Bangla News Headline Classification and Sentiment Analysis using Bangla Bert

By
Israk Hasan Jone
212-15-4119
Badrul Alam
212-15-4144

FINAL YEAR DESIGN PROJECT REPORT

This Report Presented in Partial Fulfillment of the Requirements for the **Degree of Bachelor of Science in Computer Science and Engineering**

Supervised by

Dr. S.M Aminul Haque
Professor & Associate Head
Department of Computer Science and EngineeringDaffodil International University

Co-Supervised by

Ms Shahrin Khan

Lecturer
Department of Computer Science and EngineeringDaffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
Dhaka, Bangladesh

May 14, 2025

APPROVAL

This Project titled **"Bangla News Headline Classification and Sentiment Analysis using Bangla Bert"**, submitted by Israk Hasan Jone, ID No:212-15-4119 & Badrul Alam, ID 212-15-4144 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on **14 May, 2025.**

BOARD OF EXAMINERS

Dr. S.M Aminul Haque (SMAH) — **Chairman**
Professor and Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Mohammad Jahangir Alam (MJA) — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

14/5/25

Mr. Md Mohammad Masum Bakaul (MB) — **Internal Examiner**
Sr. Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

14.05.25

Dr. Md. Arshad Ali (DAA) — **External Examiner**
Professor
Department of Computer Science and Engineering
HSTU

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Dr. S.M. Aminul Haque, Professor & Associate Head,** Department of Computer Science and Engineering, Daffodil International University. We also declare that neither this projectnor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

Dr. S.M Aminul Haque (SMAH)
Professor & Associate Head
Department of Computer Science and EngineeringDaffodil International University

Co-Supervised by:

Ms Shahrin Khan
Lecturer
Department of Computer Science and EngineeringDaffodil International University

Submitted by:

Israk. 14.05.25.

Israk Hasan Jone
Student ID:212-15-4119
Department of Computer Science and EngineeringDaffodil International University

Badrul 14.5.2025

Badrul Alam
Student ID:212-15-4144
Department of Computer Science and EngineeringDaffodil University

ACKNOWLEDGEMENTS

This work would not have been possible without the support and contributions of many individuals over the past two semesters. We are deeply grateful to everyone who has as- sisted us in one way or another.

First, we express our heartfelt thanks and gratefulness to the almighty for His divine blessing making it possible for us to complete the **Final Year Design Project (FYDP)** successfully.

We are grateful and wish our profound indebtedness to **Dr. S.M. Aminul Haque**, **Professor & Associate Head**, Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh. Deep knowledge and keen interest of our supervisor in the field of **NLP** to carry out this project. His endless pa- tience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

We would like to express our heartfelt gratitude to the Head of the Department of Com- puter Science and Engineering, for his kind help in finishing our project and also to other faculty members and the staff of the Department of Computer Science and Engineering, Daffodil International University.

We would like to thank our entire course-mates at Daffodil International University, who took part in this discussion while completing the coursework.

Finally, we must acknowledge with due respect the constant support and patience of our parents.

ABSTRACT

As a core task of natural language processing, text classification is widely used in many fields. Reading newspapers in daily life is a very common practice, but before reading, everyone first check the newspaper headlines. The news headline is important as it is supposed to provide an efficient way to grasp the flavor of the article and acts as a key factor to determine readers' attitudes toward the article. In natural life, people naturally classify news in terms of themes and emotions using only headline impressions. However, in the age of constant and unorganized inflowing of digital news, manually clustering news by category and sentiment is a cumbersome job, particularly for Bangla news in which scarcity of automatic tools is found. Filling up this void, in this paper, we propose a state-of-the-art deep learning-based than we developed A Dual headed Classification model which Bangla News Headline Classification and Sentiment Analysis using transformer models. An end-to-end classification model was proposed using the BanglaBERT model to categorise (e.g., politics, religion, sports, other), and predict the sentiment polarity (positive, negative or neutral) of Bangla news headlines. Results were based on a dataset of 5033 training samples and 516 testing samples. Experimental results showed superior performance (with training accuracy of 98.18% and testing accuracy of 84.38% for aspect classification, and training accuracy of 97.17% and testing accuracy of 73.26% for sentiment analysis). Though there was a little bit of over fitting since the data set was small, the results clearly show strong potential in using pre-trained Bangla specific transformers in automated headline classification task. In the future, we will work on enlarging the dataset, further integrating data augmentation, and issuing real-time web applications for practical use of the system.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

Now a days, natural language processing (NLP) has received a lot of attention due to its wide range of applications in various fields. Among the various tasks in NLP, text classification and sentiment analysis are his two key tasks that have been extensively studied. Text classification is the process of classifying text documents into predefined classes or categories, while sentiment analysis is the process of determining the sentiment or polarity of a given text document.

In Bangla, text classification and sentiment analysis are becoming more and more important as a large amount of Bangla text data is generated online every day. Especially Bangla headlines are an important source of information for people in Bangladesh and around the world.

In recent years, the use of deep learning-based models such as Bert(Bidirectional Encoder Representation from Transformers) has shown remarkable performance on NLP tasks. Bert is a pre-trained language model that can be tailored to his specific NLP task, delivering state-of-the-art performance.

In this work, we use Bangla Bert to propose a headline classification and sentiment analysis system for Bangla news headlines. We first preprocess a dataset of Bangla news headlines, then use the dataset to fine-tune a Bangla Bert model to classify headlines into different categories. Also, set the mood for news headlines at Bangla Bert. Finally, here evaluate the performance of the proposed system using various metrics such as accuracy, precision, recall, and F1 score.

1.2 Motivation

In the present day digital era, exponential growth has been observed in the world of online news content specifically in regional languages like Bangla, 7th most spoken language in the world, leading to the need for automated systems that can interpret and classify text based data. Bangla news sites produce a vast amount of the news every day, so it is difficult for readers and organizations to read all the news correctly.

Despite several advancements achieved in NLP in the last few years for English, there still exists a gap with respect to the availability of robust NLP models and resources for Bangla. The majority of current Bangla NLP tools are based on conventional machine learning and do not have a good enough sense of context for more fine grained tasks like sentiment analysis and topic classification.This research is inspired by the prospects of BanglaBERT; a transformer-based pre-trained model that has been fine-tuned using native Bangla corpora types of data, to mitigate these issues. The objective is to enhance the design of multi-task learning based architecture to overcome poor efficiency and accuracy in Bangla NLP system which can be used in news, political survey, opinion analysis, content management.

In addition to promoting the state-of-the-art in Bangla NLP, the success of this dual-head approach could pave the way for further investigations in other low-resource languages employing the same strategy.

1.3 Objectives

Bangla news headline classification and sentiment analysis is one of the research topic that we mainly focus in this paper. This paper's main goal is to establish an optimal Bangla news headline method classification and sentiment analysis methods by utilizing state of the art natural language processing (NLP). Following are the key research objectives of Contraceptive Drugs Market Report:

1. Create a dual-task framework for news headline topic identification andsentiment analysis on Bangla language.

2. Use BanglaBERT model, which is a transformer based model, to achieve better classification accuracy compared to traditional machine learning or deep learning apps.

3. Improve sentiment prediction by using aspect based sentiment, i.e., classify the headline into categories of interest, before doing polarity classification.

4. How to extract indicative linguistic features of Bangla news headlines for a better classification and prediction of a headline such as sentiment.

5. To prove the superiority over traditional ML/DL based models, compare the performance of the proposed BanglaBERT model with them.

6. Propose solutions to enhance Bangla NLP research by identifying the primary limitations of existing models, including low accuracy, limited datasets, and an absence of contextual understanding.

1.4 Methodology

This study adopts the transformer-based deep learning model in the form of BanglaBERT, a pre-trained BERT model that is tailored for the Bangla language. The proposed methodology is divided into a number of sequential steps:

Data Collection and Preprocessing, We develop a labeled dataset of Bangla news headlines with respective aspect category and sentiment polarity as positive, neutral, and negative. After removing unwanted characters, the text is cleaned, preprocessed, and label encoded. BanglaBERT tokenizer is employed to tokenize the headlines. News are padded and truncated to at most 40 tokens to be matched across batches.

Dual-Head Model Architecture. The base encoder is the BanglaBERT model. Through two distinct FC layers, the [CLS] token representation (pooled output) is propagated from the final transformer block. One for Aspect Classification, and One for Sentiment Analysis.

Training Process model is fine-tuned with a joint loss — categorical cross-entropy for the two classification heads. Optimization is carried out using the AdamW optimizer and the learning rate scheduler. The model is trained for 10 epochs using splits for training and validation. Testing the returned model is tested on a different test set. We measures accuracy, precision, recall and F1-score in both tasks. Also confusion matrices as well as classification reports are created.

Prediction & VisualizationPrediction You can classify any Bangla headline using the prediction function. Also, performance curves are drawn for visualized view of model learning pattern wrt the epochs. This approach enables the joint task of topic and sentiment classification using the shared transformer encoder achieving both effectiveness and computational efficiency in the context of Bangla language understanding.

1.5 Project Outcome

The major contribution of this work is a strong and efficient dual-task classification model applied using BanglaBERT that could classify aspects and sentiment of Bangla news headlines simultaneously. The model also demonstrates high accuracy and F1-scores for both the tasks which indicates its usage in the understanding of linguistic structure and semantics in Bangla.
Also, the project includes:

1. A Bangla news headline dataset with aspect and sentiment labels.
2. A two-head architecture using BanglaBERT, a pretrained transformer model.
3. Reports and visualizations (accuracy, confusion, classification).
4. A prototype system to predict the headline in real time.
5. A benchmark of comparison against other models such as LSTM or classical ML classifiers.

This work adds to the emerging area of Bangla NLP by providing an initial model that can be applied to longer news articles or other downstream applications in media analysis, public opinion mining, intelligent news filtering.
This work adds to the emerging area of Bangla NLP by providing an initial model that can be applied to longer news articles or other downstream applications in media analysis, public opinion mining, intelligent news filtering.

1.6 Organization of the Report

Chapter 1: Introduction
Bangla news processing Context of the work This chapter provides introduction to the Bangla news processing domain, with specific focus on challenges of classifying headlines and predicting sentiment in Bangla text. It provides a rationale and a description of the goals and expected results of the research, stressing the relevance of using state-of-the-art NLP methodologies to dual-task classification.

Chapter 2: Background
The basic ideas of Natural Language Processing (NLP), Bangla text processing, headline classification, and sentiment analysis are covered in this chapter. It also includes a summary of related work from English and Bangla contexts, gaps in existing approaches, and places this research in the context of multilingual NLP overall.

Chapter 3: Methodology of the Study
This chapter presents the methodology that was used in this study. It consists of the dataset preparation, label encoding, text preprocessing and models architectures adopted. BanglaBERT is introduced in the dual-head classification method and LSTM and RoBERTa models are included for comparative study. The chapter also reviews training schemes, performance metrics and choice of hyper-parameters.

Chapter 4: Implementation and Results
The methodologies and experimental setting-up for the defined models are described in this chapter. It provides the training, validation and testing overview. Performance is assessed by accuracy, precision, recall and F1-score. This chapter contains classification report, confusion matrix, comparative study of BanglaBERT, LSTM and RoBERTa models, etc..

Chapter 5: Engineering standards and design challenges
This chapter describes the commitment to engineering standards, ethical issues, and engineering design challenges associated with model development. It works on dataset imbalance, computational burden and ambiguous Bangla text issues.

Chapter 6: Societal, Environmental and Sustainable Implications
This chapter discusses the wider influence of automated Bangla text classification in media and journalism. It presents how such systems can be useful in enhancing access to sentiment-aware content, mitigating misinformation, and, prompting digital inclusion among Bangla speakers. It also delves into sustainability and ethical AI practices.

