A Sentiment-Based Comprehensive Rating Model with Extensive Dataset for Nationwide Hospital Rating in Bangladesh Using Natural language Processing and Machine Learning

by

Asiful Kanzan Auishik
19101628
Fariha Mohammed Al-Zahir
22101874
Magferah Sultana Samia
22101875
Moumita Hossain
22301083

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
BRAC University
October 2025.

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Asiful Kanzan Auishik	Fariha Mohammed Al-Zahir
19101628	22101874
Magferah Sultana Samia	Moumita Hossain
22101875	22301083

Approval

The thesis/project titled "A Sentiment-Based Comprehensive Rating Model with Extensive Dataset for Nationwide Hospital Rating in Bangladesh Using Natural language Processing and Machine Learning." submitted by

1. Asiful Kanzan Auishik (19101628)
2. Fariha Mohammed Al-Zahir (22101874)
3. Magferah Sultana Samia (22101875)
4. Moumita Hossain (22301083)

of Summer, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science in Fall 2025 Year.

Examining Committee:

Supervisor:
(Member)

Amitabha Chakrabarty, PhD

Professor
Department of Computer Science and Engineering
Brac University

Thesis Coordinator:
(Member)

Md. Golam Rabiul Alam, PhD

Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD

Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

Bangladesh has a huge number of public and private hospitals in its districts. However, there is no trustworthy, unbiased resource for patients to choose the best hospitals concerning quality of service. The existing star-based rating systems are prone to manipulation and do not capture detailed feedback from the patients. This paper proposes an advanced model of hospital rating that grades the hospitals based on online reviews of the patients, considering aspects like patient experience and quality of care. This proposed model employs NLP and ML techniques to analyze the sentiment of patient feedback and extract insights from it. It is expected to provide a data-driven hospital rating system based solely on user experiences by integrating various dimensions of hospital service quality to identify strengths and areas for improvement across the country. A large dataset of structured and unstructured reviews is collected from online platforms. Text mining and advanced NLP techniques process sentiment data. Various machine learning models, such as SVM, BERT, and CNN, are trained and validated on pre-processed data for sentiment prediction. Steps to achieving this objective involve data collection, data preprocessing, sentiment analysis, and, eventually, aspect-based sentiment analysis using zero-shot aspect detection to generate hospital ratings based on 4 aspects: treatment quality, cleanliness, affordability, and service quality. Rating generation classify sentiments as positive, negative, or neutral, thereby dynamically, and in real time, rating a hospital. This model provides a more reliable and nuanced rating system that allows for transparent comparisons across hospitals and actionable insights into strengths and weaknesses. The framework is adaptable to other sectors, such as education and retail, providing an enlarged scope of application for sentiment analysis in service quality evaluation. For sentiment prediction of the reviews, BERT proves to be the best performing model out of all the models in terms of accuracy, precision, recall, and F1 score, producing 94.1%, 94.7%, 94.1%, and 94.4% respectively. For the aspect-based ranking of hospitals, the model is most confident in detecting treatment quality as it produces the highest teacher threshold of 0.51 for this aspect. It also produces the highest precision, F1 score, and agreement of 0.98, 0.97, and 0.96, respectively, for treatment quality, whereas, both affordability and service quality score the highest recall of 0.99.

Keywords:
Sentiment Analysis (SA), Advanced Hospital Rating Systems, Natural Language Processing (NLP), Machine Learning (ML), Patient Feedback, Google Reviews, Data-driven Rating, BERT (Bidirectional Encoder Representations from Transformers), Deep Learning (DL), Deep Neural Network(DNN), Convolutional Neural Network(CNN), Support Vector Machine (SVM), Text Mining Algorithms, Scalable Framework

Acknowledgement

We want to begin by thanking The Almighty for giving us the strength and patience to finish this research. We are especially grateful to our thesis supervisor, Dr. Amitabha Chakrabarty, and to Research Assistants Mr. Fahim-Ul-Islam and Mr. Azwad Aziz. And a special "Thank you" to Mr. Partha Bhoumik for helping us every step of the way. Their guidance and support were crucial during this journey. We also appreciate our teammates and teachers for their help along the way. Finally, we sincerely thank our families and friends for their understanding and encouragement, which motivated us to complete this work.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

In recent years, online reviews have become a valuable source of feedback for assessing the quality of products and services across various industries. In many countries, this review system has flourished in the healthcare sector, allowing patients to make decisions based on quality metrics and ratings from available reviews. There are structured and detailed platforms for collecting and analyzing these reviews. For example, in Europe, a dedicated platform called "Statista R" provides rankings of hospitals based on several metrics[25]. However, such resources remain underutilized in Bangladesh. While patients in Bangladesh occasionally share their experiences in the form of comments and pictures on online platforms, these reviews are often unorganized and scattered, making it challenging to determine which hospital is the best choice. Most public and private hospitals, clinics in this country lack a structured online presence to systematically collect and manage patient feedback. Currently, Google's star rating system is used for online feedback, but this is not very effective in accurately measuring hospital performance. As a result, people of Bangladesh struggle to make informed decisions about which hospital to visit, often wasting both time and money in the process.

1.2 Motivation

One significant concern in Bangladesh's healthcare is the increasing tendency of patients to seek medical treatment abroad. Recent reports indicate that Bangladeshis spend over $4 billion annually on outbound healthcare tourism[24]. In 2023 alone, India witnessed a 48% rise in medical tourists from Bangladesh seeking treatment there[23]. This shows how people do not have much confidence and trust in local medical treatments. This happened because patients often struggle to choose which hospitals excel in specific services, cleanliness, time management, surgical options, safety, and quality of doctors. While platforms like Google Maps offer traditional star ratings, they often fail to capture the emotional depths in reviews, which results inaccurate assessment of hospital performance. Additionally, some reviews are manipulated or serve as advertisements for specific hospitals.
For developing a comprehensive rating model, we need to create a massive dataset with categorized comments and reviews that can represent the patient sentiments regarding hospitals nationwide. This study focuses on building an extensive and

well-organised dataset, combining data from online sources, as most of the existing data is unorganized and incomplete.

1.3 Problem Statement

While Bangladesh boasts of long list of both public and private hospitals, finding a way to compare them has become a problem for patients and their families. Most of the current star-based rating systems are biased, easy to manipulate, and shallow, failing to reflect critical aspects like treatment quality, expertise of the doctors, and management of the hospital. In the absence of such a system, patients usually rely on word-of-mouth or unverified online reviews, leading to uninformed choices and a complete lack of accountability in the sector.
There is a serious need for more transparency in a hospital rating system that is genuinely data-driven, reflecting real-life patient experiences. The study will focus on developing a high-level model using NLP and ML to analyze patients' feedback through online sources. Multipronged analysis will provide balanced, insightful ratings of hospitals and help patients make informed decisions while encouraging improvement in services for the hospitals.

1.4 Objective

The aim of this research work is to create a more transparent, data-driven, and patient-focused healthcare system in Bangladesh. It lays the groundwork for a hospital recommendation framework that benefits patients, hospitals, and policymakers while marking a new era of accountable and accessible healthcare quality measurement with the help of state-of-the-art AI and ML methods. It will use sentiment and aspect analysis based on real patient feedback. Traditional star ratings are limited and can give a misleading view of healthcare quality. At the core of this initiative is a new dual-ranking technique. Hospitals are graded on specific parameters like Cleanliness, Affordability, Quality of Treatment, and Quality of Service. A composite score that consolidates these parameters into a single, simple, conclusive, and regularly updated figure is also present. These grades change as soon as new reviews are added, reflecting close to real-time feedback about the performance of hospitals. The primary objectives are to allow patients to make more informed choices about their health care. A patient will be able to find such hospitals with ease that answer their questions about specialized care and general healthcare through an integrated, overall statistical data and responses in a model that is easily accessible. One of the key achievements of this thesis is the first nationwide collection, preprocessing, and organization of hospital review data of Bangladesh. This research converts scattered, unstructured, and multilingual online reviews into a clean, tokenized, and analyzable dataset. The proposed data structure allows for easy sentiment analysis and scoring based on metrics that ensure fair rankings reflective of patient experiences.
The other important aspect of this study is continuous improvement in hospital services. Unlike the static rating systems of old, the model will let rankings move with real-time performance and patient feedback. It is a dynamic system that allows hospitals to continually reassess their services in the interest of competing favorably. In this approach, the hospitals focus on patient satisfaction and quality of care and,

therefore, are motivated to bridge gaps in the services offered to these patients, thereby promoting high standards of health within the country.
The goal is to help patients with clear, detailed insights into hospitals that suit their needs, whether for general care or specialized conditions like cardiac issues, cancer, or neurological treatment. The system also encourages hospitals to continuously improve. By connecting rankings directly to performance and patient feedback, the model motivates healthcare institutions to raise service standards, improve resource use, and cut down on waste.

1.5 Methodology in Brief

This research is based on a mixed-method approach, combining both qualitative and quantitative techniques to tap into advanced data analysis models for developing a robust rating system for hospitals. Core to this methodology will be the use of Artificial Intelligence and Machine Learning algorithms in analyzing both structured and unstructured patient feedback emanating from various online sources.
These contain the opinions garnered on the review sites of hospitals, social media, and medical forums. They would be coming from various hospitals across the entire Bangladesh, which made the dataset comprehensive enough to be representative of the country's health situation. This will also allow us to document several experiences and patient opinions, which will give us better information regarding the hospital's performance.
Techniques for data analysis and processing shall be accomplished with methods in Natural Language Processing. Free-text reviews, which are unstructured data, will be cleaned and structured with the guidance of text mining programs in order to extract the prevailing sentiment indicators. The output of the sentiment analysis will provide a clear, nuanced perspective for patient satisfaction: positive, negative, and neutral comments.
It also employs different Machine Learning algorithms, i.e., the supervised learning models used for the forecasting and fine-tuning of hospital scores according to the outcomes of Sentiment Analysis: SVM and Random Forests. These models would be trained According to both historical data and real-time remarks, the more data is processed, the more precise such models become. Furthermore, the latest sentiment prediction is attainable with the assistance of Deep Learning models like RNNs, which will address the inability in the expression of sentiments and opinions by the patients via reviews.
Data will be sampled by following a stratified random sampling approach to make Representative data of any hospitals within regional and specialist areas. This ensures that the skewed basis in this rating is not tilted in favor of single institutions or, worse, some locations. For maintaining a sense of data balancing, samples may be made among different groups of patient populations-i.e., among different age and sex groups or multiple medical ailments for fairness in analytical accuracy.
After training and validation with the execution of the cross-validation technique, the model will dynamically rate the hospitals in real time based on the continuous patient feedback through sentiment analysis. This process is constant, then enabling the model to learn and be updated with time for more accurate and current reflections of performance.

1.6 Scopes and Challenges

1.6.1 Scopes

This paper will be dedicated to the construction of an overall rating scale for hospitals, which shall be data-driven and will involve the service factors across all conceivable dimensions. The model will target capturing the commentary of the patients about the quality of the care in relation to the Doctors' performance, the experiences of the patients, and the overall service at the hospital.
The geographical coverage of hospitals throughout Bangladesh is of prime importance. The development and trial will be conducted with data collected from the Internet. This allows for diversity to make the data encompass the entire spectrum of the patients' experiences in various sections and types of hospitals.
Besides that, this research is going to develop a model in the healthcare sector that is scalable. The model will provide proof of efficacy in the domain of medicine, at the same time learn its feasibility to adapt into other service industries like education, retail, and customer service. On top of it, the broad applicability also makes it thrilling to enhance the process of assessment of Cross-industry customer satisfaction.
It will also be grounded in live feeds, whereby the ranking of hospitals will be interactive based on patient surveys and quality of care. This dynamic method for ranking hospitals is likely to urge hospitals to continually endeavor to improve their service, hence the healthy competition in the health sector.

1.6.2 Challenges

While the potential of this research is vast, there are several challenges that need to be addressed for successful implementation:

- Data Quality and Availability: Most of the challenges involve the volume and quality of data. The reviews may be incomplete, inconsistent, or biased on online platforms. It requires careful curation and validation to ensure that data indeed represents a wide variety of hospital experiences.

- Sentiment Analysis Difficulty: The subtlety of human feelings is hard for NLP to express correctly. Sarcasm, ambiguity, multiple complex medical terms used make it hard to analyze patients' feedback. There is a strong need to develop a sentiment analysis model which would understand such subtlety.

