning techniques such as supervised learning and unsupervised learning. The results pointed out the advantages of supervised learning techniques, particularly deep learning BERT models which made a considerable improvement in accuracy. Multilingual datasets that make it easier for the models to be trained, new strategies for hybrid deep learning methods, and improved versions of BERT that are specialized for hate speech detection are yet to be fully developed.

Paper [23]

The paper "Towards Kannada-English POS Tagged Code-Mixed Resources" (2022) focused primarily on a corpus of manually fueled annotation for emotion and part of speech (POS) scale for code-mixed Kannada-English text. The thesis was greatly attributed to the fact of lack

of annotated data sets on Kannada-English code-mixed texts especially in the emotion tag. The presented unique corpus is annotated with emotions with POS tags which are derived from captioned Kannada-English code-mixed tweets. For emotion prediction, machine learning algorithms like SVM and bi direction LSTMs model were used while the POS tagging required CRF, Bi-LSTM, and Bi-LSTM-CRF models. The study was instrumental in achieving relatively high accuracy in respect of these tasks, thus substantiating the annotated corpus during the research stage. Future work includes the expansion of the corpus to scenarios involving multiple languages, and a thorough examination of other bilingual code-mixed language pairs.

Paper [24]

The paper "Enhancing Named Entity Recognition in Low-Resource Silicon Language" (2024) illustrates the challenges brought about by limited labelled datasets in low-resourced silicon-based Indian languages such as Tamil, Kannada, Malayalam, and Telugu. The study examines the use of multilingual resources alongside cross-lingual models while relying on transformer models like mBERT, RoBERTa, and XLM-RoBERTa. The fact that accuracy in results could be greatly improved by merging datasets of such cross-linguistic pairs as Kannada-Tamil or Telugu-Malayalam shows how effectively cross-languages strategies can work. Future work in this area focuses on creating more domain-specific models, bringing in more Dravidian languages, and optimizing transfer-learning techniques so that they can be applied over broader areas.

Paper [25]

This paper deals with the issues of sentiment analysis and offensive language detection in multilingual code-mixed datasets, which are usually noisy, informal text styles, and mixed languages, making traditional analysis methods complex. Preprocessing techniques like removing stop words and emojis, and applying word embeddings-GloVe, BERT to the data for their preparation were employed. Deep learning models like Bi-LSTM, CNN, BERT, RoBERTa, Adapter-BERT for classification tasks are used. Here, Adapter-BERT performed best with 65% accuracy on sentiment analysis and 79% on offensive language. After that, the

paper concludes by proposing future work on multitask learning and advanced methods to amplify the performance of these tasks.

Paper [26]

The paper explores techniques to identify languages in code-mixed texts at the word level. A dataset called CoLI-Kenglish was developed using Kannada-English comments from YouTube videos, with words categorized into six classes: Kannada, English, Mixed-language, Name, Location, and Other. Word embeddings were developed by merging word, sub-word, and character vectors using the Skipgram model. Four models were developed in this study: CoLI-ngrams and CoLI-vectors (ML), CoLI-BiLSTM (DL), and CoLI-ULMFiT (TL). CoLI- ngrams attained the best macro F1-score of 0.64. Future work is to enhance the quality of the dataset, develop advanced features, and add morphological features to the embedding for better performance.

Paper [27]

The paper "A Transformer-Based Approach for Abuse Detection in Code-Mixed Indic Languages" (2022) targeted the task of abusive content detection in social media texts for 13 Indic languages. Here, the researchers compare classical machine learning models to transformer-based models such as XLM-RoBERTa, MurilBERT, and indic-BERT. According to this study, a combination of BiGRU, emoji embeddings with XLM-RoBERTa achieves the best F1 score with 0.88 and the best AUC score with 0.94. This model outperforms both baseline models and individual language-specific models. The authors conclude by suggesting an expansion of the dataset and more advanced preprocessing and modeling techniques for low-resource languages as future work.

Paper [28]

This paper "Machine Learning Techniques for Sentiment Analysis of Code-Mixed and Switched Indian Social Media Text Corpus: A Comprehensive Review" published in 2022 describes some challenges and approaches towards sentiment analysis of code-mixed Indian social media texts. This review is an elaboration on the various methods applied, highlighting

how the performances of deep neural networks such as BiLSTM surpass the performance of more conventional classifiers, like SVM and Naive Bayes. The study emphasizes the importance of annotated datasets, more robust language tools, and NLP models for code-mixed sentiment analysis. Research will be further directed toward statistical reviews of ML techniques and improving resources for the processing of code-mixed data.