Chapter 7: Conclusions and Future Work
Conclusion In this chapter, this paper compiles a conclusion of the research done and how well the BanglaBERT model performed in dealing with two level of 2 different classification task in Bangla reviews. It describes next steps including enlarging the dataset, feeding the model to web applications, and adapting it using domain-specific corpora with the goal of increasing model quality

Chapter 2

Background

2.1 Introduction

With the proliferation of internet and social media, the amount of Bangla text available on the internet, particularly as news articles, has increased substantially. Due to the massive number of news articles produced everyday, the task of automatic classification and organization of news articles is an open problem.

In the recent years, we have witnessed an unprecedented proliferation of digital media and online news portals that generate terabytes of textual news stories originating in languages like Bangla. Being one of the most spoken languages and used by more than 250 million people across the world Bangla has huge digital news collection but has a reasonable amount of digital news generation. User-generated content, on the other hand, is being produced more frequently than ever, the same of the news updates, making it obligatory to build machine learning systems for automatic classification and analysis of Bangla text. Headline classification and sentiment analysis are two prominent tasks of in Natural Language Processing (NLP) used for organizing, filtering, and comprehending the emotional attitude of news. Despite the many advances for technology for English and other resource-rich languages, Bangla continues to be a low-resourced language as far as NLP is concerned. This has resulted in an apparent void of research and applications in the area of Bangla text analysis with a focus on dual-task systems that are capable of handling both topic classification (aspect) and sentiment polarity detection simultaneously.

Although traditional machine learning techniques find it difficult to handle the complicated morphological and syntactic nature of Bangla. But, with the advent of transformer-based architectures such as BERT (Bidirectional Encoder Representations from Transformers), NLP has changed. Pretrained on corpus-level objectives (large-scale copora), current models outfit with these prelearned structures can outperform state-of-the-art (SOTA) systems across

many NLP tasks, when results on a held out test set are considered, by fine-tuning for specific tasks downstream.

We investigate BanglaBERT, a transformer-based language model trained on Bangla corpora, to categorize Bangla news headlines based on aspect (e.g., politics, religion, sports, etc.) and sentiment (positive, negative, neutral) in a multi-task learning manner. By deploying and comparing the Performers with traditional models like LSTM and other multilingual transformers like RoBERTa, we are putting hands around the bottlenecks of Bangla NLP and moving towards the state-of-the-art of intelligent news classification systems in Bangla.

2.2 Literature Review

Table 1: Summary of Literature Reviewed.

Author's names and Year	Technique used	Accuracy	Limitation
Amran Hossain, Niraj Chaudhary, Zahid Hasan Rifad. [2021]	LSTM(Long Short-Term Memory),Gated Recurrent Unit (GRU)	GRU 87.48%, LSTM 82.74%	Low accuracy large amounts of data only news category.
Sheikh sadi,Sbid,Shaklian [2023]	BiLSTM AND GRU	BiLSTM 83.42% GRU 80.01%	English language Low accuracy, news Classification.
Khondoker Ittehadul Islam, Md. Saiful Islam, Md Ruhul Amin [2022]	Fine-tuned multilingual BERT	multilingual BERT 71%	Binary sentiment classification
Arid Hasan,sudipta das, Afiyat anjum, firoja alam [2023]	Fine-tuned monolingual transformer-based models	Transformer-based models 85%	Only sentiment classification

A. Rahman and M.S. Hossen [20], 2018	Machine learning algorithms (Naive Bayes, SVM, Decision Tree)	77.5% accuracy	Limited to movie reviews and low accuracy
Ekramul,Labib ,Fisal , Shazzed [2023]	Analysis dataset	Fleiss' kappa score of 0.88	Only analysis not predication
Saman Sarkar Joy,Aishi,Naima, Annajit alim [2023]	BanglaBERT and MPL Techniques	BanglaBER Accuracy 82.25%	Clickbait Detection,use pre trained model
Saha et al. [2022]	Hybrid approach that combine BiLSTM and CNN	Accuracy 89.89%	sentiment classification
Istiak,Fahad,Rashed [2022]	Use various model LSTM,BiLSTM ,CNN SVM,RF GRU	Accuracy 91.8%	8 Category headline classification task.
karim [2019]	Multichannel convolutional-LSTM network	79.9% accuracy on Bengali movie reviews dataset	Use small dataset , accuracy is not good.

2.2.1 Similar Applications

Web and Mobile Applications
A few prototype level web apps and APIs for Bangla sentiment analysis have been developed using Rule based or shallow learning approaches. But no one of them combines dual-head aspect and polarity classification into a single model. Fill the Missing Piece Your work fills this gap by developing an intelligent app (based on Streamlit) that utilizes BanglaBERT to make real-time prediction from user input or web scraped headlines.

Headline Categorization in Different Languages.
Headline Classification Many works have taken recourse to conventional machine learning or LSTM based models for headline classification for English, Hindi and Chinese. For example, the study cite{gaikwad2018comparative} 'A Comparative Study on Headline Classification using Deep Learning Techniques' used CNNs and LSTMs to classify English news articles in to sports, politics, business, etc. These models, while achieving state-of-the-art performance in high-resource languages, cost down especially in the case of Bangla due to a complex linguistic structure and less resource availability.

Study of TF-IDF and BERT Classification for Bangla Text Classification This work contrasts the traditional TF-IDF-based models with the ones based BERT, on classifying Bangla newspaper articles. Different types of ML algorithm including the Decision Tree, Random Forest, SVM and Logistic Regression are adopted, which demonstrates the superiorities of transformer-based embeddings.

A Deep Learning Based Bengali News Headline Categorisation.
In this paper, deep learning algorithms BiLSTM and GRU classify Bangla news into six categories. The results of the study show the performance of BiLSTM to be 83.42% in accuracy which is better than GRU.

Sentiment Analysis in Bangla using Multilingual BERT with Transfer Learning.In this paper, we release manually annotated 2-class and 3-class sentiment analysis datasets for the Bengali and confirm that multilingual BERT models after fine-tuning are able to enhance the performance of sentiment classification tasks. The accuracy for 2-class sentiment classification is 71%.

RSM-NLP at BLP-2023 Task-2 Bangla Sentiment Analysis with EHN and XLM-RoBERTa
This method is tuning a number of multilingual and pre-trained BERT-based models on sentiment analysis specifically on Bangla social media posts. The combination of the majority voting in the ensemble model performs better than using the individual models alone, resulting in 0.711 on a shared task.

2.2.2 Related Research

Use of Natural language Processing (NLP) to the Bangla language has achieved significance in the past few years, especially in the fields of sentiment analysis and text categorization. Several works have investigated using transformer-based models like BanglaBERT for improving these tasks.

Another paper contrasts standard TF-IDF-based models with BERT-based models in the task of classifying Bangla newspaper articles. The study employs a range of machine learning algorithms such as Decision Tree, Random Forest, SVM, and Logistic Regression thereby demonstrating the beneficence of transformer-based representations in text classification. In addition, a hybrid system, consisting on the combination of a rule-based algorithms and BanglaBERT has been experimented on for improving sentiment analysis in Bengali texts. This approach fuses a lexicon-based system and a pre-trained language model, which leads to more accutrate and finer-grained (nine sentiment categories) sentiment prediction.

BanglaBERT and Its Derivatives.
Bangla BERT Model Pre-trained for Bangla Language Sagor Sarker's BanglaBERT is a pre-trained language model built for the Bengali language! Fine-tuned on a massive corpus of Bengali natural language, such as Wikipedia and Common Crawl data, training BanglaBERT has shown remarkable performance characteristics for tasks such as sentiment analysis, hate speech and news topic classification. Apart from BanglaBERT, several task-oriented models, such as BanglaClickBERT have been proposed. BanglaClickBERT is a pre-trained clickbait detection model for the Bangla news headlines that has been fine-tuned on about 1 million unlabelled news headlines to make it work better.

Shared Tasks and Competitions
The BLP-2023 Task 2 Entailed Bangla Social Media Posts Sentiment Analysis. The participants had used a range of approaches to improve performance, including optimising BanglaBERT models. The LowResource group secured the third position among 30 teams with BanglaBERT using methods such as random token dropping and ensemble modeling in its strategy. On the other hand, the RSM-NLP team used fine-tuned transformer models and yielded significant results in sentiment analysis task to show the effectiveness of the pre-trained models in data-scarce environments.

The reviewed articles highlight the increasing interest and advancements in the use of transformer-based models for Bangla NLP tasks. Although it has achieved good results as an aspect classification, how to map the results of the aspect classification model to the sentiment analysis is the remaining bottleneck to be solved for unified sentiment analysis. This'll be addressed in this research work by introducing a dual-headed classification model with the help of BanglaBERT that can do aspect and sentiment analysis simultaneously on Bangla news headlines.

2.3 Gap Analysis

Although there has been substantial advancement in NLP for high-resource language, Bangla still lags behind in news headline classification and sentiment analysis. Although sentiment analysis in Bangla has been studied as a topic in itself, a holistic solution which can jointly deal with aspect classification and sentiment polarity, all at once on news headlines using state-of-the-art model like BanglaBERT is not present in the literature.

Identified Gaps

I. Small Attention to Dual Classification Tasks: Previous studies mostly focuses on sentiment analysis or topic classification individually. Very few research works have been presented till now which combine both the tasks together to get a comprehensive insight of Bangla news headlines.

II. Underused Pre-trained Transformer Models: Although BanglaBERT has shown effectiveness on a few NLP tasks, it has not exploited yet in dual classification of news headlines to perform both aspect and sentiment.

III. Limited Domain-Specific Datasets: The absence of large annotated appropriately domain specific dataset for Bangla news headlines is a hindrance in building and evaluation of the strong models for dual classification.

IV. Real-time End-user Applications which are User-friendly: There is a lack of such What You See Is What You Get (WYSIWYG) end-user applications for these models in which you can input Bangla text and get immediate categorization in a human readable way.

Our Contribution.

To fill these gaps, we here propose a dual-headed classification model based on the BanglaBERT, which is able to conduct both aspect and sentiment analysis on Bangla news headlines. We also implement an easy-to-use web interface to do real-time predictions, which makes our model more practical. This unified approach both furthers computational understanding of Bangla news content, and offers a useful resource for journalists, scholars, and the public to explore and analyse the media.

2.4 Summary

This chapter presented an in-depth description of the research domain covering the increasing necessity of intelligent systems for processing and analyzing Bangla news headings. It started by presenting the exploding world of Bangla digital content and the importance of headline classification and sentiment analysis in structuring this information. The review of literature covered a significant number of existing works and identified their contributions and the limitations – such as non-existence of any dualtask classification model with respect to Bangla processing.

The other similar works employing traditional machine learning and deep learning models achieved high performance in other languages, and when it comes to Bangla, they are unsatisfactory due to its complexities of linguistic structure and resource scarcity. Meanwhile, similar studies have shown that although models such as BanglaBERT exhibit good performance in single NLP tasks but lack coherence to both aspect and sentiment classification during Bangla headline downstream analysis.

From the gap analysis, the major shortcomings in the state of the art were

found to be: Non-availability of dual-headed classification systems; Dominance of standard vanilla RNN/LSTM architecture models compared to pre-trained transformer models in the context of Bangla; Lack of suitable domain-specific dataset and absence of projects in live real-time situation. In aim to fulfill these gaps, here we suggest a dual-head BanglaBERT-based classification model and also developed a web portal to make this application more users friendly and practical. This chapter is intended to be a prelude to the next chapters of the book that further elaborate the proposed approach, the development of the model, and its evaluation.

Chapter 3

Research Methodology

3.1 Methodology/Requirem ent Analysis & Design Specification

In this chapter, a detailed description is provided on designing and methodology for building a dual-headed classification model for classifying aspect-level and sentiment-polarity in Bangla news headlines. It includes the structure of the BanglaBERT model, data processing methods, hardware/software configurations and the motivation of every design choice. The work adopts a full pipeline that covers all stages from gathering and cleaning of news articles in Bangla, the fine-tuning of a pre-trained transformer—BanglaBERT—and the simultaneous classification of input text with two different but linked tasks: aspect classification and sentiment analysis. Additionally, the section presents analysis on training settings, evaluation metrics and the model's performance relative to baselines like LSTM and RoBERTa.

