PROPOSED PLAN
FOR

MAJOR PROJECT

2025

Presented To
Chandana Dev

A Proposal By
Saurav and Avinabh

SENTIMENT ANALYSIS OF ASSAMESE REVIEWS USING TRANSFER LEARNING AND DEEP LEARNING APPROACHES

PROBLEM STATEMENT

With the increasing availability of user-generated content in Assamese on social media, e-commerce platforms and news portals, understanding public opinion has become vital. However, Assamese is a low-resource language, and most state-of-the-art Natural language Processing (NLP) models are trained for English or other resource-rich languages.

The challenge lies in effectively classifying the sentiment (positive/negative) of Assamese text when limited labelled data (~2,000+ samples) is available.

OBJECTIVES

1. To preprocess Assamese text data (reviews) and prepare it for sentiment classification.
2. To implement Transfer Learning using state-of-the-art multilingual and Indic language models.
3. To implement deep learning from scratch using BiLSTM as a baseline.
4. To perform a comparative analysis between models and evaluate their performance on the dataset.
5. To identify the best-performing approach for sentiment analysis in Assamese.
6. To demonstrate the prediction of sentiment for unseen or unknown Assamese reviews.

SCOPE OF THE PROJECT

This project is focused on binary sentiment classification (Positive vs Negative). It is limited to Assamese language datasets (~2K+ reviews).

Comparative study between:

1. Transfer Learning models – MuRIL, IndicBERT and XLM-RoBERTa
2. Deep Learning model – BiLSTM (from scratch, with embeddings)

Applications:

1. E-commerce feedback analysis
2. Social media sentiment tracking
3. Regional NLP research for low-resource languages

METHODOLOGY OR WORKFLOW

The starting point of this project is preparing the dataset. The data comes in a CSV file containing three columns: ***ID*** (a unique identifier for each review), ***Review*** (the actual Assamese text), and ***Sentiment*** (the target label, either Positive or Negative). Since the *ID* column does not add value to model training, it is ignored during preprocessing. The main focus remains on the review text and its corresponding sentiment. Before moving forward, the dataset is cleaned by handling missing values, duplicate entries, or any noisy records that could negatively impact training.

After cleaning, the text preprocessing stage is applied. Assamese text can sometimes include unwanted punctuation, symbols, or inconsistencies in writing styles, so these are normalized. Tokenization is then carried out, ideally using sub-word-based tokenizers from Hugging Face, which are well-suited for low-resource languages like Assamese. To ensure effective training and evaluation, the dataset is split into three parts: training (around 70%), validation (15%), and testing (15%).

Once the data is ready, the baseline model is built using a **Bi-directional LSTM (BiLSTM)**. For this, pretrained word embeddings such as fastText (which supports Assamese) are used to represent the reviews in vector form. The BiLSTM processes these embeddings, and the model is trained to classify the reviews as Positive or Negative. This baseline provides a starting point to measure how more advanced models perform in comparison.

The next phase involves **transfer learning** with transformer-based models. Three pre-trained models are considered separately: **MuRIL** (by Google Research), **IndicBERT** (by AI4Bharat), and **XLM-RoBERTa** (by Meta AI). Each model is fine-tuned on the Assamese dataset by attaching a classification layer on top and training it to distinguish between positive and negative reviews. Their performances are then evaluated using key metrics such as accuracy, precision, recall, and F1-score.

Following this, a **comparative analysis** is carried out between the BiLSTM baseline and the transfer learning models. The aim is to highlight which model achieves the best results on the Assamese dataset, and why.

Along with the numerical scores, confusion matrices and other evaluation tools are used to gain deeper insights into model behaviour.

Finally, the best-performing model is selected for deployment. This model is used to predict sentiments for unseen Assamese reviews. For example, given the input "এই ছিবখন বহুত ভাল লাগিল" (meaning "I really liked this movie"), the model should correctly classify the sentiment as *Positive*.

Through this systematic workflow, the project demonstrates how both traditional deep learning and modern transfer learning approaches can be applied to Assamese sentiment analysis, and it identifies the most effective solution.

Evaluation Metrics

The models will be evaluated based on the following evaluation metrics:

1. Accuracy
2. Precision, Recall, F1-Score
3. Confusion Matrix

Expected Outcomes

1. A working sentiment classification system for Assamese reviews.
2. Performance comparison – showing transfer learning models outperform deep learning baseline.
3. A documented workflow and methodology useful for other low-resource language NLP projects.

Conclusion

This project demonstrates the effectiveness of transfer learning in low-resource language scenarios. While traditional deep learning models (BiLSTM) struggle due to limited data, pre-trained multilingual and Indic- specific models like MuRIL and IndicBERT are expected to achieve higher accuracy and generalizability.

The final system can serve as a foundation for regional NLP applications in Assamese and related Indic languages.