Paper [29]

The study "Sentiment Analysis on Multilingual Code-Mixed Kannada Language" (2021) addresses the challenge of sentiment analysis in code-mixed Kannada datasets, a critical issue given India's multilingual population. Traditional NLP tools often fail to handle such data effectively. The authors explored various approaches, including machine learning models, a hybrid CNN-BiLSTM model, and transformer-based BERT implementations. The ktrain-based BERT model achieved the best results with the weighted F1-score obtained being 0.66 on the validation dataset and 0.619 on the test dataset. In fact, the study calls for bigger and better-quality datasets and more model optimization for its enhanced performance in dealing with texts code-mixed with other languages. This work significantly contributes to the field of multilingual sentiment analysis, especially in the case of Dravidian languages.

Paper [30]

The paper "KanCMD: Kannada CodeMixed Dataset for Sentiment Analysis and Offensive Language Detection" from the year 2020 fills in the gaps regarding a lack of datasets for sentiment analysis and detecting offensive language detection in code-mixed text in Kannada. The authors present the KanCMD dataset, which has 7,671 annotated YouTube comments labeled with category for sentiment analysis (positive, negative, neutral, mixed feelings, and other languages) and offense language detection. They benchmarked some machine learning models, including Logistic Regression and SVM, which achieved baseline F1-scores of both tasks at about 0.66 using the former. As such, their study highlights a good potential for promoting research in languages such as Kannada, known to be under-resourced, especially by exploring multiple-task learning and further improving its performance in order to process the code-mix text.

Chapter 3

Requirement Analysis and Specification

3.1 Functional Requirements

- **Language identification for code-mixed texts**:
    - Detects and identifies languages within code-mixed texts (e.g., Kannada-English, Hindi-English).
    - Supports multiple scripts and transliterated words.
- **Sentiment analysis of code-mixed texts:**
    - Classifies the sentiment (Positive, Negative, Neutral) of code-mixed texts.
    - Handles slang, emojis, and transliterations.

3.2 Non-Functional Requirements

- **Reliable and robust under heavy loads:**
    - Handles large volumes of data and complex inputs without failure.
- **Performance:**
    - Provides quick responses for real-time analysis.
    - Scales to support multiple users simultaneously.
- **Accuracy:**
    - High precision in detecting languages and analyzing sentiments
- **Usability:**
    - User-friendly interface with clear and simple output.
- **Security:**
    - Ensures data privacy and prevents unauthorized access.

3.3 Hardware Requirements

- High-performance computing resources, such as:
- GPU-enabled systems
- Server with multi-core CPUs
- 16 GB+ RAM
- 1 TB+ storage for dataset and model outputs

3.4 Software Requirements

- Python programming language
- TensorFlow or PyTorch for deep learning
- Hugging Face Transformers for LLMs
- Scikit-learn for traditional machine learning methods
- Libraries: NLTK
- Dataset tools: Scrapy for data scraping

3.5 Cost Estimation

- Hardware Setup: ₹3000 (GPU-enabled server, storage, networking)
- Cloud Services: ₹500/month for training and storage (depending on usage)
- Software and Licenses: Open-source tools (minimal software cost)
- Data Annotation Costs: ₹1000 (manual labeling)
- Miscellaneous Costs: ₹500 (additional tools and operational costs)
- Total Estimated Cost: ₹6000 (one-time and recurring costs)