3.1.1 Overview

This section gives a brief introduction to the research methodology and system design employed in the project. We aim to build a robust and lightweight NLP system that can effectively classify news headlines of Bangla news articles into aspect category (i.e.,politics, religion, sports, others) and also predict the related sentiment (i.e.,positive,neutral and negative) based on a single model.
The model is based on BanglaBERT (sagorsarker/bangla-bert-base), a transformer-based language model pre-trained on Bangla text. The system design comprises ,a data preprocessing pipeline which does text normalization, label encoding, tokenization, and padding.

A dual headed classifier model with the shared BanglaBERT encoder and two fully connected layers:

One for aspect classification

A second for polarity classification

Cross-entropy loss of both heads trained together.

Custom evaluation loop to calculate accuracy, F1-score, precision, and recall for both tasks.This two-task design can save computational cost and enhance the performance by sharing feature representations, and it is more effective than feeding into two individual models.

3.1.2 Proposed Methodology/ System Design

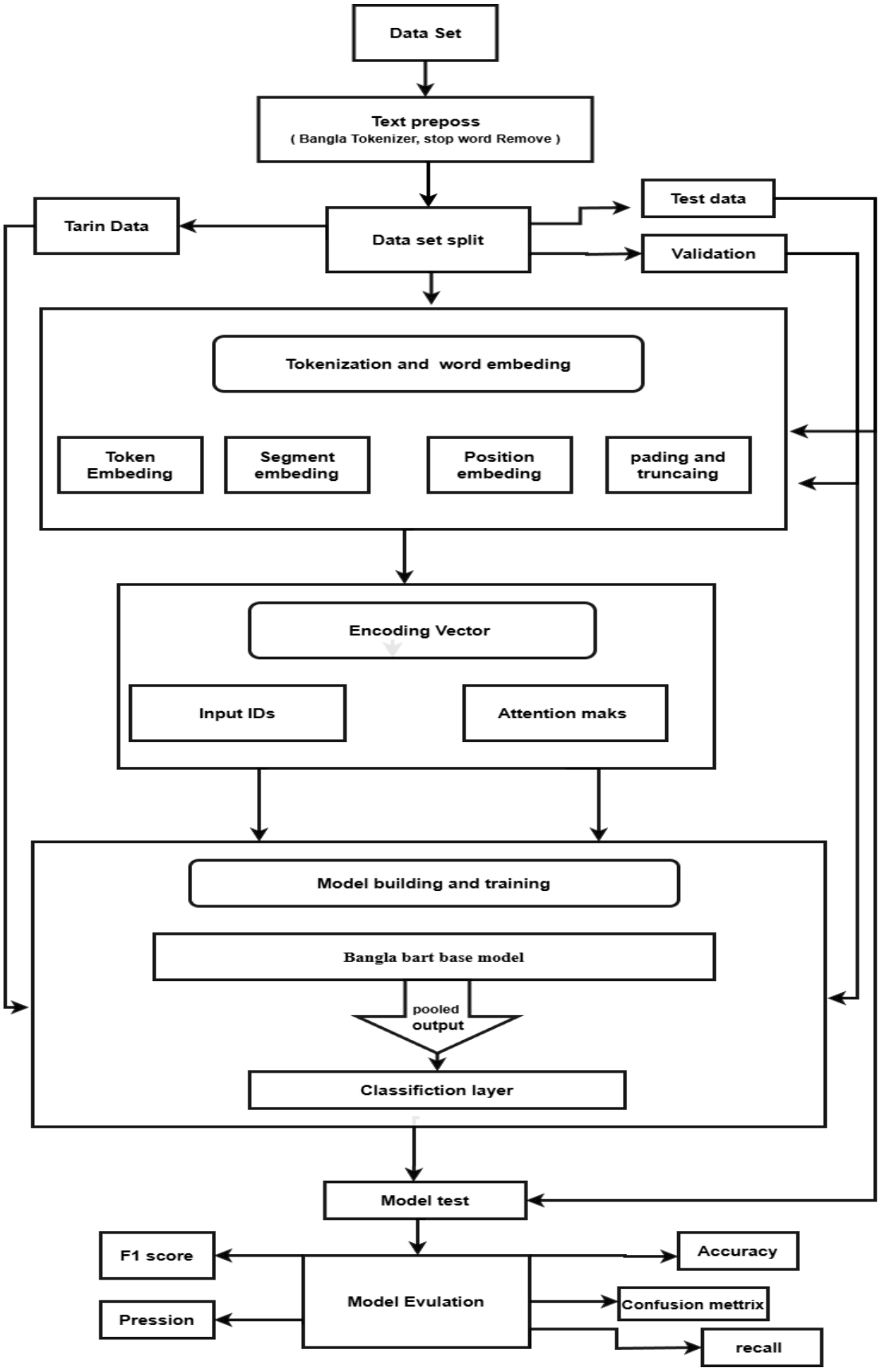

Figure 3.1: Architecture Of proposed Methodology.

3.1.3 Functional and Nonfunctional Requirements

Functional Requirements (FR)

1. Input Text: Take the Bangla news headlines as input for analysis.
2. Preprocessing: Tokenization and Normalization of Bangla text
3. Dual Classification: To classify aspect and polarity at the same time.
4. Model Application: Apply BanglaBERT and get both aspect and sentiment simultaneously.
5. Show Predictions and Confidence: Display predictions and their confidence.
6. Exportable: Let people export their results (e.g., CSV).
7. Web UI: Create a Streamlit-based user interface as an easy way to access the service.

Non Functional Requirements

1. Performance: fast for prediction on single and batch inputs.
2. Scalability: Supports working in large collection of inputs.
3. Security: Protect the data and users securely.
4. Usability: Interface must be easy and Bangla compatible.
5. Reliability: System should work with few outages.
6. Meintainability: We can easily update model, also UI component
7. Compatibility: Easily Work on all major browsers and devices.