- Bias in review: The third.challenge is one of.bias that may be built in to online reviews. Only the very dissatisfied and very satisfied patients will bother to leave a review; these, of course, are by no means representative of the average patient. This may lead to an inaccurate rating and hence destroy the validity of the review. An even spread of patient.feedback over a broad range of demographics will be important to.preventing this.

- Model Generalizability and Accuracy: While Machine Learning models do improve with time, the task of training the model to predict hospital rankings in different geographies, hospital types, and patient segments is difficult. Ensuring that the model generalizes well rather than overfitting to a trend in the data is at the core of developing a dependable rating mechanism.

- Ethical Issues on Data Privacy: The acquisition of information from patients, particularly that related to health, is a highly sensitive issue that raises ethical issues on privacy and consent. This implies adhering meticulously to data protection policies while conducting research to deal with patient information in a responsible manner.
- Adoption and Implementation: Convince the hospitals to implement this new rating system-despite concerns, especially on the part of those institutions that are resistant to changes in their established configurations of operations. Making it useable and visibly beneficial will be what is needed to succeed in overcoming barriers to gaining buy-in from these hospitals.

Chapter 2

Literature Review

2.1 Preliminaries

Choosing the right hospital is the first and most crucial decision in receiving the proper treatment. The current Stars and SERPs rating systems don't emphasize consumers' emotions enough and provide a barebone and static result. We intend to use Google reviews as our primary source of data to construct a database and use Sentimental Analysis to generate a more realistic and honest rating system for the healthcare industry of Bangladesh. Our main objective is to construct a database and run different SA models to determine which model generates the most accurate results. This database can further be manipulated to fit the different needs of future researchers. In this section, we will summarize different articles related to our work. Additionally, we will explore different approaches and algorithms, such as BERT, LDD, SVM, DNN, ML, CNN, etc. by different researchers. In this paper, we will explore how Sentimental Analysis(SA) has been used as a solution to a more reliable and dynamic rating system in various industrial sectors, and how we can further develop an extended database and improve the healthcare sector of Bangladesh using SA.

2.1.1 Sentimental Analysis

Sentimental Analysis is a Natural Language Processing task that analyzes a piece of text to determine the sentiments and emotions that text is conveying. In the era of Social Media and different online-based platforms, we can easily access publicly available data on different industries and/or subjects. Most often than not, these data are heavily convoluted and, at the first glance, fails to convey the intended message. Through Sentimental Analysis, we can process a huge amount of text data, efficiently separating the underlying meaning, and its importance, and effectively understand any opinion being carried through the text. For this reason, Sentimental Analysis is also called Opinion Mining. In the earlier stages of Sentimental Analysis, the process was performed mainly using Machine Learning and Lexicon-Based Techniques. Nowadays, a lot of research has been done using a multitude of different methods. Different TRANSFORMER based languages, with a combination of Libraries, Hybrid ML, and DL, and different combinations of these techniques are being used to further study and enhance Sentimental Analysis to better suit it for regular user consumption. The Sentimental Analysis process is

formed of multiple sub-processes - data collection, data processing, data analysis, and data visualization are the core tasks. After a piece of text has undergone this process, it is stored according to the specific sentiment it demonstrates.

2.2 Review of Existing Research

In this paper[2], the authors comprehensively compared the different approaches in multilingual SA (corpus-based, lexicon-based, and hybrid) and outlined the advantages and disadvantages of these approaches. Furthermore, it is mentioned that the same approach can generate different results depending on the dataset used and the varying details of the dataset. Similar to the previous paper[2], this paper[12] studied Five Python and R libraries - NLTK, TextBlob, Vader, Transformers (GPT and BERT pretrained), and TidyText, additionally the author implemented Four machine learning models - Tree of Decisions (DT), Support Vector Machine (SVM), Naive Bayes (NB), and K-Nearest Neighbor (KNN) in the study to apply sentimental analysis techniques. According to the author, the TRANSFORMER model using the TidyText library has the best performance on the Twitter Sentiment Analysis Dataset[21] with the following accuracy, recall, and F1-score - 0.973, 0.944, and 0.877 respectively. Another article[4] proposed supervised machine learning and an extended sentimental dictionary to perform Chinese sentimental analysis. The Naïve Bayes classifier is used, resulting in an average precision score of 0.861, an average recall score of 0.843, and an average F1 score of 0.853. Furthermore, a detailed comparison is made on how adding field-specific sentiment words can improve the experiment's outcome rather than using only the basic dictionary. This paper[5] explored Deep Learning in the field of SA. DNN, CNN, and RNN algorithms were used to perform sentimental analysis on available datasets to determine their pros and cons. The outcome of the experiment is as follows, The DNN model is the simplest to implement and the fastest to train with an average overall accuracy of 0.75 to 0.80, The CNN model is also fast, comparatively a bit slower than The DNN model, but with higher overall accuracy above 0.80, and lastly, The RNN model is determined to be the most reliable when word embedding is applied with a caveat that its computational time is also the highest. But it is to be noted that, The RNN moded with Term Frequency - Inverse Document Frequency (TF-IDF) applied results in a lower accuracy at around 0.50. Similar to the paper [5], this paper[11] used DL and additionally implemented ML to perform a multi-class sentimental analysis in Urdu text. Two publicly available datasets UCSA-21 and UCSA[8] were used to train the models. The authors' proposed mBERT classifier achieved the most promising result. An accuracy, precision, recall, and F1-score of 0.8250, 0.8135, 0.8165, and 0.8149 respectively. Another research[16] on Urdu text was conducted using a hybrid dependency-based approach. The author reported the results using SVM, Logistic Regression, Multilayer Perceptron (MLP), and Decision Tree (DT) models as well as DL models combined with dependency-based rules. The author summarized by stating that the hybrid approach outperformed the standard state-of-the-art SA method by nearly 10%. This paper[6] focuses on Natural Language Processing(NLP) and its ability to extract meaningful information more effectively. Using NLP, the authors were able to identify positive, negative, and neutral comments from a patient's experience feedback dataset with a weighted average precision, recall, and F1-score of 0.83, 0.82, and 0.82 respectively. In another paper[7], the authors used Supervised

Machine Learning with Extended Lexicon Dictionary to analyze the sentiment in Bangla text. This research focused on machine learning and creating a lexicon data dictionary from Bangla datasets while also clearly outlining the research process from creating a word dictionary to data processing, tokenization and normalization, stemming, parts of speech tagger, and finally, the algorithms used. Another paper[13] focused on ML and Lexicon-based approaches in SA and the comparison of the approaches. While a detailed comparison has been presented by the author in his study, the author concluded the experiment by stating that the supervised ML approach yielded 0.85 accuracy in his research. Similar to the paper[7], this paper[14] also focuses on Bangla text, in the context of the Ukraine-Russia war, but instead of a normal approach where ML and DL are dominantly used, the authors of this article trained and used BERT transformer models, namely multilingual BERT (mBERT), Distil-mBERT, XLM-RoBERTa, XLM-RoBERTa-large, BanglaBERT, to explore the outcome of the different approach. Based on the accuracy of 0.86 with a 0.82 F1 score, BanglaBERT outperforms all baseline and other transformer-based classifiers. This paper[10], comparatively studies SA using NLP and ML on US Airline Tweet data on Kaggle as a dataset, named Twitter US Airline Sentiment[1]. A detailed comparison is made between classification algorithms with Bag-of-Words(BoW) and classification algorithms with Term Frequency-Inverse Document Frequency (TF-IDF), where SVM using BoW classifier has an accuracy of 0.77 with an F1-score of 0.75 and SVM using TF-IDF classifier has an accuracy of 0.77 with an F1-score of 0.74. Similarly, the Logistic Regression algorithm, using BoW resulted in an overall accuracy of 0.77 with an F1-score of 0.77, and using TF-IDF resulted in an overall accuracy of 0.77 with an F1-score of 0.76. This research[10] concludes that using the Bag of-Words classification algorithm generates a slightly better F1 score with the same accuracy in Sentimental Analysis. Another paper[20] researched a BERT-CNN-based approach in SA on an IMDB movie review dataset[9]. The author proposed a deep learning model in combination with BERT and CNN and compared the results with the sole BERT model. The result determines that the BERT-CNN model does not surpass the pure BERT model in terms of accuracy (BERT CNN accuracy 0.929 and BERT accuracy 0.930). Still, an interesting pattern is observed that the BERT-CNN model is good at identifying negative emotions while the pure BERT model is good at identifying positive emotions. Two more papers[19] [17] studied different TRANSFORMER based language models and made a comparison outlining the results. The later paper did Bangla Sentiment Analysis using Weighted and Majority Voted Fine-Tuned Transformers. Amongst all the models, DistillBERT and BnaglaBERT performed the best with Accuracy, Precision, Recall, and F1-score 0.701, 0.687, 0.701, and 0.701 for both respectively. This paper[22] proposed a different idea of foreign language sentimental analysis by translating those languages into a base language, English, and then, applying sentimental analysis. Since the tools available for English text sentimental analysis are more widely available, this explores that opportunity. The text translation was performed using LibreTranslate and Google Translate, and sentiment analysis was performed using Twitter-Roberta-Base, Bertweet-Base, GPT-3, and a novel Proposed Ensemble model. Amongst these, Google Translate, with the Proposed Ensemble model achieved overall highest scores across all metrics. An accuracy of 0.8671, a precision of 0.8091, a recall of 0.8122, an F1-score of 0.8106, and a specificity of 0.5713.Paper[15] evaluates multilingual sentiment analysis tech-

niques for under-resourced languages, emphasizing that deep learning models like CNN and RNN are used in more than 60% of the studies. The paper identifies the drawbacks of each technique used in sentiment analysis and concludes that a combination of deep learning models such as CNN, Bi-LSTM, and LSTM can significantly enhance multilingual sentiment analysis and proposes a DL learning framework independent of any MT-based methods. For the extraction of six individual emotions - happy, sad, tender, excited, angry, and scared - from Bangla text for both article and sentence. Paper[3] implements two machine learning techniques - Naive Bayes and Topical approach- where the Topcial approach surpasses the Naive Bayes with an accuracy of over 0.90. Paper[26] compares the semantic stability of BERT and Word2Vec across four time spans (3, 5, 10, and 20 years) using 20 years of People's Daily articles, concluding that BERT maintains stronger overall semantic stability over both short and long durations due to scoring higher SD, MTS and LNS and lower RSC in every period compared to Word2Vec.

2.3 Summary of Key Findings

Table 2.1: Overview of Studies on Sentimental Analysis

Ref.	Task	Model	Dataset	Metrics	Limitation
[15]	Evaluation of multilingual sentiment analysis techniques for under-resourced languages	Deep learning methods, Machine learning methods, Lexicon methods, Machine translation based methods	Dataset	N/A	Review conducted by only a few researchers, hence there may be bias in selecting studies for comparison
[3]	Extraction of emotion from Bangla text by Naive Bayes Classification and Topical approach	Naive Bayes Classification, Topical Approach	N/A	Topical Approach $\rightarrow$ Acc. > 0.90	Lack of smooth data and complexity of Bangla Language Need for more labeled data.
[26]	Comparison of Word2Vec and BERT in terms of semantic consistency	Word2Vec, BERT	Dataset	BERT's SD $>$ Word2Vec's SD, BERT's MTS $>$ Word2Vec's MTS, BERT's RSC $<$ Word2Vec's RSC, BERT's LNS $>$ Word2Vec's LNS	BERT may occasionally fail to capture gradual semantic shifts, BERT'S high computational cost

Ref.	Task	Model	Dataset	Metrics	Limitation
[2]	A comprehensive comparison between different approaches in multilingual SA and their advantages and disadvantages	Corpus-based, Lexicon-based, Hybrid	Dataset 1, Dataset 2	N/A	Small number of trained data
[4]	Chinese Text Sentiment Analysis through Supervised ML and Extended Sentimental Dictionary	Naive Bayesian , SVM	Dataset 1, Dataset 2, Dataset 3	Pre. 0.861, Rec. 0.843, F1 0.853	Weights of sentiment words need to be further refined Limited generalization of sentiment classifier.
[5]	Deep Learning in the field of Sentiment Analysis	DNN, CNN, RNN	Dataset 1, Dataset 2	DNN → Acc. 0.75 - 0.80, CNN → Acc. >0.8, RNN + TF-IDF → Acc. 0.50	Computational cost, Problem of aspect sentiment analysis
[6]	NLP to extract necessary and meaningful information from text	Neural Network with Keras DL Library	N/A	Pre. 0.83, Rec. 0.82, F1 0.82	Data imbalance an complexity due recipient's to lack of knowledge of "healthcare" terms, Easy to miss context of certain words
[7]	Supervised ML and Extended LLD to analyze sentiment in Bangla texts	Extended LLD, BTSC, SVM	N/A	N/A	Huge volume of sentimental dictionary required for a larger dataset
[14]	BERT TRANSFORMER model to analyze Bangla text in the context of the Ukraine-Russia War	mBERT, Distil-mBERT, XLM-RoBERTa, XLM-RoBERTa-large, BanglaBERT	Dataset	BanglaBERT → Acc. 0.86, F1 0.82	Shortage of Bangla datasets and class imbalance issue