Chapter 4

Design

4.1 High level design

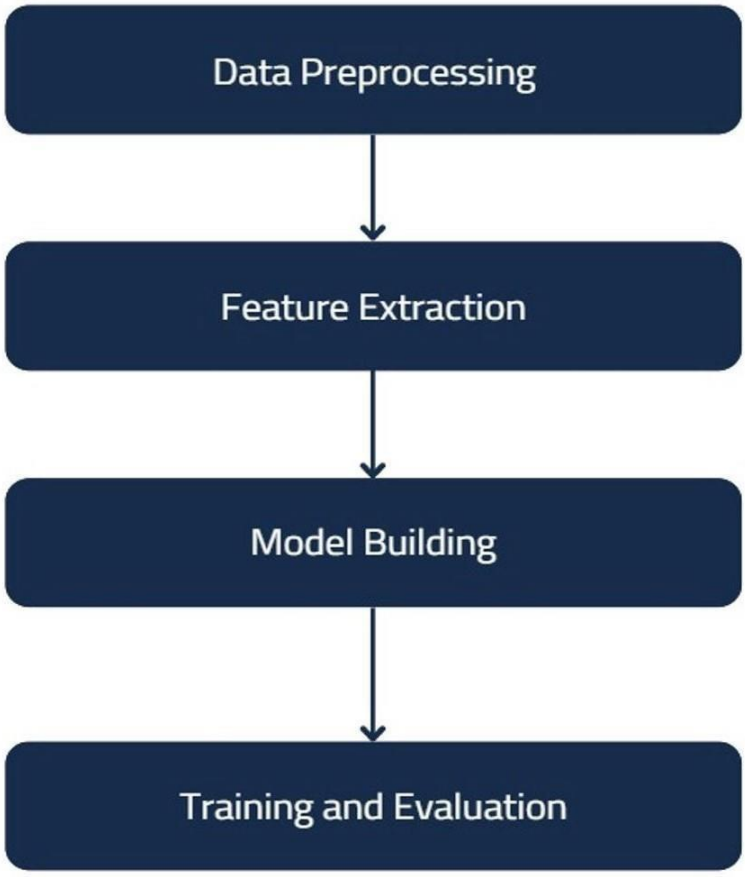


Fig 4.1 High Level Design

Figure 4.1 High-Level Design represents the overall workflow for the project, highlighting key stages: Data Preprocessing (cleaning and organizing raw data), Feature Extraction (deriving relevant features for analysis), Model Building (designing and training machine learning models), and Training and Evaluation.

4.1.1 System Architecture

The system architecture for sentiment analysis of code-mixed text involves preprocessing, feature extraction, and classification stages.

Preprocessing includes tokenization, normalization, and language identification.

Feature extraction uses techniques like Bag of Words, and embeddings.

Classification employs machine learning models (e.g., SVM, NB) and deep learning models (e.g., BiLSTM).

4.1.2 Abstract Specification of Sub-systems

Preprocessing Sub-system: Handles noise reduction, tokenization, and language tagging for code-mixed data.

Feature Extraction Sub-system: Converts textual data into numerical vectors using methods like embeddings and n-grams.

Classification Sub-system: Applies machine learning or deep learning models to classify sentiments as positive, negative, or neutral.

4.1.3 Interface Design

Input Interface: Accepts code-mixed text data, either from files, social media APIs, or manual input.

Preprocessing Interface: Displays cleaned and tokenized text, with identified language tags.

Output Interface: Visualizes sentiment classification results with metrics like accuracy, precision, recall, and F1-score.

4.2 Methodology

Step 1: Data Collection from platforms like social media (e.g., Twitter, Facebook) using APIs and scraping tools.

Step 2: Preprocessing for normalization, including spelling corrections, noise reduction, and tokenization.

Step 3: Feature Extraction through techniques like TF-IDF, embeddings, or lexicon-based analysis.

Step 4: Sentiment Classification using models such as SVM, or BiLSTM.

Step 5: Evaluation of model performance using accuracy, F1-score, and recall metrics

Chapter 5
Conclusion

In conclusion, this project focuses on enhancing language identification and sentiment analysis for code-mixed data, which is often found in informal communication, particularly in social media. Traditional systems like rule-based approaches and machine learning models such as SVM and Naive Bayes are limited in handling the complexity of code-mixed text. Deep learning models, like BiLSTM, provide better performance by capturing context and sequence-based patterns, but they still require large datasets and may struggle with code-switching.

The proposed solution involves using BiLSTM with attention mechanisms to improve language identification and sentiment analysis. Additionally, fine-tuning pre-trained multilingual models like mBERT and XLM-R can effectively handle multiple languages and better analyze mixed-language data. These models leverage extensive multilingual corpora to understand and process code-mixed text more efficiently.

Future work will include fine-tuning these models for domain-specific data, such as social media comments or online reviews, to improve their accuracy and generalization. By combining language identification with sentiment analysis in a single pipeline, this project aims to develop a robust system capable of handling the complexities of code-mixed language data, improving applications like social media monitoring, customer feedback analysis, and multilingual content processing.