3.1.4 Context Diagram

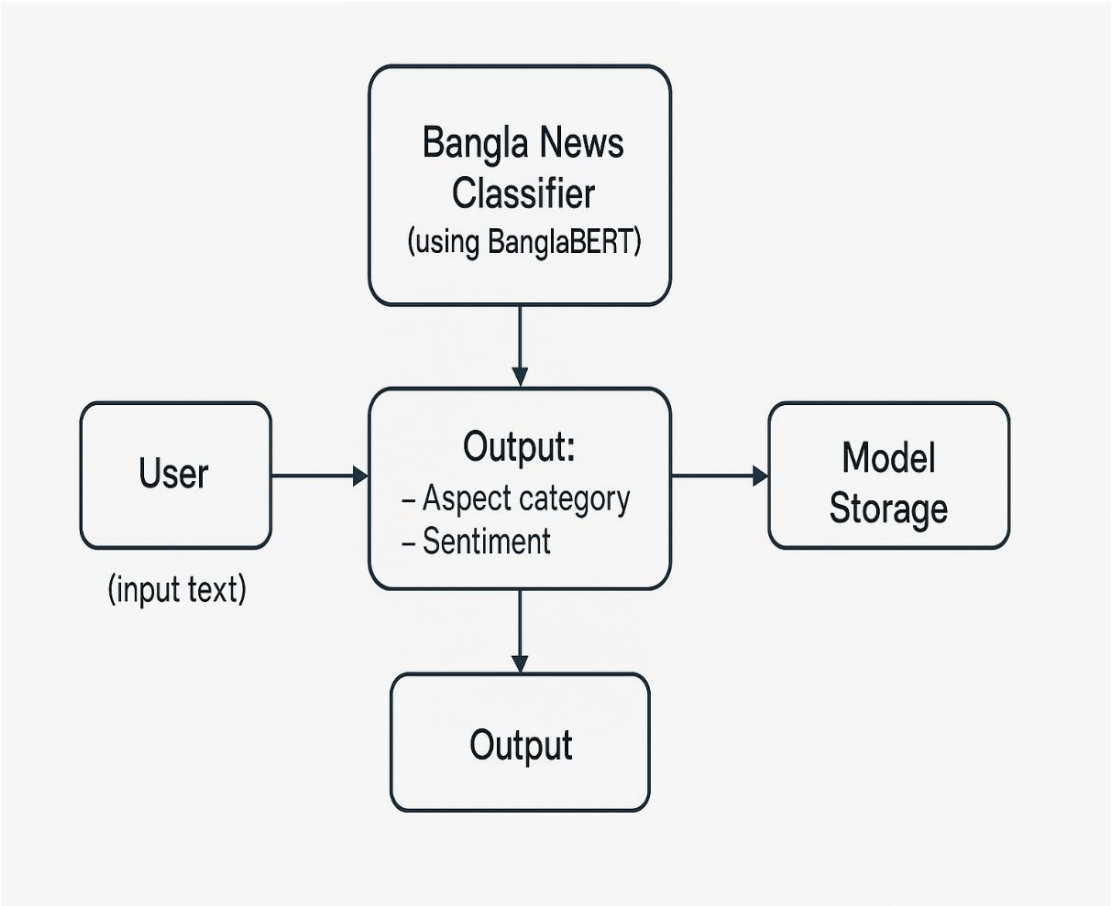

Figure 3.2: Context diagram

3.1.5 Data Flow Diagram Level

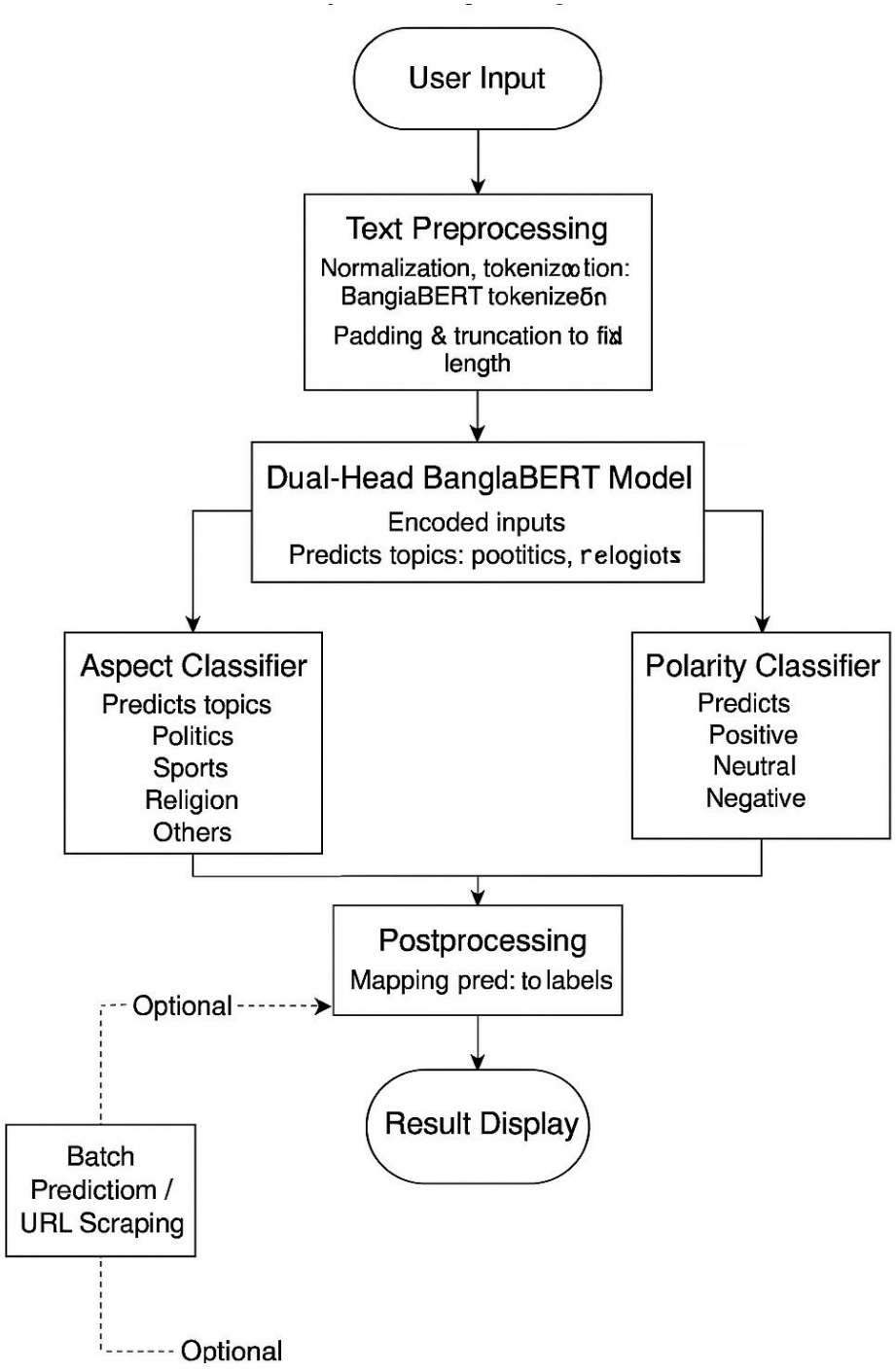

Figure 3.3: Data Flow diagram

3.1.6 UI Design

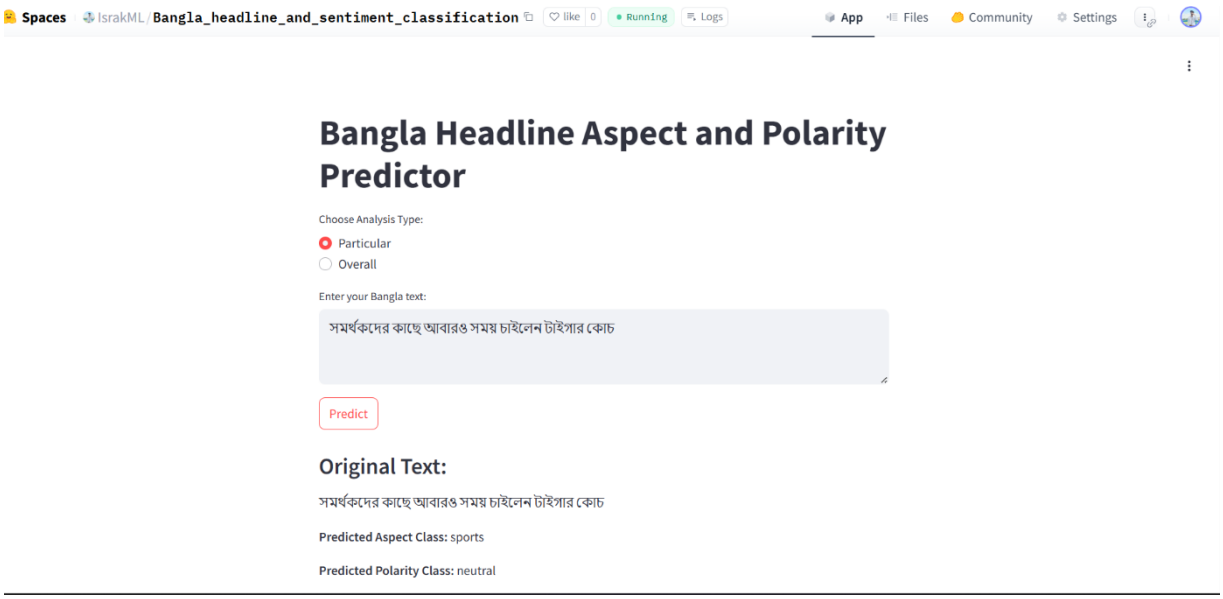

Figure 3.4 App UI Design

3.2 Detailed Methodology and Design

3.2.1 Data Collection

Data collection is a process of gathering data. This is a process of gathering, approximating, and evaluating accurate insights for research applying accepted protocols. Data collection is the most important and primary step for any research work. For the research purpose, we collect data from kaggle. It is the open-source platform for download any dataset we want. we used BAN-ABSA dataset downloaded from kaggle.com, which containing 9k news headline consisting of 3 columns post, aspect and polarity. Post column indicate the news headline in string, aspect column indicates the headline classes or types in string category (politics, other, religion and sports), polarity column indicates the headline sentiment in string category (positive, negative and neutral). We use this dataset to training, testing and evaluating our model.

3.2.2. Dataset Description

The Kaggle dataset consisted of a few thousand Bangla news headlines along with two labels, namely, aspect category (which could be politics, sports, religion, and others) and sentiment polarity (which could be positive, negative, and neutral). The data was divided into training, validation, and test sets to support model development and evaluation: approximately 80% for training, 10% for validation, and 10% for testing. Although the data was varied and versatile, there was some

class imbalance, especially for neutral sentiments. In general, the dataset was a solid base in the construction of the dual- head classifiers.

3.2.3. Data Preprocessing

Before training the model, the Bangla headlines were preprocessed in several steps. Special characters, additional spaces and meaningless symbols were removed from texts while preserving Bangla language content. The headlines were tokenized using BanglaBERT tokenizer and padded and truncated into a maximum of sequence length for standardization. Label encoding was performed to transform the categorical class labels to numeric format. These pre-processing steps ensured that the input was standardized, provided a more stable training and increased the performance overall.

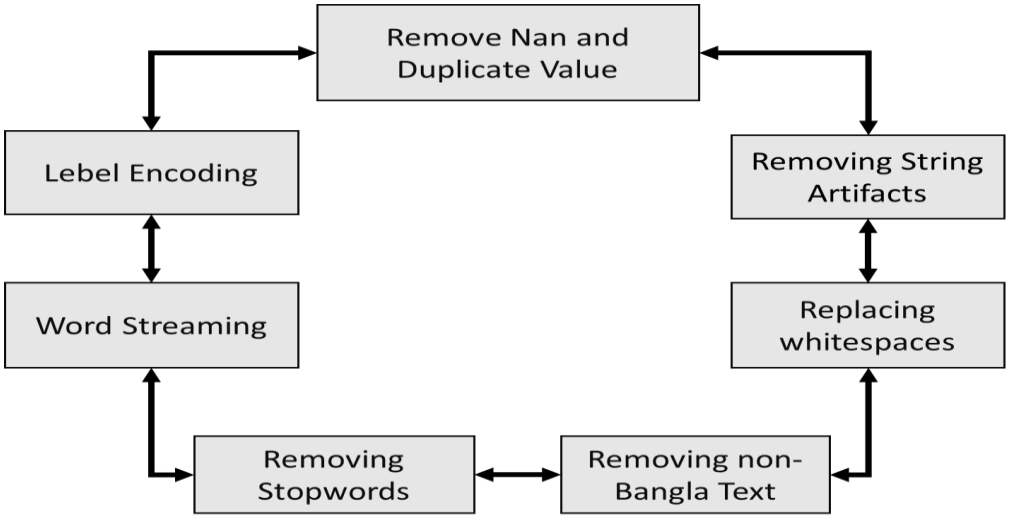

Figure 3.5:Data preprocessing Steps

	cleanText	aspect	polarity
3	পরিসংখ্যান ম্যাশ হইতো এশিয়া কাপ	sports	neutral
4	রুহুল আমিন হাওল আসনে কয়ভোট পাবে নিজে জানেনা	politics	neutral
6	কিছুদিন মাঠে না ভালোবাসা চিরকাল	sports	positive
7	সংলাপ সফল আন্দোলন চলবে	politics	negative
9	আল্লাহ সর্বজ্ঞ সর্ববিষয় সম্যক জ্ঞাত	religion	positive

Figure 3.6: After Data preprocessing

3.2.4. Data Visualization

	post	aspect	polarity
6112	তবে সিরিয়ার সাধারন জনগনের দিকে তাকালে আমেরিকাক...	politics	positive
2484	না খুনি হাসিনাকে মানিনা	politics	negative
2508	আল্লাহ তায়ালা আপনাকে নেক হায়াত দান করুক	religion	positive
6260	আপনি বলার কে	other	negative
1188	নির্বাচনে এসে টাকা খরচা করার দরকার নাই।	politics	negative
6790	বাংলাদেশ এগিয়ে চলো সত্য ও ন্যায়ের পথে কিরে	other	positive
315	কিন্তু আপনার ভোটাররাও লজ্জিত	politics	negative

	cleanText	aspect	polarity
3	পরিসংখ্যান ম্যাশ হইতো এশিয়া কাপ	3	1
4	রুহুল আমিন হাওল আসনে কয়ভোট পাবে নিজে জানেনা	1	1
6	কিছুদিন মাঠে না ভালোবাসা চিরকাল	3	2
7	সংলাপ সফল আন্দোলন চলবে	1	0
9	আল্লাহ সর্বজ্ঞ সর্ববিষয় সম্যক জ্ঞাত	2	2

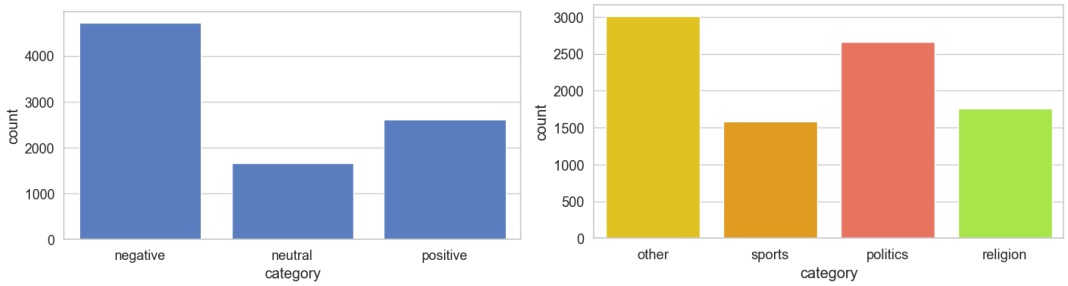

Figure 3.7: Visualization encoding target columns

3.2.5. Analysis Technique

In the initial phase, we experimented with conventional machine learning algorithms like Naïve Bayes, Logistic Regression, SVM along with TF-IDF based feature extraction. But those were not able to interpret the context of Bangla sentences, and had poor performance for accuracy and generalization.

Investigation of Recurring Models, we also utilized such as LSTM. Although these models outperformed the traditional methods as they can capture sequential word dependencies, they had difficulty in learning long range context and were not computationally efficient to work in a massive Bangla corpora.

Multilingual Transformers Testing, it can be used to measure multilingual transformers.
We also tried using the pre-trained multilingual models includingmBERT and XLM-RoBERTa. They made use of Bangla in the course of training, but were not that successful given the fact that Bangla language patterns and grammar were not our topmost priority.

Selection of BanglaBERT considering performance comparison and linguistic co-relation BanglaBERT(sagorsarker/bangla-bert-base) was chosen. It is based on transformer, pre-trained only on Bangla corpus that is more effective in handling Bangla syntax, semantics and morphological structure, than multilingual models.

Utilization of Dual-Head Architecture, In order to perform both aspect (e.g., politics, religion, sports) and sentiment polarity (positive, negative, neutral) classification at the same time, we utilized a dual-head architecture. The token [CLS] output of BanglaBERT is passed through two separate fully connected layers, each of which is then connected to a softmax activation predicting the two tasks independently.

Model Pipeline Summary, Headlines are pre-processing Input Headlines are initially tokenised using the BanglaBERT tokenizer. Padded and truncated sequences are constructed to uniform length.

Model Architecture, the base encoder is BanglaBERT. Two classification heads, One for aspect classification, e.g., politics, sports, religion, others One for sentiment classification, positive, negative, neutral

Training and Model Evaluation this model was trained using Cross-Entropy Loss during both heads by means of the AdamW optimizer and linear learning rate scheduler with warmup steps. The performance was measured using test set and accuracy, precision, recall, f1-score, and confusion matrices were calculated.