Ref.	Task	Model	Dataset	Metrics	Limitation
[10]	Comparison between NLP and ML in Sentimental Analysis	SVM, LR with BoW	Dataset	SVM + Bow → Acc. 0.77, F1 0.75; SVM + TF-IDF → Acc. 0.77, F1 0.74; LR + BoW → Acc. 0.77, F1 0.77; LR + TF-IDF → Acc. 0.77, F1 0.76	Need for a more generalized and robust model for similar datasets
[11]	DL and ML to perform multiclass Sentimental Analysis in Urdu text	mBERT, SVM, LR, RF	Dataset	mBERT → Acc. 0.8250, Pre. 0.8135, Rec. 0.8165, F1 0.8149	Limited resources of data in Urdu
[20]	BERT-CNN based approach in Sentimental Analysis	BERT-CNN, BERT	Dataset	BERT-CNN → Acc. 0.929, BERT → Acc. 0.930	Research only focuses on text features, hence user characteristics like age, gender, or reputation are ignored
[12]	Comparison of different Libraries, Models, and Algorithms to determine their differences	NLTK, TextBlob, Vader, TRANSFORMER (GPT, BERT), Tidytext, Decision Tree, SVM, Naive Bayes, KNN	Dataset	TRANSFORMER + TidyText → Acc. 0.973, Rec. 0.944, F1 0.877	N/A
[19]	Different TRANSFORMER Moder comparison	N/A	Dataset	N/A	Majority of fine-tuned models on the IMDb dataset start to overfit at a certain number of epochs which may lead to biased models

Ref.	Task	Model	Dataset	Metrics	Limitation
[17]	Different TRANSFORMER Moder comparison in Bangla text Sentimental Analysis	RoBERTa (Base), Distill BERT, HF-PT BERT-1, HF-PT BERT-2, HF-PT BERT-3, Bangla BERT (Small), Bangla BERT (Large) , Bangla BERT (Base), Banglish BERT	N/A	Distil BERT → Acc. 0.701, Pre. 0.687, Rec. 0.701, F1 0.701; Bangla BERT → Acc. 0.701, Pre. 0.687, Rec. 0.701, F1 0.701	Limited resources of data in Bangla
[16]	Urdu text Sentimental Analysis using Hybrid dependency-based approach	DNN, CNN, SVM, LR, TFN	Dataset	N/A	Limited resources of data in Urdu, The issue of unclassified sentences
[22]	Translate foreign language to English and perform Sentimental Analysis	LibreTranslate, Google Translate, Twitter-Roberta-Base, Bertweet-Base, GPT-3, A novel Proposed Ensemble model	Dataset 1, Dataset 2, Dataset 2	Google Translate + Ensemble Model → Acc. 0.8671, Pre. 0.8091, Rec. 0.8122, Spec. 0.5713	Translation errors in machine translation tools
[13]	ML and Lexicon-Based approaches in Sentimental Analysis and their comparison	Supervised ML, Lexicon-Based	N/A	Acc. 0.85	Research mainly focused on ML techniques and Lexicon-based methods for sentiment analysis on Twitter data

2.4 Research Gap

There exist several research gaps when it comes to the development of hospital rating systems by utilizing machine learning and natural language processing. Firstly, most datasets do not extend beyond Dhaka-based hospitals which fail to train and design a more inclusive and generalized model for hospitals across Bangladesh. Another limitation is the lack of datasets in Bangla, and as a low-resource language, its sentiment analysis using multilingual models and machine translation tools might produce unreliable results. Thirdly, traditional sentiment analysis techniques harbor the problem of aspect-based sentiment analysis where they struggle with the association of sentiments with specific aspects of the hospital's performance e.g hygiene, discharge procedures, staff behavior. Additionally, implicit sentiment detection poses a major challenge when it comes to text that lacks explicitly expressed sentiments, which may lead to misclassification of reviews by many models, however, it can be addressed by deep learning models that take into account the context of the text for better detection of subtle expressions of sentiments.

Chapter 3

Requirements, Impacts and Constraints

3.1 Final Specifications and Requirements

This study will try to rank hospitals based on sentiment analysis, identify top services at each hospital, and detect common complaints and praises from the patient. It is observed that public hospitals have the best doctors, but often lack cleanliness and resources such as- space for bed, blood, and so on. The private hospitals focus on the quality of service, but some lack in expert employers. Overall, the management system of hospitals seems more arranged in the capital city than the other cities.

3.2 Societal Impact

Implementing a reliable sentiment-based ranking system for hospitals will address a significant concern for the people of Bangladesh: making informed healthcare decisions that can save both time and money. If the public begins to trust local health services more, they will be less inclined to seek care outside the country. This shift can reduce the overall economic burden on the nation, benefiting society both economically and medically. Additionally, doctors and hospital authorities will likely feel more motivated and responsible in their roles.

3.3 Environmental Impact

This study will have several positive environmental impacts in Bangladesh. First, by enabling people to make informed decisions, individuals will no longer need to visit multiple doctors for a single disease or health problem. This will reduce unnecessary hospital visits, leading to less crowded hospitals and making it easier to manage and maintain hospital waste. Since our dataset is derived from online resources, there was no need for paper-based feedback, which promotes carbon-free digital health records. Another significant improvement will be resource allocation for hospitals. Hospital authorities can identify where resources are most needed with the feedback and rankings provided. This ensures better resource distribution and sustainability.

3.4 Ethical Issues

The project aims to uphold all ethical polices regarding academic integrity. data confidentiality and such to ensure the development of a reliable hospital ranking sytem without threatening the safety of any individuals or organizations.

3.4.1 Ethical Considerations Concerning Data

- Publicly available patient reviews will be used for this project, maintaining data confidentiality.
- Personal information such as name and contacts will not be stored or disclosed. All reviews and data will be anonymous.
- Intentional manipulation of reviews of any hospital will be avoided.

3.4.2 Ethical Considerations in Artificial Intelligence

- The models may develop biases due to pre-trained data, hence, their performance will be monitored throughout the project for any observed bias in sentiment and aspect-based sentiment analysis
- The produced hospital rankings will be used for research purposes only and not be publicly marketed.
- All models, parameters, and thresholds used will be thoroughly documented for accountability.

3.4.3 Academic and Professional Responsibilities

- All code, content, design process and analysis will be developed and conducted by the thesis group with proper citation of tools and libraries used, codes reused or adapted, and existing research papers that helped, maintaining academic integrity and responsibility.
- Results and evaluation of them will be presented as they are, without exaggerating or understating.

3.5 Project Management Plan

The objective of this project is to collect online reviews of hospitals across the eight divisions of Bangladesh and, based on them, generate scores of each hospital in 4 aspects: cleanliness, affordability, treatment quality, and service quality. Additionally, we create a system that produces the top 3 hospitals in each of the aforementioned 4 aspects. The scores will present the performance of the hospitals in those 4 areas, which, along with the ranking system, will enable patients or visitors to determine the best hospital that will meet their needs, whether they are looking for affordable treatment or treatment in a hygienic environment.

3.5.1 Analysis and Design Methodology

The online reviews of all the hospitals across the entire Bangladesh were scraped from Google Maps using a web scraper tool and built into a massive dataset comprising of 8 parts, each part containing reviews from its corresponding division among the 8 divisions. Every review was manually labelled with either of the 3 sentiments - +1 (positive), 0 (neutral), and -1 (negative). Each member of our group voted for one sentiment per review, and if necessary (if it were a tie), we would repeat this voting process several times. Thus, each review was labelled with the most voted sentiment to minimize bias.

Initially, we trained 5 machine learning models - Decision Tree Classifier, Support Vector Machine (SVM), Logistic Regression, Random Forest, and Gaussain Naive Bayes - and 2 deep learning models - BERT and CNN - on the dataset to perform sentiment analysis. The performance of these models was then evaluated and compared using 4 metric - accuracy, F1 score, precision, and recall. Additionally, we also produced confusion matrix for each of the models to display the most common positive, negative, and neutral words for a division used in the dataset.

In order to achieve our main objective - generating scores of each hospital in terms of cleanliness, affordability, treatment quality, and service quality - we implemented a zero-shot aspect-based sentiment pipeline on the preprocessed reviews. Zero-shot uses the XLM-RoBERTa-large model and the NLI (Natural Language Inference) to perform aspect-based sentiment analysis without needing to be trained on the dataset since it is pretrained on multilingual text and fine-tuned on the XNLI dataset.

The aspects that were present and absent in a review were detected by zero-shot. It did so by producing a score for each aspect, e.g cleanliness score, and compared it with a threshold value to determine whether the aspect was mentioned in the review. Then, the aspects that were detected were assigned the sentiment of the review itself, and those that were not detected were assigned 0. Hence, we obtained 4 scores for 4 aspects in each review e.g cleanliness score, affordability score, treatment quality score, and service quality score.

The fraction of reviews that mentioned an aspect, e.g cleanliness, and the average score for that aspect were found. The average score was then mapped into a value in the range 0-1 and multiplying that mapped value with the fraction of reviews produced a raw aspect score per hospital, e.g raw cleanliness score. The 4 raw computed scores were normalized to produce scores in the range 0-1 for fair comparison among the hospitals. These scores were used to rank the top 3 hospitals per aspect. For each hospital, a base score was computed using the weights of all 4 aspects and their normalized scores, and another confidence value calculated. Finally, a final score per hospital was generated using the base score and the confidence score, which decided the rank of the hospital. These scores were used to produce the top 10 hospitals in terms of all the aspects.

Besides review-level zero-shot aspect detection, additionally, we also implemented sentence-level zero-shot aspect detection. The key difference between the two is that the former assigned the review's sentiment to the aspects that were present, whereas the latter detected aspects per sentence and assigned them the sentiment of that particular sentence. Moreover, we conducted a comparative analysis between the two by producing the biggest differences (rises and drops) in ranks generated from using review-level aspect detection and sentence-level aspect detection of hospitals. The

differences in their performances were also assessed in terms of precision, recall, F1 score and agreement rate. It was found that sentence-level zero-shot aspect detection more accurately ranks the hospitals since it preserves aspect-specific sentiment by assigning the detected aspects per sentence the sentence-specific sentiment. Hence, it reduces the chances of missing subtle aspects in a review and identifies the sentiment of the aspect more accurately in case of a mixed review. However, review-level zero-shot aspect detection carries a higher chance of missing aspect details and the aspect sentiments may get averaged which results in less accuracy in ranking hospitals.

3.5.2 Resources

Training the machine learning models - Decision Tree Classifier, Support Vector Machine (SVM), Logistic Regression, Random Forest, and Gaussain Naive Bayes - on CPU will be efficient enough, however, using Colab Pro will faciliate the process due to its longer sessions and higher RAM to produce larger grids. The deep learning models - BERT and CNN - require GPU without which, their training would consume much more time on a regular CPU. Through Colab Pro, we can access T4/A100 High Ram GPUs, which will reduce epoch time and improve throughput. All the csv files, produced plots and so on will saved to Google Drive for easy access and storage. Apart from Colab Pro costing us $9.99, there is not much expense as free resources are readily available for the project.

3.5.3 Timeline

The most time-consuming part of this project will be to scrape reviews of all hospitals across all the 8 divisons of Bangladesh and then, manually labeling them with sentiment scores. These will take around a month to complete if we focus heavily and consistently on the project. The rest of the work will involve developing the code to train the models and produce results which will take less time with available resources such as GPUs.

3.6 Risk Management

Several risks that the project might pose are:

- The online hospital reviews collected from Google Maps may be inconsistent, incomplete, contain spelling mistakes or punctuation errors which may degrade the performance of the model and produce less accurate results. Removing emojis, blank rows, punctuation marks and so on could make the reviews much cleaner before feeding them to the model.
- A number of hospitals have barely any reviews or very less reviews compared to others which can lead to unstable or unreliable rankings. A minimum review filter could be applied during the ranking of hospitals in order to exclude hospitals that have a review count below the minimum review value.
- Zero-shot model is computationally expensive and its implementation on our huge dataset might be time consuming or cause memory errors. This can be mitigated by GPU acceleration.