References

[1] Ahmad, G.I., Singla, J., Anis, A., Reshi, A.A. and Salameh, A.A., 2022. Machine learning techniques for sentiment analysis of code-mixed and switched indian social media text corpus: A comprehensive review. International Journal of Advanced Computer Science and Applications, 13(2).

[2] Ansari, M.A. and Govilkar, S., 2018. Sentiment analysis of mixed code for the transliterated Hindi and Marathi texts. International Journal on Natural Language Computing (IJNLC) Vol,

[3] Astuti, L.W. and Sari, Y., 2023. Code-Mixed Sentiment Analysis using Transformer for Twitter Social Media Data. International Journal of Advanced Computer Science and Applications, 14(10).

[4] Hashmi, E., Yayilgan, S.Y., Hameed, I.A., Yamin, M.M., Ullah, M. and Abomhara, M., 2024. Enhancing multilingual hate speech detection: From language-specific insights to cross-linguistic integration. IEEE Access.

[5] Dou, L., Liu, Q., Zeng, G., Guo, J., Zhou, J., Lu, W. and Lin, M., 2024. Sailor: Open Language Models for South-East Asia. arXiv preprint arXiv:2404.03608.

[6] Kumar, S., Akhtar, M.S., Cambria, E. and Chakraborty, T., 2024. SemEval 2024--Task 10: Emotion Discovery and Reasoning its Flip in Conversation (EDiReF). arXiv preprint arXiv:2402.18944.

[7] Buscemi, A. and Proverbio, D., 2024. Chatgpt vs gemini vs llama on multilingual sentiment analysis. arXiv preprint arXiv:2402.01715.

[8] Kharde, V. and Sonawane, P., 2016. Sentiment analysis of twitter data: a survey of techniques. arXiv preprint arXiv:1601.06971.

[9] Md. Mostafizer Rahman, Ariful Islam Shiplu, Yutaka Watanobe, Member, IEEE, and Md. Ashad Alam

[10] Adak, A., Pradhan, B. and Shukla, N., 2022. Sentiment analysis of customer reviews of food delivery services using deep learning and explainable artificial intelligence: Systematic review. Foods, 11(10), p.1500.

[11] Abuein, Q., Ra'ed, M., Migdady, A., Jawarneh, M.S. and Al-Khateeb, A., 2024. ArSa-Tweets: A novel Arabic sarcasm detection system based on deep learning model. Heliyon, 10(17).

[12] Li, W., Shao, W., Ji, S. and Cambria, E., 2022. BiERU: Bidirectional emotional recurrent unit for conversational sentiment analysis. Neurocomputing, 467, pp.73-82.

[13] Yadav, A., Garg, T., Klemen, M., Ulcar, M., Agarwal, B. and Sikonja, M.R., 2024. Code-mixed Sentiment and Hate-speech Prediction. arXiv preprint arXiv:2405.12929.

[14] Sharma, N.A., Ali, A.S. and Kabir, M.A., 2024. A review of sentiment analysis: tasks, applications, and deep learning techniques. International journal of data science and analytics, pp.1-38.

[15] Ratican, J. and Hutson, J., 2024. Advancing Sentiment Analysis Through Emotionally-Agnostic Text Mining in Large Language Models (LLMS). Journal of Biosensors and Bioelectronics Research.

[16] Upadhye, A., 2024. Sentiment Analysis using Large Language Models: Methodologies, Applications, and Challenges. Int. J. Comput. Appl, 186, pp.30-34.

[17] Perera, A. and Caldera, A., 2024. Sentiment Analysis of Code-Mixed Text: A Comprehensive Review. Journal of Universal Computer Science (JUCS), 30(2).

[18] Yigezu, M.G., Tonja, A.L., Kolesnikova, O., Tash, M.S., Sidorov, G. and Gelbukh, A., 2022, December. Word level language identification in code-mixed Kannada-English texts using deep learning approach. In Proceedings of the 19th International Conference on Natural Language Processing (ICON): Shared Task on Word Level Language Identification in Code-mixed Kannada-English Texts (pp. 29-33).

[19] Chakravarthi, B.R., Jose, N., Suryawanshi, S., Sherly, E. and McCrae, J.P., 2020. A sentiment analysis dataset for code-mixed Malayalam-English. arXiv preprint arXiv:2006.00210.