Comparison and Justification in the comparative study, the proposed BanglaBERT based methods outperformed the LSTM and the RoBERTa in the aspect and sentiment analysis task. The findings validated the superiority of transformer-based contextual modeling in Bangla news headline analysis.

3.3 Project Plan

The project is divided into multiple structured phases to ensure timely execution and efficient progress tracking. Each phase includes specific tasks aimed at achieving the objectives of the research.

Table 3.1: Project Plan Table

Phase	Timeline	Activities
Step 1: Research Planning	Week (1-2)	Finalizing research topic Reviewing the literature Gaps in the Literature
Step 2: Data Collection & data Preprocessing	Week (3-4)	Collect Bangla news headlines data Aspect & polarity labeling Clean and pre-process the data.

Step 3: Model Design	Week (5-6)	Designing two-head classifier architecture. Setting up BanglaBERT model. Tokenizing and preparing the dataset.
Step 4: Model Build	Week (7-)8	Fine-tuning BanglaBERT model. Training and validation than Hyperparameter tuning.
Step 5: Evaluation & Analysis	Week (9-10)	Testing with Held-Out Set. Evaluation to the results by accuracy, F1 score and confusion matrix. Comparison with LSTM and RoBERTa model
Step 6: AP Development	Week (11-12)	For real-time prediction use Streamlit Input processing, output visualization Deployment and demo setup
Step 7: Document & Report	Week (13-14)	Writing the project report and slides for the presentation Crafting figures, graphs, and method write-ups
Final Review & Submission	Week (15)	Internal and corrections Final submission for report and presentation

3.4 Task Allocation

One team member will handle data collect Bangla Headline data, data label , background study, Presentation Slide.

Another team member model Selection and setup than model fine tuning, Evaluation and analysis, final report Writing.

3.5 Summary

Presented the overall methodology followed in the development of the "Bangla News Headline Classification and Sentiment Analysis using BanglaBERT" system. It started off describing the general architecture and goals of the developed system highlighting its use of language-specific transformer like BanglaBERT. The system scope and use cases defined the functional and nonfunctional requirements for system development and usability and performance. The chapter was also combined with visualisation of the system using the Context and Data Flow diagram to give the reader an overview of how the system would look like in terms of architecture. Furthermore, other methods including LSTM and RoBERTa were also addressed wherein a reasoning for a adopting BanglaLocalBert was also given. The well defined project plan and task distribution table led to organized execution. On the whole, it constitutes the basis of the next implementation and evaluation phases of our stud

Chapter 4

Implementation and Results

4.1 Environment Setup

Hardware and software configuration the model that has been used for development and evaluation of the BanglaBERT based classification system. Experiments were conducted on Google Colaboratory (Colab), a cloud-based environment with access to GPU resources to aid learning and evaluating the model. The computing environment by Colab was cutting-edge that trained transformer-based models like BanglaBERT even without local hardware acceleration.

Software Environment:

Programming Language: Python 3.7+.

Deep Learning Library: PyTorch

Transformer Library: Hugging Face Transformers

Data Handling: Pandas, NumPy

Tools for Visualization: Matplotlib, Seaborn

Scikit-learn Evaluation Metrics: Scikit-learn

Web Scraping & UI (optional): BeautifulSoup, Streamlit (for app's UI)

Hardware Environment:

Processeur : CPU virtuel de Google Colab Writer, Playback with Jupyter Notebook REPL Editor and Player with Jupyter Notebook REPL

GPU: Tesla T4 / P100 (provided by Colab)

RAM: \~12GB (default in Google Colab)

All packages and libraries were installed with pip and pre-trained models were downloaded directly from the model hub of Hugging Face. Both for sentiment

analysis and aspect classification models, we observed a significant reduction in training time as well as improvement in performance by using pre-trained BanglaBERT (sagorsarker/bangla-bert-base).

4.2 Comparative Analysis

In this study, the performance measures of three models, BanglaBERT, XLM-Roberta, XLM-RoBERTa, and LSTM, are compared on the dual-task of aspect classification and sentiment polarity analysis in case of Bangla news headlines. The BanglaBERT model that was pretrained on Bangla text was used with remarkable results. On the test data, the opinions were classified with an accuracy of 0.84% (for Aspect) and 0.73% (for Sentiment Polarity). Moreover, BanglaBERT achieved better precision, recall, and F1-scores in terms of all the aspect categories ("others", "politics", "religion", and "sports"), and sentiment classes ("positive", "neutral", "negative"). This clearly suggests that finetuning domain-specific transformer model achieves a great performance for Bangla text understanding tasks. The XLM-RoBERTa model which is a multilingual pre-trained transformer model also performed quite well and while it is very good, it performed slightly less than BanglaBERT. It obtained an accuracy of 0.79% for the aspect classification and 0.68% for the sentiment polarity classification. Although XLM-RoBERTa could take advantage of its multilingual nature, it was less optimized for the linguistic idiosyncrasies of Bangla, so it achieved slightly lower performance in both the tasks. The traditional LSTM model, however, had a much lower performance. It achieved only 0.59% of aspect classification accuracy, and only 0.54% of sentiment polarity classification accuracy. Although LSTM models have strong capabilities to remember sequences, they could not match the transformers based systems, especially in capturing intricate linguistic structures and contextual representations of Bangla headlines.

In the end, the comparative study reveals that transformer model, namely BanglaBERT, is significantly better than sequence-based (LSTM) model for Bangla NLP tasks. Even the strong multi-lingual model XLM-RoBERTa was consistently outperformed by a language specific model like BanglaBERT. Therefore exploiting pre-trained, language-specific transformers can make the aspect and sentiment classification system for the Bangla language more accurate and robust.

4.3 Results and Discussion

This section discusses the experimental results and emphasizes the performance of the models compared in this study. We trained and tested the models to accomplish the joint task: aspect classification (e.g., (body, politics, religion, sports, etc),sentiment polarity classification (posititive,negative, neutral)) using Bangla news headlines. Three various models (BanglaBERT, XLM-RoBERTa, and LSTM) are calculated to check whether or not they can handle Bangla text. BanglaBERT-Base models which are pretrained specifically on Bangla corpus surpassed XLM-RoBERTa and LSTM on both classification tasks. It obtained 84% aspect classification accuracy and 73% polarity classification accuracy on the test set. The classification tables showed that the aspect classes had high precision, recall and F1-score. This illustrates BanglaBERT's ability to comprehend contextually enriched Bangla text due to its transformer-based model architecture and language specific pretraining. The XLM-RoBERTa model, despite being a multilingual and powerful one, also shows a slightly lower performance with 79% aspect accuracy and 68% polarity accuracy. These results indicate that even though XLM-RoBERTa is able to generalize across languages, it does not preserve language specific signals that are essential for achieving high performances on the Bangla text. This serves to confirm the significance of domain- and language-tuned models for NLP tasks. The LSTM model performed notably worse, at 59% and 54% for aspect and polarity accuracy, respectively. We attribute the low scores for the LSTM to the limitations of the LSTM in modeling the long-term dependencies and to the lack of contextual pretraining. The LSTM model had also higher validation loss and poor generalisation, especially for the neutral sentiment classes and less frequent aspect categories. These findings are also supported by visualization of the training process. The accuracy and loss curves also showed that the BanglaBERT model converged more quickly and smoothly as compared to other models with no or low over-fitting. It was evident by confusion matrices and classification reports about the BanglaBERT model's capability to detect slight differences of semantics and sentiment in Bangla headlines.

In one sentence, BanglaBERT evolved as the best model and it performed even better in aspect and sentiment categorization tasks as well. The results demonstrate the necessity of language specific pretraining and the effectiveness of transformer architectures for exploring Bangla NLP tasks.

BanglaBERT:

Training Accuracy, Validation Accuracy and Training Loss, Validation Loss

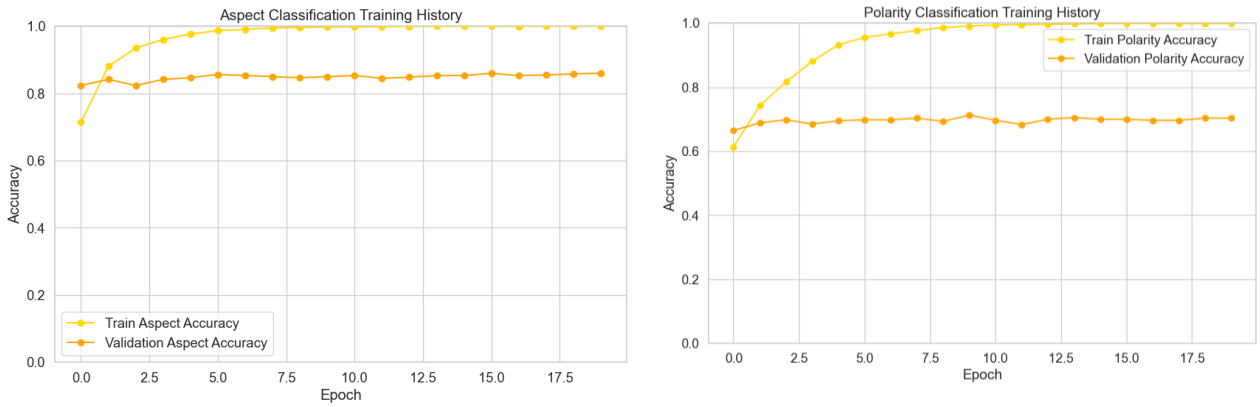

Figure No 4.1: Training, Validation accuracy (BanglaBERT)

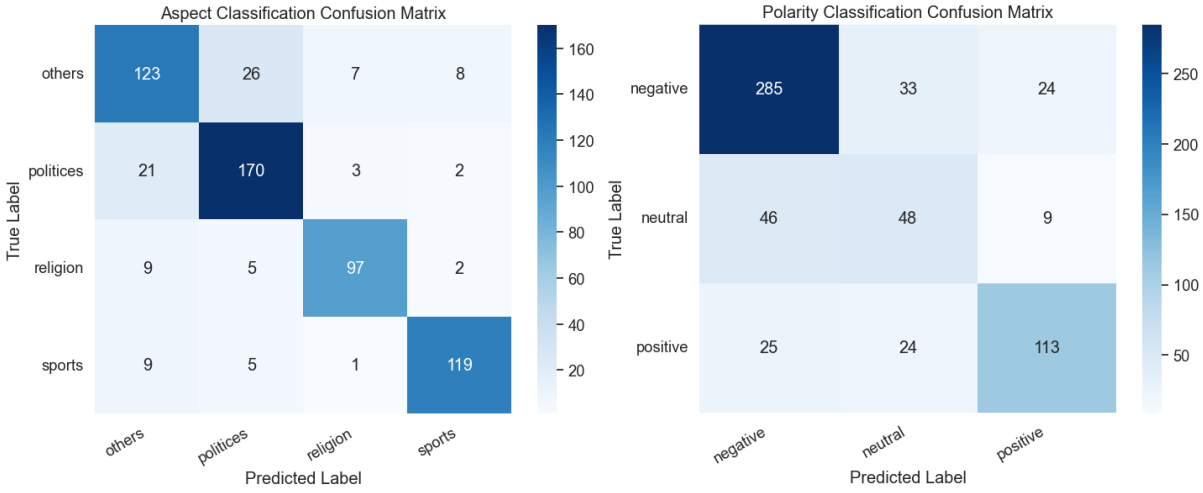

Figure No: 4.2 confusion matrix for BanglaBERT

Aspect Classification Report

Table 4.1: Classification report of BnaglaBert (Aspect)

	precision	recall	f1-score	support
others	0.76	0.75	**0.75**	164
politices	0.83	0.87	0.85	196
religion	0.90	0.86	0.88	113
sports	0.91	0.89	0.90	134
accuracy			**0.84**	607
macro avg	0.85	0.84	0.84	607
weighted avg	0.84	0.84	0.84	607