- Collecting reviews, manually labeling them with sentiments, implementing models on the entire dataset, testing and tuning may take longer than planned. To minimize the risk of delay, we will prioritize working with the hospitals of the capital city, Dhaka, due to it having the most number of hospital visitors and reviews and focus on outputting the hospital rankings by implementing review-level zero-shot aspect detection first (rather than sentence-level) before further refinements.

3.7 Revised Project Plan and Rescheduling

Throughout the project, we may encounter several challenges which can lead to adjustments to the original plan in order to complete it within the given timeline. The following revisions could transpire:

- Simultaneous implementation of zero-shot model on the hospital reviews of all 8 divisions of Bangladesh may turn out to be computationally slow. It can be adjusted by running zero-shot on a single division first and debugging any issues before moving on to the other divisions. This will greatly reduce chances of failure regarding all the 8 divisions at once.
- Implementing sentence-level zero-shot aspect detection is complex and time-consuming, since it looks for aspects per sentence in a single review, therefore, will take time to process all the reviews. Hence, review-level zero-aspect detection could be applied in its stead, which detects aspects per review, thus, processes faster.

3.8 Economic Analysis

The project aims to build a cost-efficient hospital ranking system by applying aspect-based sentiment analysis on reviews of hospitals across Bangladesh through utilizing mostly free and open-source tools.

Resource	**Estimated Cost**
Hardware (Personal laptop)	Existing
Software (Python, Transformers, Scikit-learn, NLTK, Pandas)	Free
Colab Pro	$9.99

3.8.1 Cost-Benefit Estimation

Aside from a small cost in GPU to implement zero-shot and deep learning models, the most of the cost will amount to the time investment into manually labeling the dataset concerning all the 8 divisions of Bangladesh, data preprocessing and model testing. However, these costs will lead to greater benefits:

- An automated, scalable hospital ranking system.

- Zero-shot model does not require to be trained anew unlike some other models.
- A reusable pipeline for similar or other domains.

Chapter 4

Methodology

4.1 Work Plan

The work plan started with the collection of data. As the main purpose of this research is to provide a reliable, unfiltered, and honest rating system that can help users with making the right decision, we need a huge amount of data. We began with collecting and storing that data. The data was be gathered from the Google review section's comment portion. After collection, we started data processing. In this part, we translated the data into a base language (English), tokenized and normalized the data, and moved on to the analysis part. Multiple machine learning models (Decision Tree Classifier, SVM, Logistic regression, Random forest, Gaussian Naive Bayes), deep learning models (CNN, BERT) and large language models (xlm-roberta-large-xnli, twitter-xlm-roberta-base-sentiment, XLM-Roberta-Base) was be used to analyze the data and the outcome was be recorded to determine individual model's advantages and disadvantages. Further we analyzed entire reviews by ABSA(Aspect- Based Sentiment Analysis) , to extract sentiment for specific aspects. This helped fine-grained ranking of hospitals based on different service categories. Based on the best-performing method, we had a system that can analyze a variety of human emotions from text data and generate a result accordingly.
A similar work[18] has been done before, but that research used a smaller dataset, only the reviews of the hospitals in Dhaka City were used as the data for that research[18]. We intend to create our own expanded dataset and replicate and improve the results. Future work will include on-site offline data collection at the very end, processing and analyzing that data in a similar fashion and making a detailed comparison on which hospitals have improved its healthcare facilities over the years

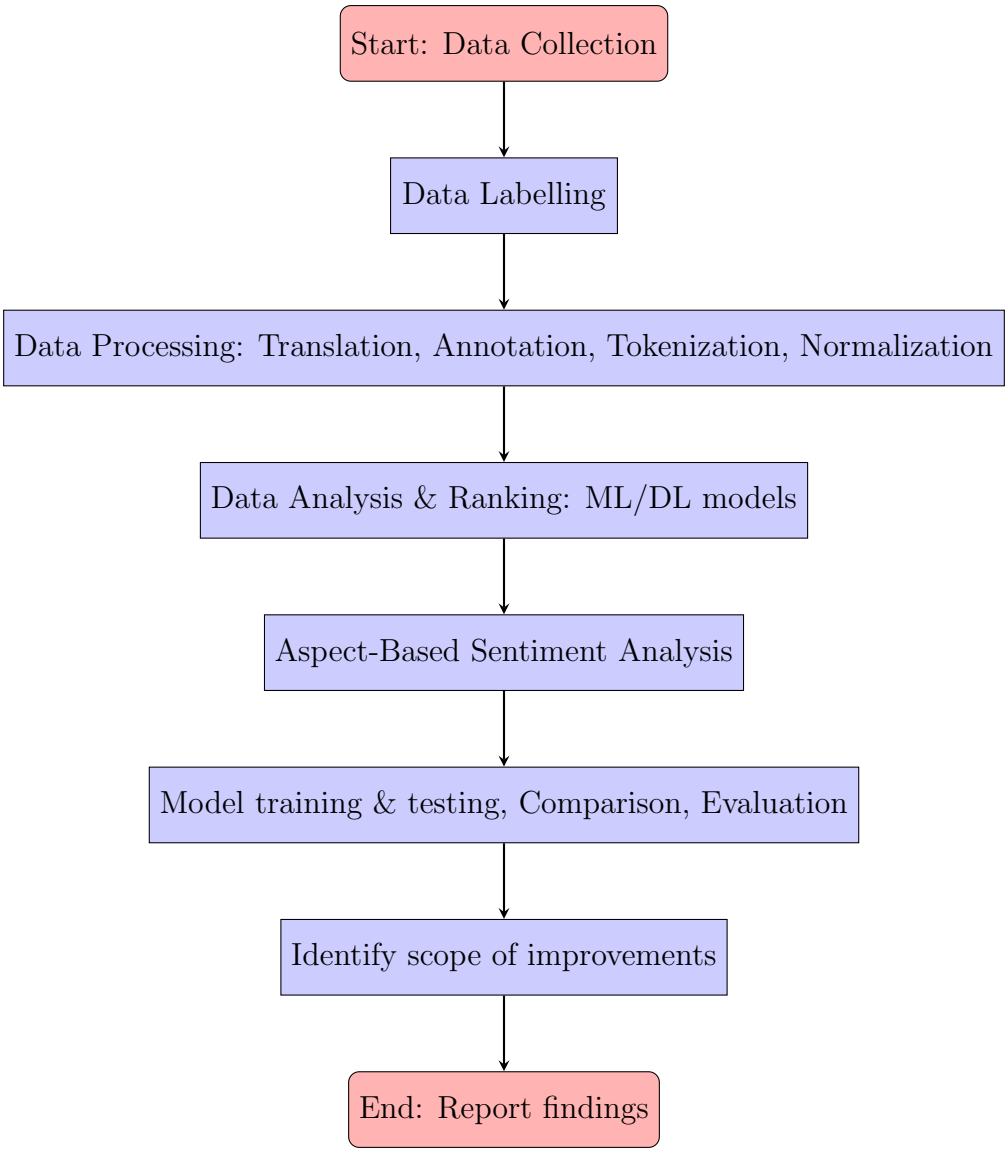

Figure 4.1: Work Plan Flowchart

4.2 Dataset

The Dataset was collected using Python's Selenium library by scrapping the Google Reviews of all the hospitals in Bangladesh. After the data collection was completed, data labelling was done in muliple steps. Firstly, the was divided equally amongst the researchers. Each researcher labelled a portion of the dataset. But one major flaw of this process is, it could introduce bias to the labelling process. Thus, to minimize the bias as much as possible, we choose an "Anonymous Voting" process and relabelled the whole dataset again. This time, each researcher labelled the whole dataset anonymously. Afterwards, the labelled data from each member was collected and the majority vote of a single review was selected as the final label/sentiment for that review. Finally, this process was repeated again, to resolve any ties between the positive, negative, and neutral sentiment for any given review.

4.3 Data preprocessing

Data preprocessing is a crucial step prior to the implementation of machine learning models on a dataset. Oftentimes, real-world data harbor imperfections, such as punctuations, emojis, missing values, outliers, features with different scales, and so on, which, if not rectified, might lead to several issues. The model might learn incorrect patterns in the presence of missing values, give less accurate results if punctuations and emojis add noise, overfit to the outliers, become more biased to features with larger values - all of these degrade the performance of a model. Hence, for the best results, our dataset of Google reviews of hospitals across Bangladesh has undergone the following stages of preprocessing.

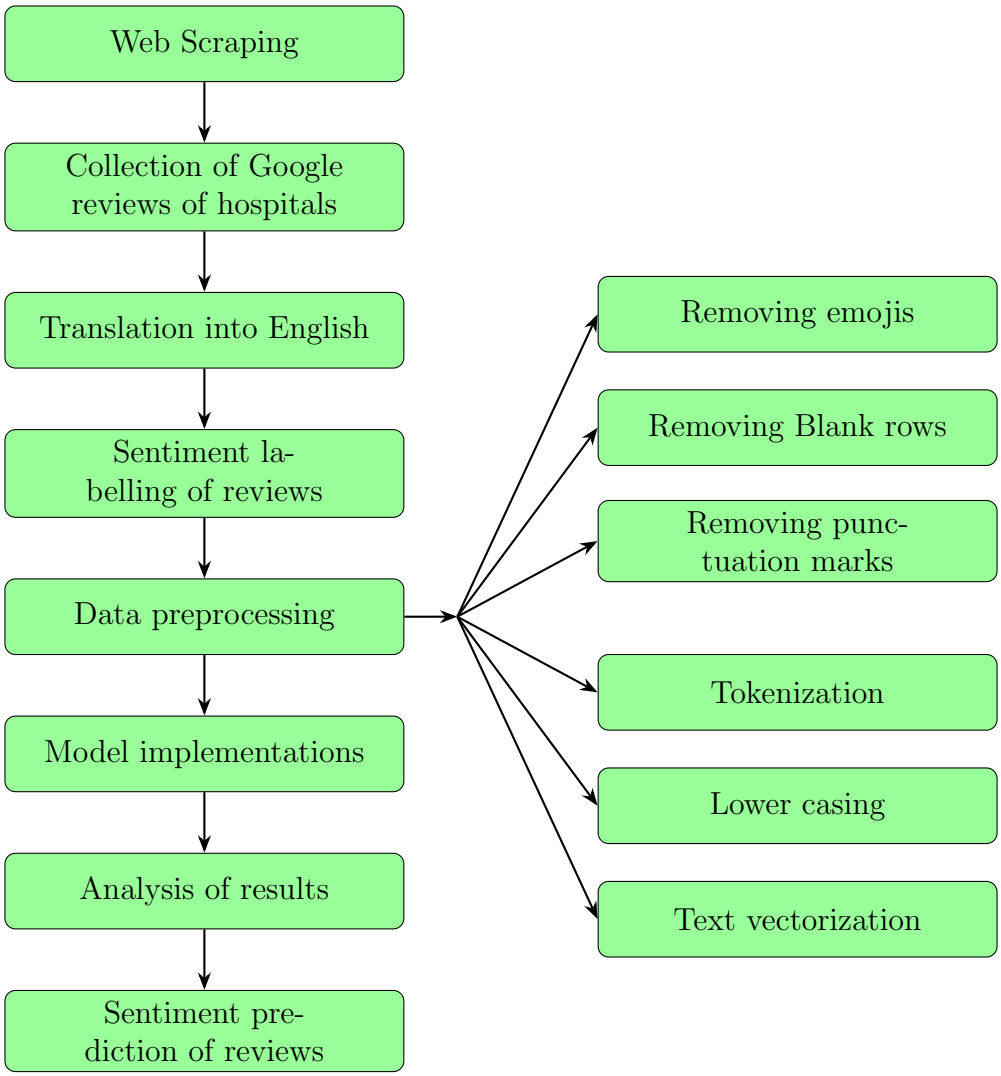

Figure 4.2: Work Flow diagram

4.3.1 Removing Emojis

A number of emojis were used by the people to express their sentiments regarding the hospitals. For example, “smiley” emojis expressed a positive experience, whereas “angry” emojis conveyed dissatisfaction regarding a hospital’s sector. All the emojis were removed from the dataset.

4.3.2 Removing Blank Rows

Removing emojis and duplicate reviews leaves unnecessary blank rows, which were also removed from the dataset.