[20] Srinivasan, R. and Subalalitha, C.N., 2023. Sentimental analysis from imbalanced code-mixed data using machine learning approaches. Distributed and Parallel Databases, 41(1), pp.37-52.

[21] Haque, R., Islam, N., Tasneem, M. and Das, A.K., 2023. Multi-class sentiment classification on Bengali social media comments using machine learning. International journal of cognitive computing in engineering, 4, pp.21-35.

[22] Jahan, M.S. and Oussalah, M., 2023. A systematic review of hate speech automatic detection using natural language processing. Neurocomputing, 546, p.126232.

[23] REDDY, A.A., 2022. Emotion Prediction and Parts-of-Speech Tagging Resources and Experiments for Kannada-English Code-Mixing (Doctoral dissertation, International Institute of Information Technology Hyderabad).

[24] Panchadara, K., 2024. Enhancing Named Entity Recognition in Low-Resource Dravidian Languages: A Comparative Analysis of Multilingual Learning and Transfer Learning Techniques. Journal of Artificial intelligence and Machine Learning, 2(1), pp.1-7.

[25] Shanmugavadivel, K., Sathishkumar, V.E., Raja, S., Lingaiah, T.B., Neelakandan, S. and Subramanian, M., 2022. Deep learning based sentiment analysis and offensive language identification on multilingual code-mixed data. Scientific Reports, 12(1), p.21557.

[26] Shashirekha, H.L., Balouchzahi, F., Anusha, M.D. and Sidorov, G., 2022. CoLI-machine learning approaches for code-mixed language identification at the word level in Kannada-English texts. arXiv preprint arXiv:2211.09847.

[27] Bansal, V., Tyagi, M., Sharma, R., Gupta, V. and Xin, Q., 2022. A Transformer Based Approach for Abuse Detection in Code Mixed Indic Languages. ACM transactions on Asian and low-resource language information processing.

[28] Ahmad, G.I., Singla, J., Anis, A., Reshi, A.A. and Salameh, A.A., 2022. Machine learning techniques for sentiment analysis of code-mixed and switched indian social media text corpus: A comprehensive review. International Journal of Advanced Computer Science and Applications, 13(2).

[29] Dutta, S., Agrawal, H. and Roy, P.K., 2021, December. Sentiment Analysis on Multilingual Code-Mixed Kannada Language. In FIRE (Working Notes) (pp. 908-918).

[30] Hande, A., Priyadharshini, R. and Chakravarthi, B.R., 2020, December. KanCMD: Kannada CodeMixed dataset for sentiment analysis and offensive language detection. In Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in social media (pp. 54-63).

APPENDIX A: Details of publications

Author Names: N/A

Paper Title: N/A

Name of the Conference or Journal: N/A

Place of the conference or Vol No. , Issue No. , Page No.s of Journal: N/A

Date of Conference or Date of Publication: N/A

APENDIX B: POs and PSOs Mapped

B.M.S. College of Engineering.

Department of Computer Science and Engineering.

Attainment of POs and PSOs

Batch no.: B56

Date: 27-01-2025

Project Title: Sentimental Analysis for Code-Mixed Language Using Deep Learning

PROGRAM OUTCOMES

PO	Level (3/2/1) 3-High 2-Medium 1-Low	Justification if addressed
PO1	3	Engineering knowledge: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.
PO2	3	Problem analysis: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.
PO3	3	Design/development of solutions: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.
PO4	3	Conduct investigations of complex problems: Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.
PO5	3	Modern tool usage: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.
PO6	3	The engineer and society: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.

PO7	1	Environment and sustainability: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.
PO8	3	Ethics: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.
PO9	3	Individual and team work: Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.
PO10	3	Communication: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.
PO11	1	Project management and finance: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.
PO12	3	Life-long learning: Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

PROGRAM SPECIFIC OUTCOMES

PSO	**Level (3/2/1)** **3-High** **2-Medium** **1-Low**	**Justification if addressed**
PSO1	3	Apply Software Engineering Principles and Practices to provide software solutions
PSO2	1	Design and Develop Network, Mobile and Web based Computational systems under realistic constraints
PSO3	3	Design efficient algorithms and develop effective code.

APENDIX C: Plagiarism report