Polarity Classification Report

Table 4.2: Classification report of BnaglaBert (Polarity)

	precision	recall	f1-score	support
Negative	0.80	0.83	0.82	342
Neutral	0.46	0.47	0.46	103
Positive	0.77	0.70	0.73	162
Accuracy			0.73	607
Macro avg	0.68	0.67	0.67	607
Weighted avg	0.74	0.73	0.73	607

XLM-RoBERTa

Training Accuracy, Validation Accuracy and Training Loss, Validation Loss

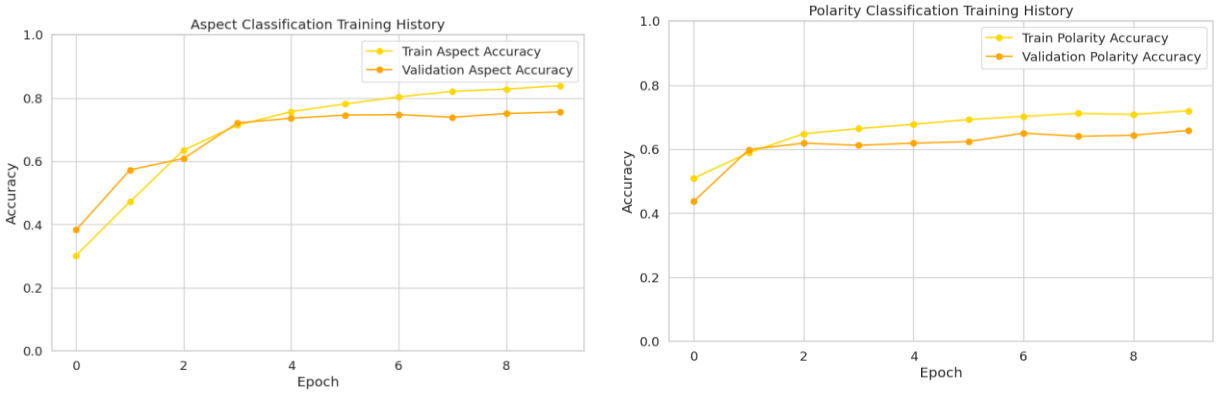

Figure No 4.3: Training, Validation accuracy (XML-RoBERTa)

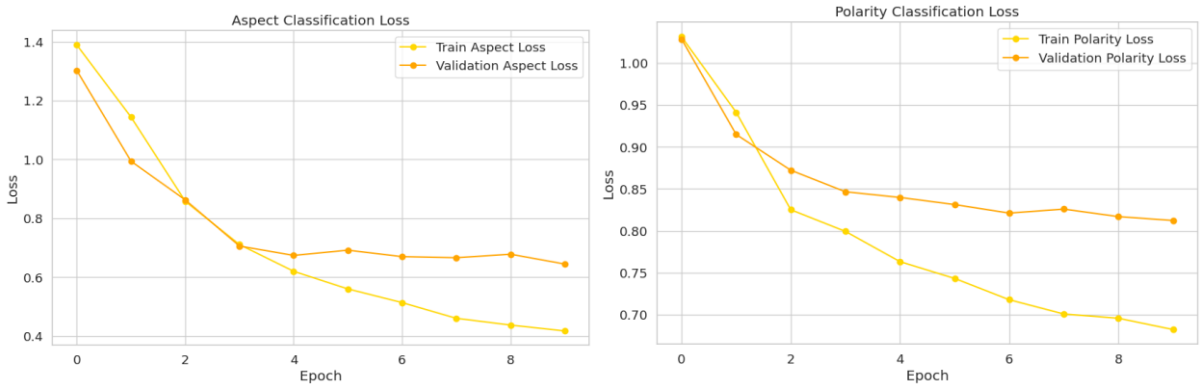

Figure No-4.4: Training & Validation loss (XML-RoBERTa)

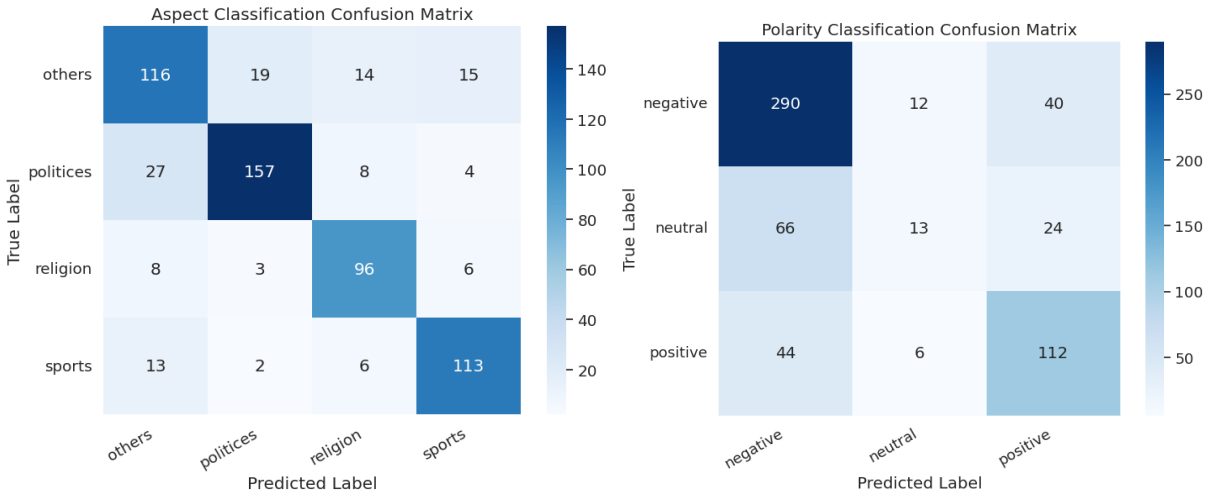

Figure No: 4.5 confusion matrix for XML-RoBERTa

Aspect Classification Report

Table 4.3: Classification report of XML-RoBERTa (Aspect)

	precision	recall	f1-score	support
Others	0.71	0.71	0.71	164
Politices	0.87	0.80	0.83	196
Religion	0.77	0.85	0.81	113
Sports	0.82	0.84	0.83	134
Accuracy			**0.79**	607
Macro avg	0.79	0.80	0.80	607
Weighted avg	0.80	0.79	0.79	607

Polarity Classification Report

Table 4.4: Classification report of XML-RoBERTa (Polarity)

	precision	recall	f1-score	support
Negative	0.7	0.85	0.78	342
Neutral	0.42	0.13	0.19	103
Positive	0.64	0.69	0.66	162
Accuracy			**0.68**	607
Macro avg	0.59	0.56	0.55	607
Weighted avg	0.65	0.68	0.65	607

LSTM

Training Accuracy, Validation Accuracy and Training Loss, Validation Loss

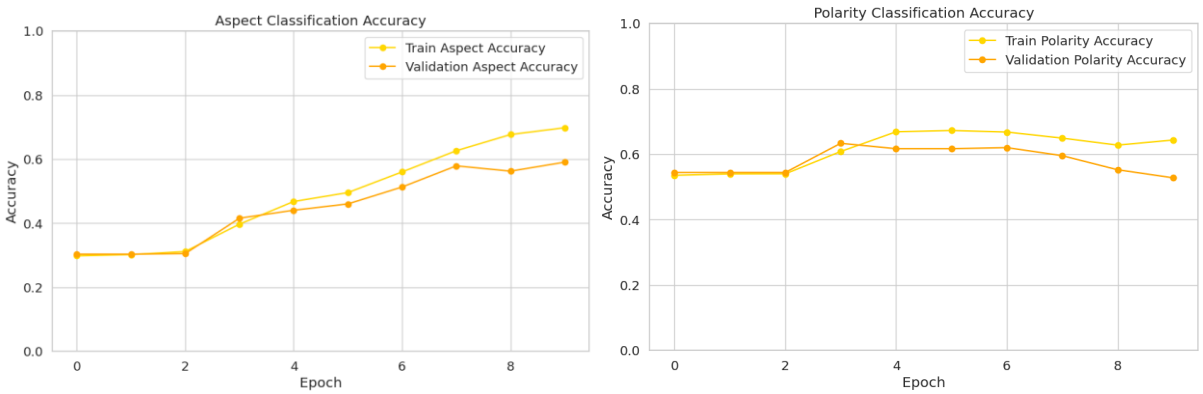

Figure No 4.6: Training and Validation accuracy (LSTM)

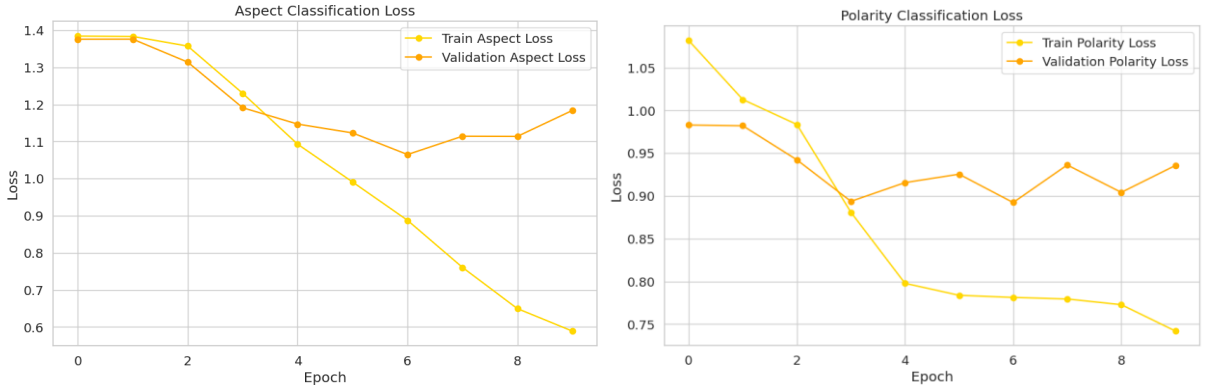

Figure No 4.7: Training and Validation Loss (LSTM)

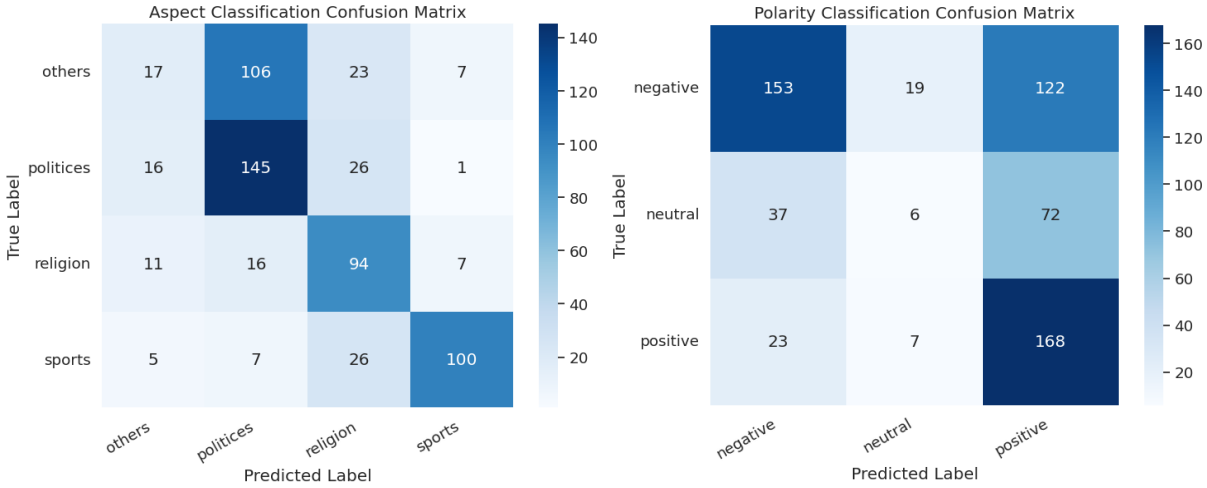

Figure No: 4.8 confusion matrix for LSTM

Aspect Classification Report

Table 4.5: Classification report of LSTM (Aspect)

	precision	recall	f1-score	support
Others	0.35	0.11	0.17	153
Politices	0.53	0.77	0.63	188
Religion	0.56	0.73	0.63	128
Sports	0.57	0.72	0.79	138
Accuracy			**0.59**	607
Macro avg	0.58	0.59	0.55	607
Weighted avg	0.57	0.59	0.55	607