4.3.3 Removing Punctuation Mark

For the initial Sentiment Analysis part, we used multiple ML models and some DL models. During this part of the work, there were a significant number of punctuation marks that were also used to convey a variety of sentiments regarding a hospital, similar to how emojis were used. All the punctuation marks were removed from the dataset. Afterwards, for aspect detection, our pipeline relies on natural sentence boundaries and transformer tokenization. Thus, we preserve punctuation, only normalize whitespace/case, strip HTML/zero-width characters, and lightly compress elongations (e.g., "soooo"→"soo").

4.3.4 Tokenization

The dataset was tokenized, where each review was broken down into smaller units called tokens. The tokenized input was fed into the machine learning models for better processing of the reviews.

4.3.5 Text Vectorization

Machine learning models require the input in a numeric format rather than just raw text to be able to understand and learn from it. Hence, the dataset has undergone text vectorization, where words have been converted into numbers for the models to perform sentiment analysis.

4.3.6 Transformer Tokenization

We use the model's own tokenizer (XLM-RoBERTa) to split text into subword pieces. Doing this, even rare words, misspellings, and Banglish code-mixing are handled by composing smaller chunks, so there's no out-of-vocabulary issue. The tokenizer outputs token IDs and an attention mask (real text vs padding), which go straight into the Transformer to get contextual embeddings. In practice we tokenize per sentence, keep punctuation, and pad/truncate to a sensible max length.

4.3.7 Lower Casing

All the reviews in the dataset were converted to lowercase for several reasons. Words such as "Patient" and "patient" should be treated the same; however, machine learning models might perceive them differently, if not converted to either the uppercase or lowercase. Converting the reviews to uppercase increases the vocabulary size of the models, so lowercase is preferable as it reduces memory usage, enhances performance, and has better readability than uppercase.

4.4 Model Specifications

4.4.1 Decision Tree Classifier

A Decision Tree organizes data by splitting it into subsets based on the values of various features. Each internal node evaluates a specific feature(such as a word), while the leaf nodes indicate the class labels, which, in our case, are “Positive”, “Negative” and “Neutral”. It works for both categorical and numerical data and is the best choice for comparative analysis with other models.The model employs metrics like gini impurity or entropy or information gain to identify the most suitable features to split on. The model, paired with TF-IDF or Bag-of-Words(BoW) vectorization, ensures a powerful text classification, offering straightforward understanding and visualization. For our dataset, we used TF-IDF as it is better than BoW because of its weighting mechanisms. So, rare but important words are assigned higher scores while common words are assigned lower scores.
How Decision Tree works can be illustrated with an example review picked from our very own dataset: “Need to improve services.” We assume that the vocabulary constructed from the most common and important words from the dataset is [“good”, “bad”, “worst”, “improve”, “service”, “need”, “waiting”], depending on which, the review is vectorized to [0.0,0.0,0.0, 0.61,0.44,0.53,0.0]. The decision tree splits based on the specified metrics using the assigned weights. This process is visualized in the accompanying figure. Based on the prediction path taken by the model, it interprets this as a signal of dissatisfaction, leading to a prediction of negative sentiment.

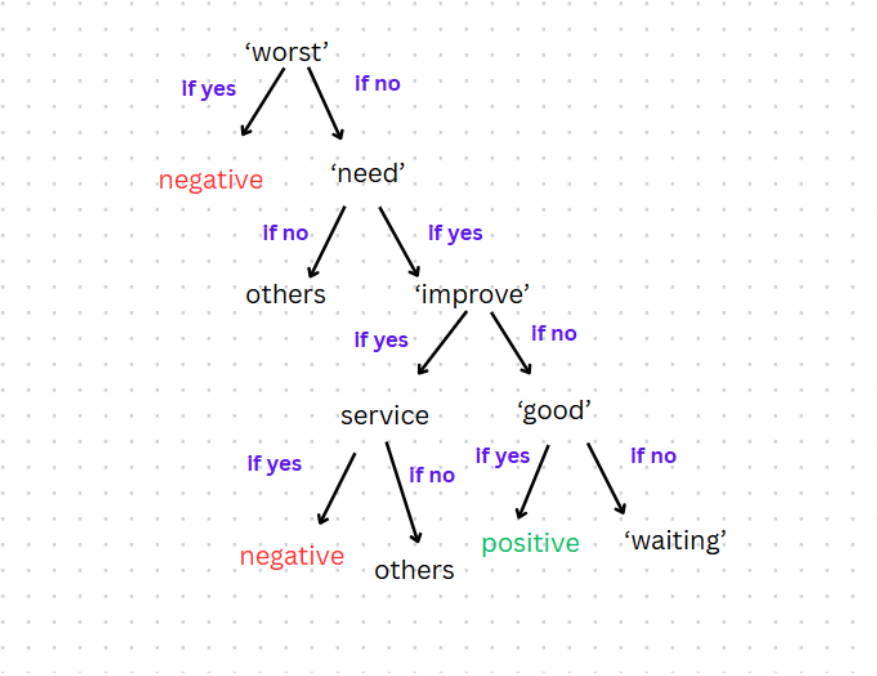

Figure 4.3: Example TF–IDF decision tree splitting on words like “need” and “improve.”

4.4.2 SVM

The Support Vector Machine identifies a hyperplane that optimally separates different classes in a high-dimensional space. It is particularly well suited for handling sparse data, such as text vectors. Additionally, it can use kernels to change data into higher dimensions. SVM does not use conditional logics like decision trees, rather it uses a hyperplane equation such as w1*x1 + w2*x2 +...+wn*xn + b, where 'w' is the weight vector of a word learned during the training phase, 'x' is the TF-IDF score of the word, and 'b' is a bias value. The model predicts a sentiment (class) based on the comparison between the computed value and a threshold parameter.
As with the former models, the entire dataset is preprocessed and vectorized before before feeding it to the SVM model. The operations of SVM are described using the previous example of our dataset: "Need to improve services" with the constructed vocabulary ["good", "bad", "worst", "improve", "service", "need", "excellent"]. The review is vectorized to [0.0,0.0,0.0, 0.61,0.44,0.53,0.0]. During the training phase, the model learns that high scores for "improve", "need", and "bad" often lead to negative sentiment and high scores for "good", "excellent" lead to positive sentiment. The vectors are mapped into feature space and the model checks which side of the hyperplane the features (words) fall on. The boundary space is visualized in the accompanying figure. Plotting the vectors, we see the words of our example fall on the negative side. Thus, SVM predicts the review as a negative sentiment.
One drawback of this model is that it struggles with neutral sentiment especially if there is an overlap in feature space between classes. When a sentence has both positive and negative words, it may not confidently classify it, because neutral reviews often lie near the boundary. To overcome that, we use confidence score (SVC) where low confidence is identified as neutral sentiment. Overall, SVM often works surprisingly well on text data and predicts accurately.

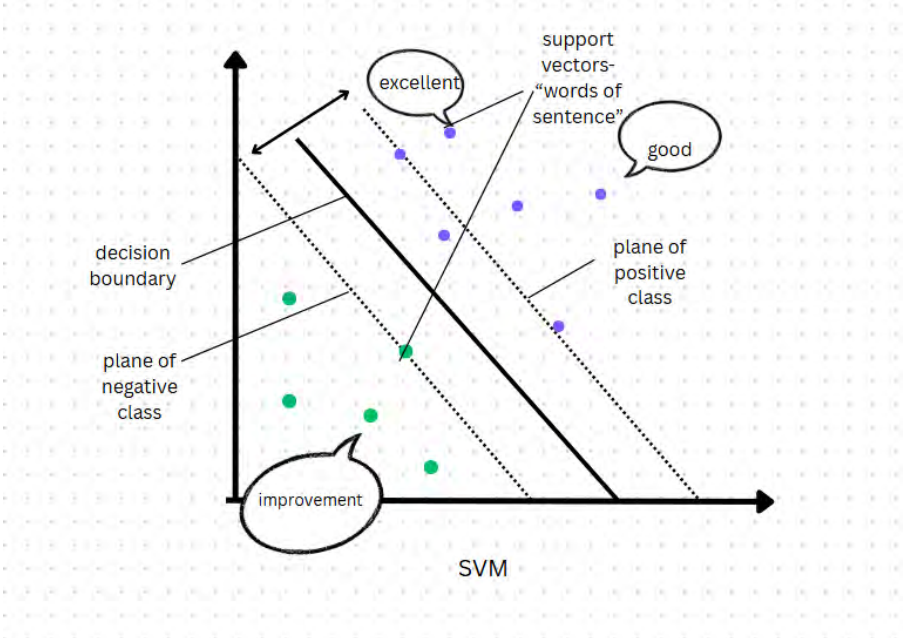

Figure 4.4: SVM separating positive and negative word vectors via a maximal margin hyperplane.

4.4.3 Logistic regression

Logistic regression is a supervised machine learning algorithm that effectively categorizes data into distinct groups. Unlike linear regression, it predicts a continuous value, a probability that an input belongs to a specific class. It is particularly advantageous for scenarios with two possible outcomes. We modified the model into three class classifier, according to our dataset. During training, the model learns from the labeled dataset, while in testing, it predicts outcomes for new, unseen data. Logistic regression uses a continuous loss function along with gradient descent to approach the optimal solution.
The mechanism of the model can be described using the same example - "Need to improve services" vectorized to [0.0,0.0,0.0, 0.61,0.44,0.53,0.0] using the constructed vocabulary ["good", "bad", "worst", "improve", "service", "need", "excellent"]. The model learns the weights for each feature and biases from training, and computes a linear score using the equation z = w.x + b. This score is mapped to a probability using the sigmoid function. A comparison between the value of probability and a threshold value classifies the review as "Negative".

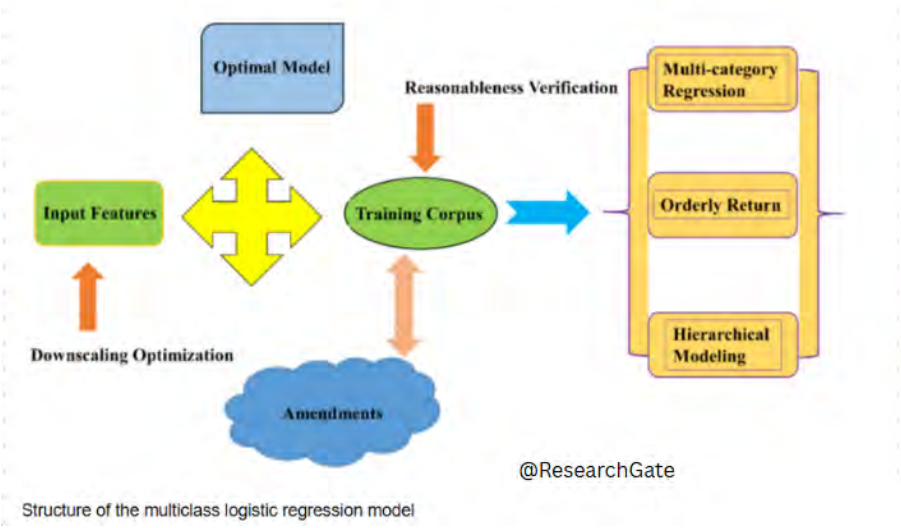

Figure 4.5: Multiclass logistic regression with a softmax output for three sentiment labels.

4.4.4 Random forest

Random forest is another machine learning algorithm that can be used for classification in the context of sentiment analysis. It is called a forest due to a number of decision trees at work where each tree perceives the review differently and predicts a sentiment. The sentiment that is voted by the majority of the trees is the final, decided sentiment for that particular review.
Each tree picks a random subset of features from the training data to train itself, and chooses the best feature(s) from that subset to split on based on the information gain of the features. The higher the information gain of a feature, the more suitable it is to split on.If we take the example review, "Need to improve services", then, using TF-IDF, it is vectorized to [0.0, 0.0, 0.0, 0.61, 0.44, 0.53, 0.0] using the same example vocabulary: ["good", "bad", "worst", "improve", "service", "need",

"waiting"], which has been assumed to be constructed from common words in the dataset. Now, to predict the sentiment of this review, if we consider three trees at work, then we can assume that tree 1 picks the features, "improve", "need", "waiting", tree 2 picks "good", "bad", "service", tree 3 picks "worst", "waiting", "need". Each tree will now check different conditions on word weights, follow different prediction paths that the conditions create and, in the end, produce sentiments that may differ. Suppose, tree 1 produces "Negative", tree 2 produces "Neutral" and tree 3 produces "Negative". Since the majority of the trees predict the review to be "Negative", the final output is "Negative".