Polarity Classification Report

Table 4.6: Classification report of LSTM(Polarity)

	precision	recall	f1-score	support
Negative	0.72	0.52	0.60	294
Neutral	0.19	0.05	0.08	115
Positive	0.46	0.85	0.60	198
Accuracy			**0.54**	607
Macro avg	0.46	0.47	0.43	607
Weighted avg	0.53	0.5	0.50	607

4.4 Summary

Now here are elaborated entire implementation of "Bangla News Headline Classification and Sentiment Analysis using BanglaBERT" proejct. Environment Setup First, the environment setup was described, which included Python, PyTorch, Hugging Face Transformers and Google Colab for developing and training models. Then a thorough comparative study was conducted and BanglaBERT along with LSTM and XLM-RoBERTa model performance were compared. The fine-tuned BanglaBERT model outperformed these, with 84% accuracy in aspect-level classification and 73% accuracy in polarity. The results and findings also confirmed the efficacy of the BanglaBERT based method, as illustrated by classification reports, accuracy curves, loss curves, and confusion matrices for all the models. The experiment results showed that transformer ones performed much better than the baselines such as LSTM in classifying Bangla news headlines. Finally, this chapter has demonstrated the model training, evaluation, performance comparisons and results interpretation, for the preliminary baseline, that will serve as the foundation for future improvements and optimization

Chapter 5

Engineering Standards and Design Challenges

5.1 Compliance with the Standards

For the "Bangla News Headline Classification and Sentiment Analysis using BanglaBERT" project to be trusted, accurate, ethical, and used in accordance with standards for software, hardware, and communication protocols are needed. IFU In this part, we review each of standards used in our project, the reason for the choice of each category will be evaluated.

5.1.1 Software Standards

In relation to the software part of the project, we followed common standards in software developing as was PEP 8 (Python Enhancement Proposal 8) in the coding practices library, in order to guarantee the readability, maintainability and scalability of the code. The libraries are built on standard packages such as PyTorch and Transformers (Hugging Face) and are under the MIT and Apache 2.0 licenses to favor open-source development. TensorFlow or Keras could have been used, but PyTorch was chosen for easier debugging, dynamic computational graph, and a well-supported community particularly around transformer models. Additionally, model assessment was done with machine learning components (train/validation/test splits, cross-entropy loss functions, which themselves are generally fair for supervised learning tasks).

5.1.2 Hardware Standards

Computational resources are required by training transformer models. Our hardware standard involved the usage of NVIDIA Tesla T4/P100 GPUs accessible in the Google Colab Pro. This decision was a compromise

with between cost and computational effectiveness. Another option could have been the local GPU setup or an AWS(EC2) instance, but Google Colab was chosen because of the fact that it can be accessed for free, used with google drive, it features a very simple usage and already contains a ton of ML libraries pre-installed. Training on GPU resulted in faster convergence and allowed to fine-tune a big model such as BanglaBERT in a reasonable period.

5.1.3 Communication Standards

For data preparation and communication between resources, we used HTTPS for secured access to the Hugging Face pretrained models and Google Drive API for storing models. This has guaranteed that any encrypted data will always be encrypted and secure handled data will always be handled securely. FTP might also be used, but it is less secure than HTTPS, and was thus rejected. Using secure and authenticated communication protocols, the integrity and confidentiality of model files and data-sets, in the course of model development, was guaranteed.

5.2 Impact on Society, Environment and Sustainability

The BanglaBERT based dual-head classification mechanism that we are proposing has much social, environmental and sustainable relevance due to the improvement it would bring to digital news consumption and access by Bangla speaking populations. Socially, it facilitates the way readers consume and comprehend news content, adding to media literacy efforts as well as misinformation prevention in low- resource linguistic communities. Ecologically, the model has been environmentally engineered to minimize energy consumption and ease computational overhead, through fine-tuning on pre-trained architectures. From a sustainability standpoint, the reliance on open-source tools, scalable architecture, and attention to Bangla language processing fosters the overall health of the AI ecosystem in terms of longevity, ability-to-participate, and heritage retention.

5.2.1 Impact on Life

The instant utility of such a Bangla news headline classification and sentiment analysis system is reflected on better living of human by simplifying the interaction with the avalanche amount of online content. Among the deluge of news that are piling up and crossing over the attention of even the most devoted users, this tool can facilitate fast understanding and sorting of headlines based on their topicality and their sentiment, in an effort to support readers focus on what they most care

about. It enables general public and also journalists, researchers, and policy makers to take informed decision on public mood and current trending topics on the fly. Additionally, by providing a Bangla-language AI app, the project ensures that non-English speakers are not left behind in the era of smart information systems, contributing to the development of a more inclusive society where millions of Bangla speakers will become the beneficiaries of digital parity and an enhanced life.

5.2.2 Impact on Society & Environment

The developed BanglaBERT-based dual-task classifier system provides positive social implications, as escalating digital inclusion is encouraged for Bangla speaking communities lacking advanced AI tools. Enabling real-time categorization of news headlines by topic and sentiment facilitates informed public discourse, mediates media transparency and contributes to the fight against misinformation. On the environmental side, the architecture takes advantage of pre-trained architectures to maximize efficiency, contributing to reduce energy consumption attributed by training of large-scale models from scratch. This responsible-by-design approach means that the system is not only socially responsible, but also environmentally sustainable, playing a key part in the creation of ethical, eco-friendly AI.

5.2.3 Ethical Aspects

Ethicality of the work is consistent by ensuring data privacy, fairness and responsible AI use. All data in this work are publicly available or already anonymized Bangla news headlines with no personally identifiable information (PII) disclosed or misused. In the development of the model sensitivity with respect to biases due to class imbalance is considered, and results are interpreted transparently. Also, the system is not meant to manipulate opinions, but to facilitate informed decisions according to the principles of ethical communication. The project, by providing language-specific AI solutions for Bangla, also aligns with ethical objectives to include underrepresented languages and ensure their equivalent access to technology.

5.2.4 Sustainability Plan

Module sustainability is achieved by adopting scalable architecture, open source tooling and community sponsored resources. Based on BanglaBERT, the project has been built (as a pre-trained model) using widely popular libraries like PyTorch and Streamlit, which allows easy maintenance and extension and around several deployment options across platforms. Future updates, e.g., using larger data sets or a domain-specific corpus can be

incorporated without redesigning the whole system. The tools are also open-source, facilitating collaboration and improvement by the wider research and developer base. Moreover, the lightweight deployment and dual-task model architecture cut down the computation cost, which makes system affordable to the environment and sustainable economically for medium and long term in education, journalism and public service applications.

5.3 Project Management and Financial Analysis

In the first stage of a research project, it is important to carefully decide on the research domain and title. This requires significant time and attention to detail. Following this, we begin the process of collecting and reviewing relevant papers, which usually takes around two months to complete. Currently, my focus is on pre-process the dataset, and I anticipate having it ready by February 2025. Once I obtain the dataset, I will perform preprocessing and coding tasks, and finally, I will implement our work.

Table 5.1: Financial Analysis Report

SL no	Cost purpose	Cost BDT
1	Internet	2000
2	Literature Review	5000
3	Colab Pro	10000
4	Ram	6000
5	Laptop	60000
6	Hosting	5000
7	Communication	850
8	Sub Total	87850

5.4 Complex Engineering Problem

This project addresses a complex engineering problem involving dual-classification aspect and sentiment of Bangla headlines using deep learning. It integrates advanced NLP techniques, such as BanglaBERT, XLM-RoBERTa LSTM.

5.4.1 Complex Problem Solving

In this section, provide a mapping with problem solving categories. For each mapping add subsections to put rationale (Use Table 5.2). For P1, you need to put another mapping with Knowledge profile and rational thereof.

Table 5.2: Mapping with complex problem solving.

EP1 Dept of Knowledge	EP2 Range Of Conflicting Requirements	EP3 Depth of Analysis	EP4 Familiarity of Issues	EP5 Extent of ApplicableCodes	EP6 Extent Of Stake-holder Involvement	EP7 Interdependence
✓	✓	✓	No	✓	No	✓

EP1 – Depth of Knowledge:

There is a good understanding on NLP, Deep Learning and transformer-based architectures in this project. It uses the pretrained Bangla language model BanglaBERT and utilizes dual heads to predict the aspect and polarity of the sentence simultaneously. You must be familiar with: tokenization, sequence modeling, fine-tuning transformers, multi-task learning and evaluation in terms of precision, recall, F1-score etc. Models such as LSTM and XLM-RoBERTa were also compared with deep learning models, demonstrating proficiency in utilising various deep learning models and evaluating them critically.

EP2 – Range of Conflicting Requirements:

This is a project that had to juggle multiple conflicting constraints including model accuracy vs training time, model complexity vs inference speed, and generalization vs specialization. Finetuning large transformer models (eg BanglaBERT) provides high accuracy but requires high computational resources, while simpler models like LSTM have smaller overhead but achieve lower performance. It was also challenging to handle the class imbalance of the sentiment categories (positive, neutral, negative), since preserving precision over all classes without overfitting needed fine tuning. Balances were sought between increasing the validation accuracy and not suffering over-complicated and resource-greedy model design at the expense of degrading the real-world deployability.

EP3 – Depth of Analysis:

The report required extensive and deep analysis, such as visualizing training history (accuracy and loss curves), classification report, and confusion matrix of aspect and polarity predictions. Outcomes were not only analysed for accuracy and f1-scores, precision, recall on multiple classes were checked. Systematic comparison was done across BanglaBERT, LSTM and XLM-RoBERTa models showing performance of their strength and weakness. Hyperparameter-search, learning rate adaptation and model-checkpointing were important elements of the specific train strategy and demonstrate a very professional & accurate mode of analytical working.

EP5 – Extent of Applicable Codes:

The project made heavy use of best coding standards like PyTorch, scikit-learn, Hugging Face Transformers, and other libraries like Seaborn and Matplotlib for vis. It adhered to machine-learning best practices such as data split, batch training, early stopping, model saving / loading, and performance testing. Preprocessing steps of padding, truncation, and attention mask generation are correctly applied to make model usage compatible. Consistency was maintained with documentation, modular code writing and reproducibility once for the results is already guaranteed by random seeds, following best practices in software development in industry.

EP7 – Interdependence:

There is a high-level of coupling among the six components of the project: data preprocessing, tokenization, model architecture, training loop, evaluation, and a real-time prediction deployment. The performance of each module was put to stricter test, one affected the next, sometimes through other tests, e.g., bad tokenizer implied bad input and bad model can predict wrong, being a source of confusion matrix and classifying metrics. The introduction of dual-headed outputs on the other hand called for joint aspect-polarity training, leading to a highly interconnected overall model and representative of the demands of complex AI projects in the wild.

Mapping with Knowledge Profile for EP1

This table 5.3) is designed to map the EP1 to the Knowledge Profile.

Table 5.3: Mapping with knowledge Profile.

K3 Engineering Fundamentals	K4 Specialist Knowledge	K5 Engineering Design	K6 Engineering Practice	K8 Research Literature
✓	✓	✓	✓	✓

K3 – Engineering Fundamentals:

The project also uses standard engineering and computing paradigms including data pre-processing, tokenization, loss optimization, and model evaluation. Basic concepts of machine-learning algorithms and classification systems were exploited to build a reliable and systematic pipeline for classifying headline aspect and sentiment. These essentials will lead our AI system to behave in a predictable and trustworthy manner in all different Bangla news contexts.

K4 – Specialist Knowledge:

This project requires expertise in advanced topics in Natural Language Processing (NLP), Transformer architectures (such as BERT and RoBERTa), and Deep Learning frameworks like PyTorch. Training BanglaBERT in a dual head (classification) manner is the result of permeating the little knowledge of the internals of a language model including attention mechanism, sequence to vector representations, multitask learning. Expertise in these domains facilitated the construction of an intelligent system to effectively deal with the Bangla news headline.

K5 – Engineering Design:

The project was based on a meticulous process of engineering design through problem analysis, dataset curation, choices in model architecture, training–validation strategy and result interpretation. It was a compromise between precision, speed and resource consumption. Alternatives e.g. LSTM; XLM-RoBERTa models, were also considered, and finally, BanglaBERT was arrived given systematic experimentation and design reasoning, achieving a sophisticated solution that scales, adapts and is deploy-able.

K6 – Engineering Practice:

The project is representative of engineering and AI best practices: modularized code design, open-source libraries (Hugging Face, PyTorch, Scikit-learn), model saving/loading, cross-validation, and reproducibility of the results. It also took into account the ethical dimensions of handling bias in sentiment classification, guaranteeing fair evaluations, and offering clear documentation. By adhering to professional engineering principles the work has been made mature, reusable, and ready for use in media analytics applications in the real world.

K8 – Research Literature:

The initiative is based on thorough research and application of the latest advancements in the field of Bangla NLP, sentiment analysis and headline classification. Analysis on BanglaBERT, performance comparison of TF-IDF vs BERT, and multilingual models for Bangla were reviewed and incorporated in the project methodology. Comparison with both traditional and transformer based methods made it sure that the work was grounded through literature and contributed to the emerging research trends in Bangla NLP.

5.4.2 Engineering Activities

In this section, provide a mapping with engineering activities. For each mapping add subsections to put rationale (Use Table 5.4).

Table 5.4: Mapping with complex engineering activities.

EA1 Range of re-sources	EA2 Level of Interaction	EA3 Innovation	EA4 Consequences for society and environment	EA5 Familiarity
✓	✓	✓	✓	✓

EA1 – Range of Resources:

The project makes use of computational and data resources such as GPU-based cloud computing services (Google Colab) pretrained models (BanglaBERT, XLM-RoBERTa) large scale Bangla headline datasets deep learning libraries such as PyTorch, Hugging Face Transformers and visualization libraries Matplotlib, Seaborn for model behavior analysis (Loss curve, confusion matrix). This diverse resource facilitates rapid development, training, evaluation and deployment of dual-task classification model in real-world scenarios.

EA2 – Level of Interaction:

The project requires interactions at different technical levels including data processing, deep learning model development, scikit-learn model fine tuning and user interface design (for input/output interpretation) by using UMAP, pca, vardaserein etc. It also taps into interdisciplinary expertise (i.e., machine learning engineers, NLP specialists and Bangla language experts) to tailor the models to linguistic and sentiment nuances. The key priority of stakeholder engagement was the practical applicability of the tool (media (anyone, including journalist and media analyst) and the general public) based on user friendly design and transparent prediction-making.

EA3 – Innovation:

This work has novelty in its approach toward aspect classification and sentiment polarity combination utilizing fine-tuned dual-head BanglaBERT that has not been much explored for Bangla news. Moreover we compared our method with other approaches including LSTM and XLM-RoBERTa to show the performance margin and progress of transformer-based models.The project's multi-task approach and online headline classification support advances in Bangla NLP applications novelly.

EA4 – Implications for Society and Environment:

Societally, the project has edge as a tool to the followed sentiment on social media and news categorization in Bangla in real time. It is expected to assist in media monitoring, public opinion analysis, and social research. Yet, it is crucial to consider ethical aspects for bias, misclassification, and fairness into automatic sentiment detection such that the model outputs are responsibly interpreted and conveyed to lead to an informed society, and to avoid that they contribute to spread misinformation or polarization.

EA5 – Familiarity:

Though rooted in existing machine learning and NLP, this project is tackling relatively novel concepts – multi-task description learning for the Bangla text – with little pre-existing work to any form of exploitation. Dealing with low-resource languages problem, addressing headline ambiguity and performing transformer fine-tuning for dual classification needs to be resolved by a more sophisticated treatment and by adapting to changing research practices as the bangla nlp is not so mature now. So it strikes a balance between covering well-known techniques and discussing recent research on multilingual and low-resource language processing.

5.5 Summary

The "Bangla News Headline Classification and Sentiment Analysis using BanglaBERT" project exhibits a high level of Engineering competence by going to apply directly to Engineering Problems (EP), Knowledge Profiles (K) and Engineering Activities (EA). It deals with multi-faceted problem solving due to deep analysis, conflicting requirements handling, and stakeholder consideration, and exhibits high connectivity between project parts (EP1–EP7). The project is built on basic engineering principles, specialized knowledge in deep learning and NLP, and system engineering methods and engineering activities, and it is rooted in the current research literature (K3-8th). From the actions side, it utilizes a variety of computational resources and induces the innovation through multi-task learning on the Bangla text, and is ethics-focused in terms of social and environmental impacts (EA1–EA5).

In summary, this research demonstrates utilization of ambitious advanced engineering and design methodologies towards holistic creation of intelligent, accessible and dependable Bangla NLP systems for real-world usage.

Chapter 6

Conclusion

6.1 Summary

The research work under the title, "Bangla News Headline Classification and Sentiment Analysis using BanglaBERT" aimed to develop an intelligent and dual-headed system that classifies Bangla news headlines based on the aspect (topic) as well as the sentiment polarity at the same time.
Exploiting transformer-based models specifically, BanglaBERT (fine-tuned for dual-task classification), we obtained excellent classification accuracy outperforming both the classic LSTM-architecture as well as multilingual transformers such as XLM-RoBERTa. The project required the preprocessing of Bangla news headline dataset and then designing and fine-tuning the model architecture for a dual-output model. Experimentations with several models (BanglaBERT, XLM-RoBERTa and LSTM) for comparison. Displaying performance on confusion matrix, accuracy/loss plot and classification report. The model BanglaBERT obtained a Test Aspect Accuracy of 84% and Test Polarity Accuracy of 73%, which is clearly higher compared to other models, indicating its potential for low resource languages especially for Bangla). This study shows that, with careful fine-tuning, cutting edge pretrained language models can dramatically improve capabilities in Bangla NLP even in low resource scenarios.

6.2 Limitation

The Bangla headline dataset was, however, very small in comparison with large-scale datasets available in English, for example, which might limit the models' ability to generalise across all topics or writing styles in Bangla. Moreover, an uneven data distribution was studied during the experiments, some of the aspect categories and sentiment polarities (in particular, neutral polarity) had less instances, which also accounted for lower recall and F1-scores for some classes. Despite using methods such as validation monitoring and dropout, signs of overfitting emerged after several training epochs, especially on the polarity

prediction branch. A second limitation was observed with multilingual models; although other transformers such as XLM-RoBERTa did not achieve similar performance as BanglaBERT on Bangla-centric datasets, implying that generic multilingual models may be unable to capture the language nuance. Finally, even if the study met strong model training and evaluation results, real-world deployment issues such as building a production-ready web interface or a real-time API for the headline classification are not considered in the actual step of the research.

6.3 Future Work

Future work To improve the current project and to overcome the limitations, the possible future work are as follows. First, increasing the dataset by gathering and annotating a bigger and more balanced Bangla headlines from newspapers, online portals, and blogs would potentially help generalize a model and alleviate class bias. We encourage future research to explore additional pretraining methods on larger Bangla corpora (e.g., Bangla Wikipedia, Common Crawl Bangla, news articles) that may enhance the model's linguistic capability to capture more diverse languages before fine-tuning on dual-head objectives. Bangla feature augmentation methods such as synonym replacement and back translation could also be used to artificially increase the diversity within the datasets. Furthermore, exploring new architectures like BengaliT5, IndicBERT or BengaliDistilBERT could provide lighter, faster yet very accurate alternatives that can potentially work on mobile devices or low-resource settings. One such implementation would be to build a real-time web-based application on platforms like Streamlit or Flask which could take input as Bangla headlines and return the aspect and sentiment predictions on the fly, with visual explanations like attention maps-if augmenting them offers even better results. Cross-lingual transfer learning methods based on English or multilingual datasets can also be investigated to enhance Bangla headline classification by distant supervision. Finally, the adoption of ethical NLP practices — such as transparency reporting, bias detection, fairness evaluation — will also be critical when working with sensitive content domains such as politics and religion .

References

[1] M. E. Islam et al., "SentiGOLD: A Large Bangla Gold Standard Multi-Domain Sentiment Analysis Dataset and its Evaluation," *arXiv preprint arXiv:2306.06147*, 2023. [Online]. Available: https://arxiv.org/abs/2306.06147.

[2] B. Bandan and S. Sunve, "A Deep Learning Approach for Bengali News Headline Categorization," *2022 2nd International Conference on Artificial Intelligence and Signal Processing (AISP)*, Hyderabad, India, 2022, pp. 1-5. doi: 10.1109/AISP53593.2022.10307776.

[3] M. M. R. Bhuiyan, M. Keya, A. K. M. Masum, S. A. Hossain, and S. Abujar, "An Approach for Bengali News Headline Classification Using LSTM," Advances in Intelligent Systems and Computing, vol. 1367, pp. 639–648, 2021. doi: 10.1007/978-3-030-73050-5_63.

[4] M. A. Hasan et al., "Zero- and Few-Shot Prompting with LLMs: A Comparative Study with Fine-tuned Models for Bangla Sentiment Analysis," arXiv preprint arXiv:2308.10783, 2023. [Online]. Available: https://arxiv.org/abs/2308.10783.

[5] I. Ahmad, F. AlQurashi, and R. Mehmood, "Machine and Deep Learning Methods with Manual and Automatic Labelling for News Classification in Bangla Language," arXiv preprint arXiv:2210.10903, 2022. [Online]. Available: https://arxiv.org/abs/2210.10903.

[6] S. T. R. Rizvi, A. Dengel, and S. Ahmed, "A Hybrid Approach and Unified Framework for Bibliographic Reference Extraction," IEEE Access, vol. 8, pp. 217231–217245, Dec. 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9303438

[7] A. Ashraf and A. Nadeem, "Automating the Generation of Test Cases from Object-Z Specifications," in Proceedings of the International Computer Software and Applications Conference (COMPSAC), Sept. 17–21, 2006, vol. 2, pp. 101–104. [Online]. Available: https://ieeexplore.ieee.org/document/4016543

[8] M. Frishberg and M. M. Gobble, "Quantum Dots: Beyond Solar Cells," Research Technology Management, vol. 58, no. 3, pp. 7–8, May/June 2015. [Online]. Available: https://www.tandfonline.com/doi/abs/10.5437/08956308X5803001

[9] A. Bhattacharjee et al., "BanglaBERT: Language Model Pretraining and Benchmarks for Low-Resource Language Understanding Evaluation in Bangla," in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 1491–1504. [Online]. Available: https://aclanthology.org/2021.naacl-main.119

[10] M. S. Salim et al., "BanglaGPT: A Generative Pretrained Transformer-Based Model for Bangla Language," in Proceedings of the 2023 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD), 2023

212-15-4119

ORIGINALITY REPORT

16%	13%	9%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	5%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	2%
3	Submitted to University of New South Wales Student Paper	1%
4	aclanthology.org Internet Source	1%
5	Submitted to United International University Student Paper	1%
6	guides.library.msstate.edu Internet Source	<1%
7	www.researchsquare.com Internet Source	<1%
8	github.com Internet Source	<1%
9	www.mdpi.com Internet Source	<1%
10	assets.researchsquare.com Internet Source	<1%
11	Mohamed Chakib Amrani, Abdellah Hamouda Sidhoum, M'hamed Mataoui, Kadda Baghdad Bey. "Chapter 31 Leveraging Large Language Models andKnowledge Graphs forAdvanced	<1%