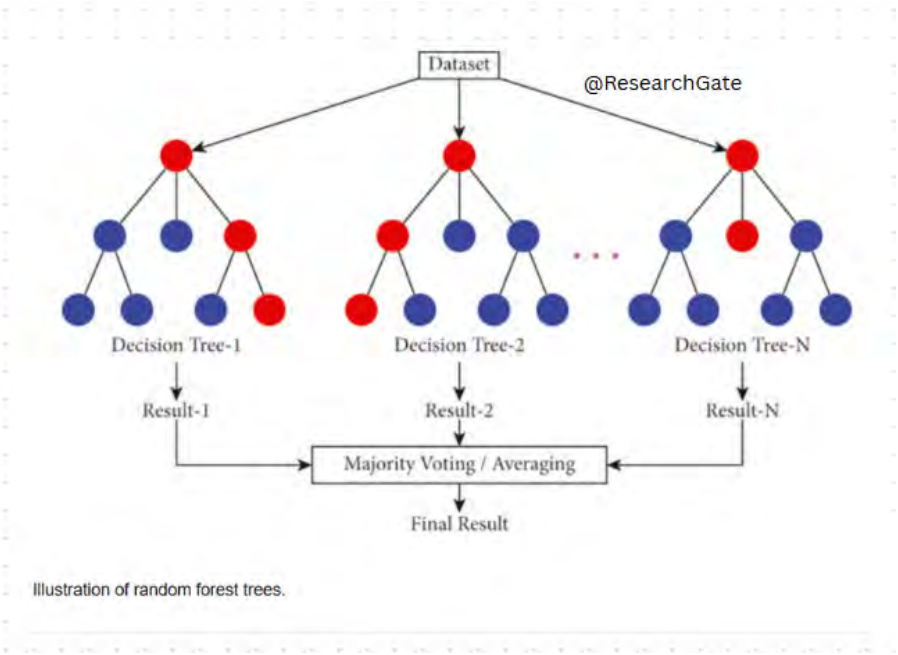

Figure 4.6: Random forest ensemble of decision trees with majority-vote aggregation.

4.4.5 Gaussian Naive Bayes

Gaussian Naive Bayes is another machine learning algorithm that can be used to classify a review into one of the three sentiments: Positive, Negative, Neutral. The algorithm implements Bayes' Theorem and naively assumes that the features (words) in a review are independent of each other. Specifically, it assumes that the values of the features are normally distributed. It studies the mean and standard deviation of each feature within each class, the prior probabilities of each class and calculates the probabilities of a given input belonging to the "Positive" class, "Negative" class and "Neutral" class, and after comparison, decides that the class (sentiment) that produces the highest probability becomes the final output.
As with all the other models, the preprocessed reviews are first vectorized using TF-IDF before feeding the vectorized input to the Gaussian Naive Bayes model. Now, suppose, we want to predict the sentiment of the example review, "Need to improve services", which has been vectorized into [0.0, 0.0, 0.0, 0.61, 0.44, 0.53, 0.0] using TF-IDF. Only the features "improve", "service" and "need" have values which are not "0.0". The model will calculate P(feature | class) for each feature within each class using their values and multiply them with their respective prior probabilities:

$P(\text{Positive} \mid \text{Input}) = P(\text{Positive}) * P(\text{improve} \mid \text{Positive}) * P(\text{service} \mid \text{Positive})\ P(\text{need} \mid \text{Positive})$

$P(\text{Negative} \mid \text{Input}) = P(\text{Negative}) * P(\text{improve} \mid \text{Negative}) * P(\text{service} \mid \text{Negative})$
$P(\text{need} \mid \text{Negative})$

$P(\text{Neutral} \mid \text{Input}) = P(\text{Neutral}) * P(\text{improve} \mid \text{Neutral}) * P(\text{service} \mid \text{Neutral})\ P(\text{need} \mid \text{Neutral})$

The sentiment with the highest resultant probability will be attached as the final sentiment to the example review, "Need to improve services."

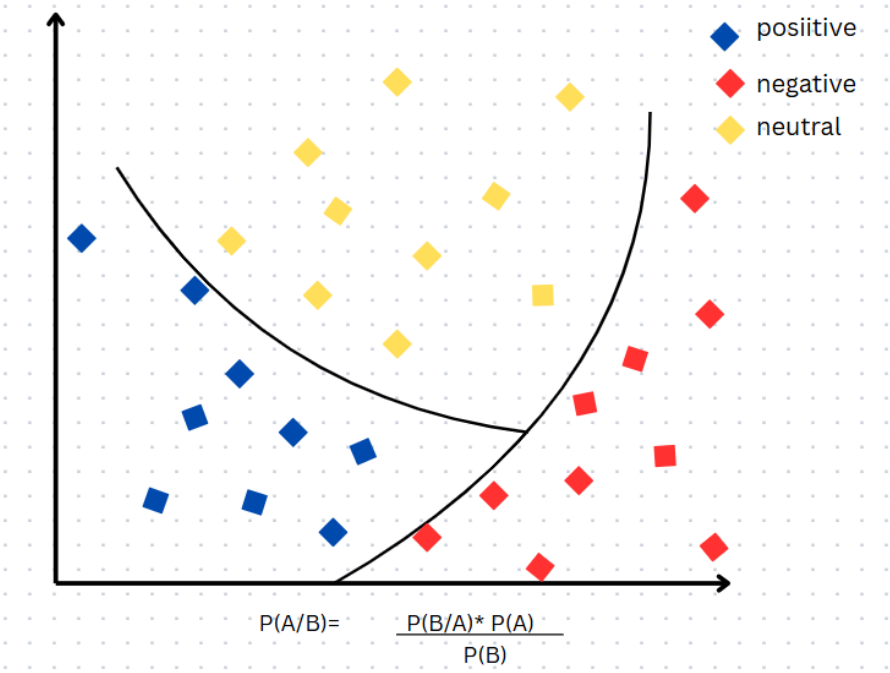

Figure 4.7: Gaussian NB decision boundaries partitioning positive, neutral, and negative regions.

4.4.6 CNN

Although CNN(Convolutional Neural networks) model is originally designed for images, they work very well for text classification, including sentiment analysis . It extracts important features from word embeddings, which are special representations of words as vectors. It has mainly four layers- input, convolutional, pooling, fully connected. Through these layers it captures local patterns, detects features such as "bad service", "good hospital", "not clean" and so on. This model performs well for short to medium length texts.
We explain the stages of CNN through the example from our dataset: "Need to improve services". At first we convert text into word embeddings - [w1, w2, w3, w4] - which captures meanings of words. Next, the convolutional layer uses filters called "kernels" to extract features or word combinations from embedded words.

For example, [‘need’, ‘to’], [‘to’, ‘improve’], [‘improve’, ‘services’] with corresponding feature scores [0.12, 0.45, 0.89]. Pooling layers then prune and keep only the feature with max score. This step helps to prevent overfitting. Finally, fully connected layers predict sentiments, assuming scores of positive: 0.05, negative: 0.85, neutral: 0.10. Hence, the final prediction is negative which scores the highest.

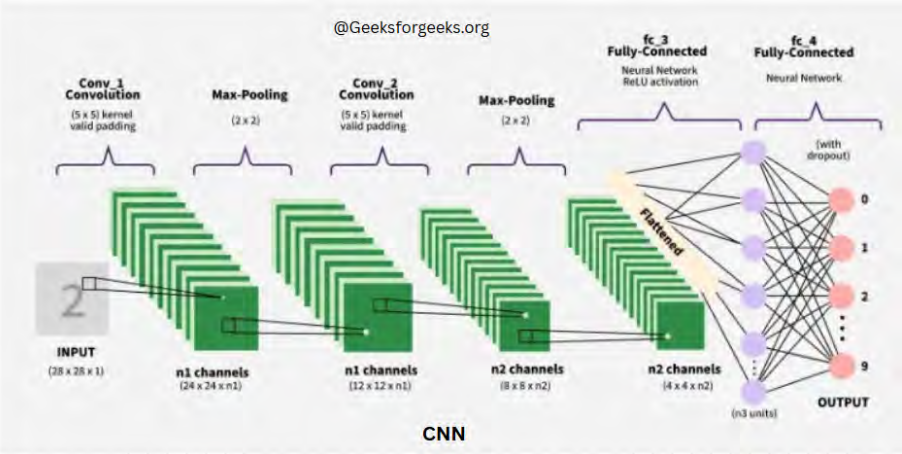

Figure 4.8: Text CNN: embedding, convolution + pooling, and fully-connected layers with softmax.

4.4.7 BERT(Bidirectional Encoder Representations from Transformers)

BERT(Bidirectional Encoder Representations from Transformers) is a highly advanced deep learning model, developed by Google, that excels in understanding the context of words within a sentence by taking into account both preceding and following words. This bidirectional approach allows it to achieve state-of-art results in many tasks. Unlike traditional language models that process text in a unidirectional manner (either left-to-right or right-to-left), BERT evaluates the entire context of a word by considering all words in a sentence simultaneously. For our dataset, we use BERT, to capture the depth of emotions or sentiments in the hospitals' reviews. To reach our goal of making an advanced rating system for hospital rankings, this model contributes significantly.
To illustrate how BERT works, we take an example review from our dataset: “Need to improve services.” At first, it takes the whole sentence as input, then tokenizes it into subword units:

```
['[CLS]', 'need', 'to', 'improve', 'service', '#s', '[SEP]']
```

Next, it embeds each token into a dense vector and applies multi-head self-attention layers where every token attends to every other token. This means BERT understands how “improve” relates to “need” and “services.”
Finally, we assume the model predicts

$$\text{positive} = 0.15, \quad \text{neutral} = 0.20, \quad \text{negative} = 0.65.$$

Among these, negative sentiment has the highest value, so it is selected as the final prediction. Hence, BERT outputs contextualized embeddings for each word. Even if “improve” appears in some positive feedback later, BERT uses “context” to decide.

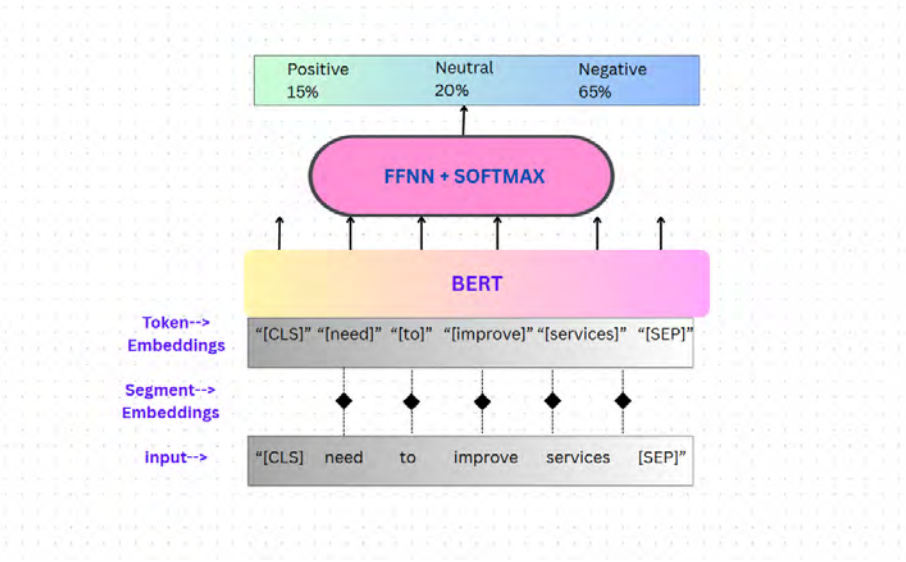

Figure 4.9: BERT classifier: token and segment embeddings → Transformer encoder → softmax head.

4.4.8 Zero-shot "Teacher" (XLM-RoBERTa-Large, XNLI)

We use a multilingual natural-language-inference (NLI) model, joeddav/xlm-roberta-large-xnli, as a zero-shot aspect detector at the sentence level. For every sentence, we ask four short hypotheses like "This sentence is about cleanliness and hygiene," "... affordability and cost," "... treatment quality and medical care," and "... service quality and staff behavior." The model returns a probability (0-1) for each hypothesis, effectively telling us how much that sentence supports the idea that it's about the aspect. Because it's multilingual and trained on entailment, it generalizes well to English + Banglish/code-mix without any task-specific labels. We chose this teacher because it gives good coverage with zero annotation, and operating at the sentence level avoids the "topic dilution" that happens when whole reviews mix multiple issues.

4.4.9 Distilled Student Classifier (XLM-RoBERTa-Base)

To make inference fast and scalable, we distill the teacher into a lighter student model, xlm-roberta-base, that directly predicts four aspect probabilities from a sentence in one forward pass. We train it using the teacher's soft scores as targets with a multi-label BCE-with-logits loss, lightly weighting clearer examples higher. The result is a compact model that mimics the teacher's behavior but runs much faster, practical for large datasets or CPU-only environments. We chose this student to keep the final system efficient, consistent, and deployable, while preserving most of the teacher's accuracy.

4.4.10 Sentence-level Sentiment (CardiffNLP XLM-R Base)

Alongside aspect detection, each sentence goes through cardiffnlp/twitter-xlm-roberta-base-sentiment, which classifies negative/neutral/positive. We map these to 1/0/+1 so that every aspect mentioned can be paired with a direction of opinion. This model

is robust to short, noisy text and code-mixing, which matches the style of user reviews. We chose it so our analysis doesn't just say what is being discussed (e.g., cleanliness) but also how people feel about it.

4.4.11 Thresholding / Calibration (per-aspect)

Both student and teacher output probabilities, then we transform them into explicit "present/absent" aspect tags with per-aspect cutoffs. We start with a good default and then automatically tune the cutoff of each aspect from the data (Otsu's method), with the decision boundary aligning with the distribution of true scores. This simple calibration step makes the tags consistent and understandable across aspects without the need for human-labeled thresholds.

4.5 Implementation of Selected Design

4.5.1 Aspect Detection Pipeline

Our pipeline starts with a sentence-level zero-shot *teacher* based on the multilingual XLM-RoBERTa-Large model fine-tuned for natural-language inference joeddav/xlm-roberta-large-xnli. For every sentence in a review, we pose four short hypothesis statements; "This sentence is about cleanliness and hygiene," "...affordability and cost and billing," "...treatment quality and medical care," and "...service quality and staff behavior and waiting time." The teacher returns a probability in $[0, 1]$ for each hypothesis, which we store as zs_<aspect>_score . Working at the sentence level avoids the "topic dilution" that happens when whole reviews discuss multiple issues at once, and the multilingual backbone is robust to English–Banglish code-mixing. We run teacher inference in small batches (typically 8 sentences per batch) to fit Colab-scale GPUs.

To make the system practical for large datasets, we distill the teacher into a lighter *student* multi-label classifier built on XLM-RoBERTa-Base. The student takes a sentence as input and predicts four probabilities directly (one per aspect). We train it to mimic the teacher's soft scores using a multi-label BCE-with-logits loss with sample weighting (each sentence is weighted by the teacher's mean confidence so clearer cases matter a bit more). Training is done in 15 epochs, learning rate 2×10^{-5}, weight decay 0.01, batch sizes 16 (train), 32 (eval), logging every 50 steps, and evaluation at every end of epoch, with best model loaded by default. This produces teacher-like behavior at a tiny fraction of inference time, which is essential when we move to sentence-level processing.

Apart from aspect detection, each sentence is passed through a sentence-level sentiment model, cardiffnlp/twitter-xlm-roberta-base-sentiment. It outputs neg/neu/pos, which we map to $-1/0/+1$. This gives us polarity to every aspect mention: if a sentence is about cleanliness, e.g., we also know the author's positive or negative opinion about it. Afterwards, at review level, we average sentiment only over the sentences actually mentioning that aspect, so we don't propagate emotion from elsewhere in the review.

Since both student and teacher produce probabilities, we convert them to raw present/absent tags with per-aspect thresholds. We start with a sensible default (0.40) and then adapt each aspect's cutoff *unsupervised* from the score distribution

over all sentences (Otsu's method, with a simple two-Gaussian alternative too). On our data this produced last Otsu thresholds of 0.3875 for service quality, 0.4075 for affordability, 0.5125 for treatment quality, and 0.3375 for cleanliness, predictably stricter for treatment quality (which had a more definite positive mode) and looser for service quality (which was noisier). For student diagnostics and agreement plots we also draw a plain 0.50 decision rule on the student outputs.

We carry on pre-processing light but effective for noisy, code-mixed reviews: stripping HTML and zero-width characters, normalizing whitespace and case, and lightly compressing long spellings common in Banglish ("sooooo" $\rightarrow$ "so"). We sentence-split reviews with NLTK, falling back on a simple punctuation regex otherwise. The architecture is aimed at localizing aspect mentions precisely at the sentence level, since it is there that customers actually mention explicit grievances or praise.

Finally, we aggregate sentence-level indicators to reviews and hospitals in a clear way. A review "has" an aspect if *any* of its sentences goes beyond that aspect's threshold; the review's aspect sentiment is the *mean* over just those sentences. For every hospital, we compute an aspect raw score as

$$\text{raw}_{a,h} \;=\; \big(\text{mention fraction of aspect } a \text{ in hospital } h\big) \;\times\; \frac{\overline{s}_{a,h} + 1}{2},$$

then min–max normalize between hospitals to obtain comparable C, A, TQ, SQ scores in $[0, 1]$. Global opinion S is computed from review-level labels and also scaled to $[0, 1]$. For rankings to be stable we (i) keep only those hospitals with ≥ 5 reviews, and (ii) average scores with an added volume-based confidence component by a log transform and $\alpha = 0.85$ so that very high-volume hospitals are marginally more trusted without overwhelming the signal. The final score is a fixed-weighted average with transparent weights. S=0.25, C=0.20, A=0.15, TQ=0.25, SQ=0.15, reflecting a defensible priority order (clinical effectiveness and overall satisfaction matter most, cleanliness next, with affordability and service quality still important). We intentionally keep these weights fixed because we lack ground-truth "best hospital" labels, and fixed weights make the ranking auditable and reproducible; the *learning* in our system happens earlier, when the student model is trained to produce high-quality aspect probabilities from sentences.

Chapter 5

Result Analysis

5.1 Preliminary Result Analysis

The dataset comprising Google reviews of hospitals across the entire Bangladesh was used to train 5 machine learning models - Decision Tree Classifier, Support Vector Machine (SVM), Logistic Regression, Random Forest, and Gaussian Naive Bayes - and 2 deep learning models - BERT and CNN. The performance of these models was then evaluated by assessing their accuracy, F1 score, precision, recall, and confusion matrix for each of the 8 divisions of Bangladesh - Rangpur, Rajshahi, Dhaka, Mymensingh, Sylhet, Barisal, Khulna, and Chittagong. Furthermore, each division was also made to produce three types of word clouds - Positive word cloud, Negative word cloud, and Neutral word cloud - in order to showcase the most common words in the dataset used for a division that result in a positive, negative, and neutral sentiment, respectively.

5.1.1 Evaluation Metrics

1. **Accuracy:** Accuracy measures the overall effectiveness of a classifier, i.e. the fraction of all correct predictions (true positives and true negatives) out of the total number of predictions:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

where

- TP: True Positives,
- TN: True Negatives,
- FP: False Positives,
- FN: False Negatives.

2. **Precision:** Precision quantifies the exactness of the classifier by measuring how many of the instances predicted positive are actually positive:

$$\text{Precision} = \frac{TP}{TP + FP}$$

3. **Recall (Sensitivity):** Recall (or sensitivity) measures the completeness of the classifier by computing the fraction of actual positives that were correctly identified:
$$\text{Recall} \;=\; \frac{TP}{TP + FN}$$

4. **$\mathbf{F_1}$ Score:** The F_1 score is the harmonic mean of precision and recall, providing a single balanced metric:
$$F_1 \;=\; 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

5. **Word Cloud.** A word cloud is a visual summary of text data where each word's size reflects its frequency or importance, allowing quick identification of prominent terms.

6. **Confusion Matrix.** The confusion matrix is a tabular summary of classification outcomes, listing counts of true positives, true negatives, false positives, and false negatives to give a complete picture of model performance.

5.1.2 Rangpur

In the case of the Rangpur division, the Random Forest model scores the highest accuracy of 81.9% and the highest recall of 81.9% with BERT scoring the highest precision of 82.2% and the highest F1 score of 80.1%. A comparison among the performance of all the models and the confusion matrices is shown in figure 5.1 and in figure 5.3 respectively; along with the positive, negative, and neutral word clouds for the Rajshahi division in figure 5.2.

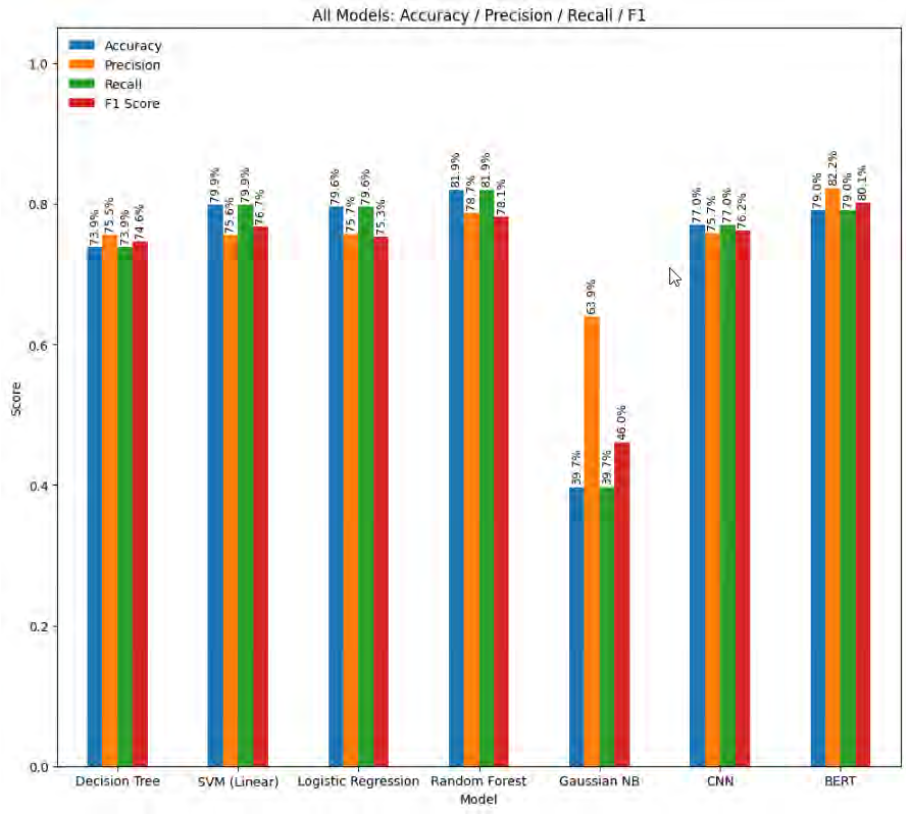

Figure 5.1: Comparison of all models' performance for the Rangpur division.

Figure 5.2: Negative, Neutral, and Positive word cloud for the Rangpur division.

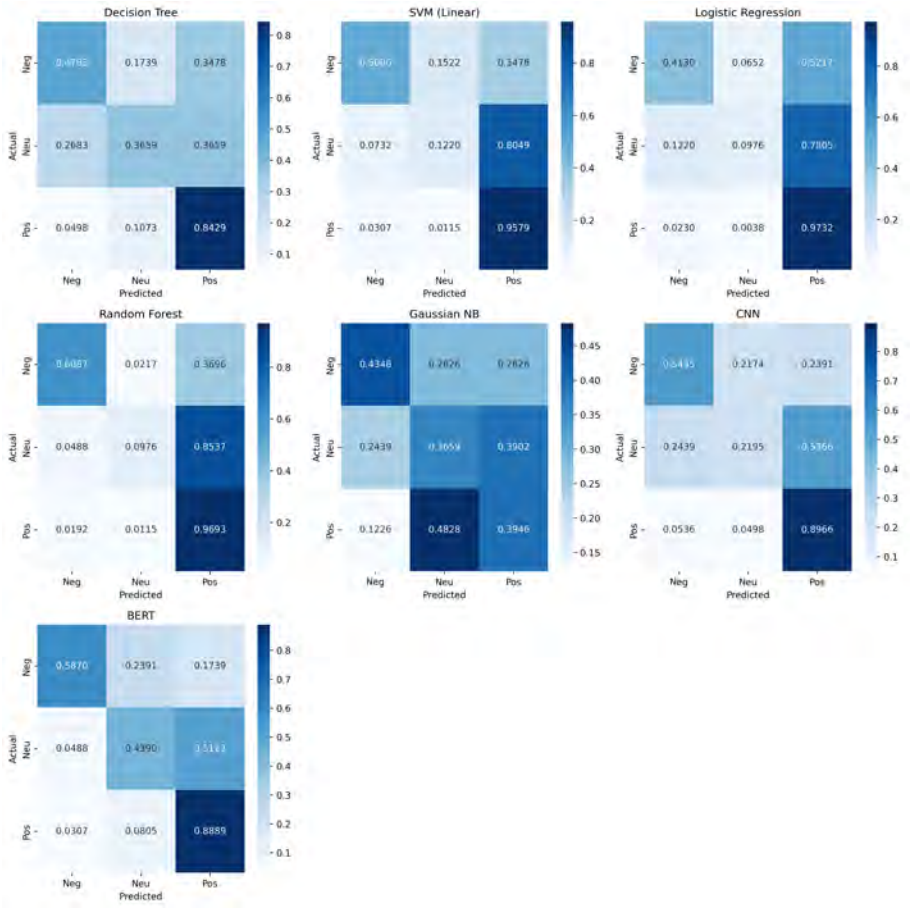

Figure 5.3: Confusion matrices of all models for the Rangpur division.

5.1.3 Rajshahi

For the Rajshahi division, BERT performs the best in terms of all four metrics with an accuracy of 93.1%, precision of 92.8%, recall of 93.1%, and F1 score of 92.9%. A comparison among the performance of all the models in figure 5.4 and the confusion matrices is shown in figure 5.6 along with the positive, negative, and neutral word clouds in figure 5.5 for the Rajshahi division.

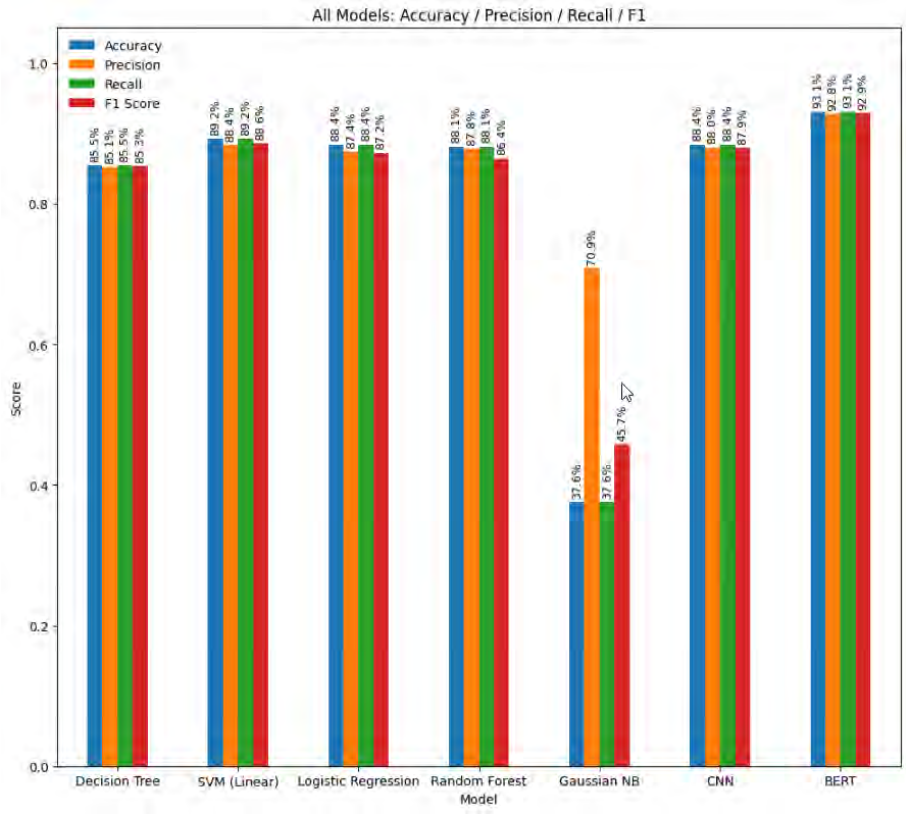

Figure 5.4: Comparison of all models' performance for the Rajshahi division.

Figure 5.5: Negative, Neutral, and Positive word cloud for the Rajshahi division.

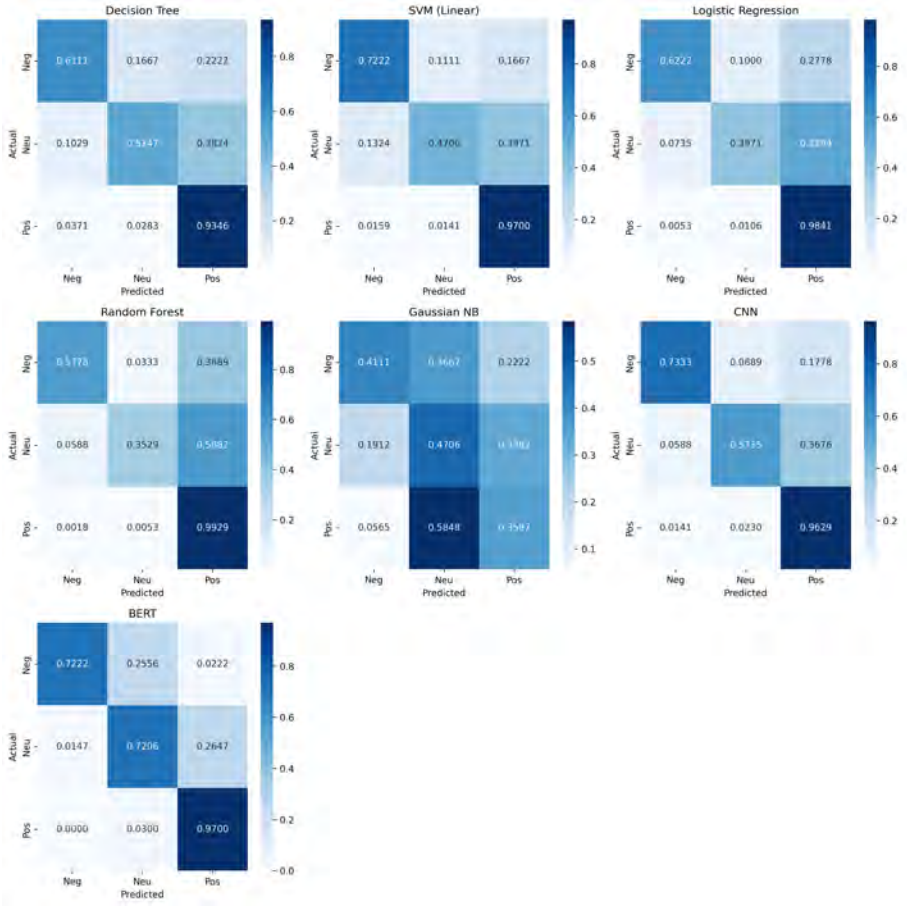

Figure 5.6: Confusion matrices of all models for the Rajshahi division.

5.1.4 Dhaka

For the Dhaka division, BERT, yet again, performs the best with an accuracy of 86.6%, precision of 86.5%, recall of 86.6%, and F1 score of 86.4%. A comparison among the performance of all the models in figure 5.7, and the confusion matrices is shown in figure 5.9, along with the positive, negative, and neutral word clouds in figure 5.8 for the Dhaka division.

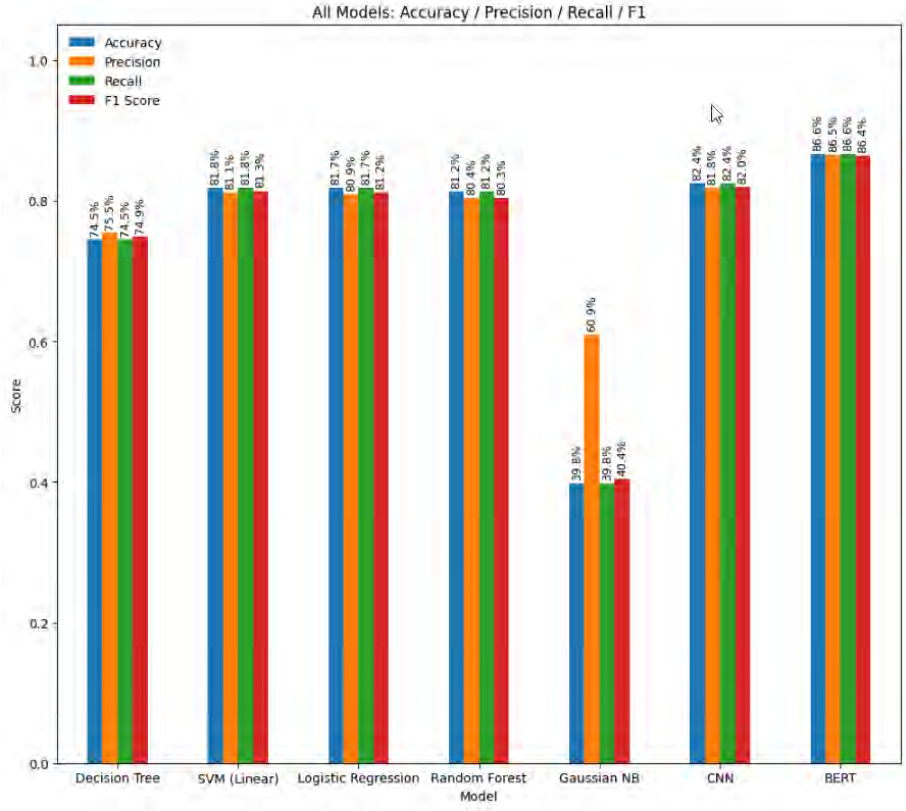

Figure 5.7: Comparison of all models' performance for the Dhaka division.

Figure 5.8: Negative, Neutral, and Positive word cloud for the Dhaka division.

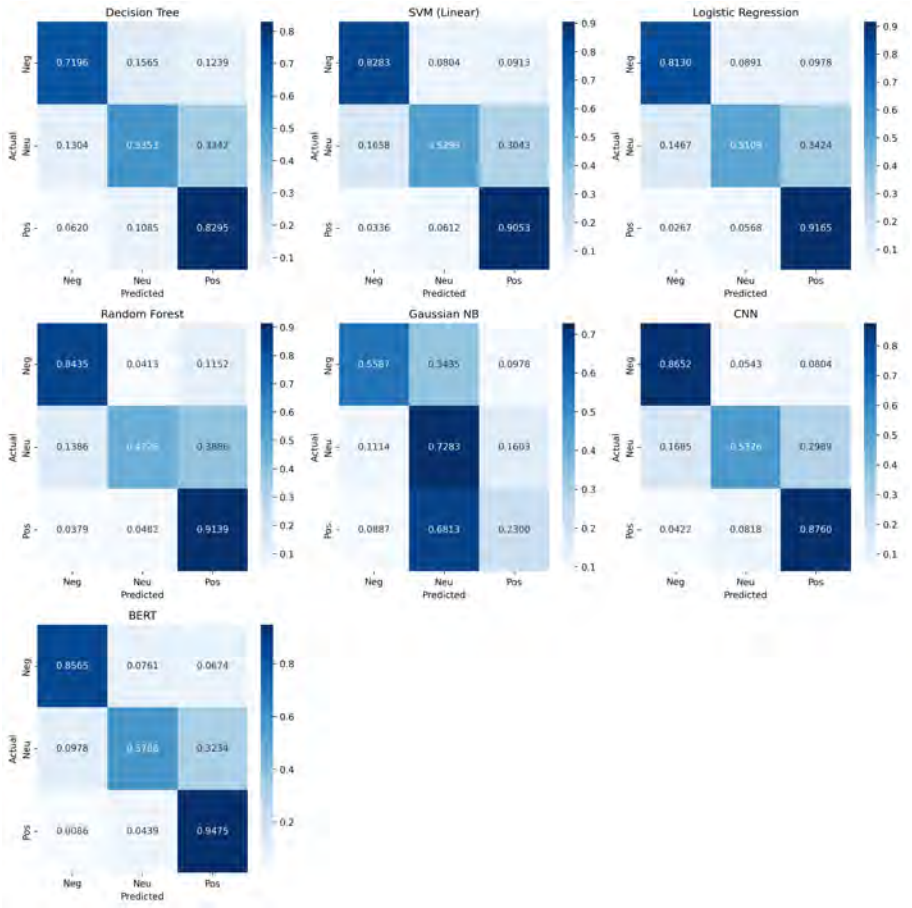

Figure 5.9: Confusion matrices of all models for the Dhaka division.

5.1.5 Mymensingh

In the case of the Mymensingh division, the SVM model achieves the highest precision of 87.6%. However, BERT computes the highest accuracy, recall, and F1 score of 88.9%, 88.9%, and 87.3%, respectively. A comparison among the performance of all the models in figure 5.10 and the confusion matrices is shown in figure 5.12 along with the positive, negative, and neutral word clouds in figure 5.11 for the Mymensingh division.

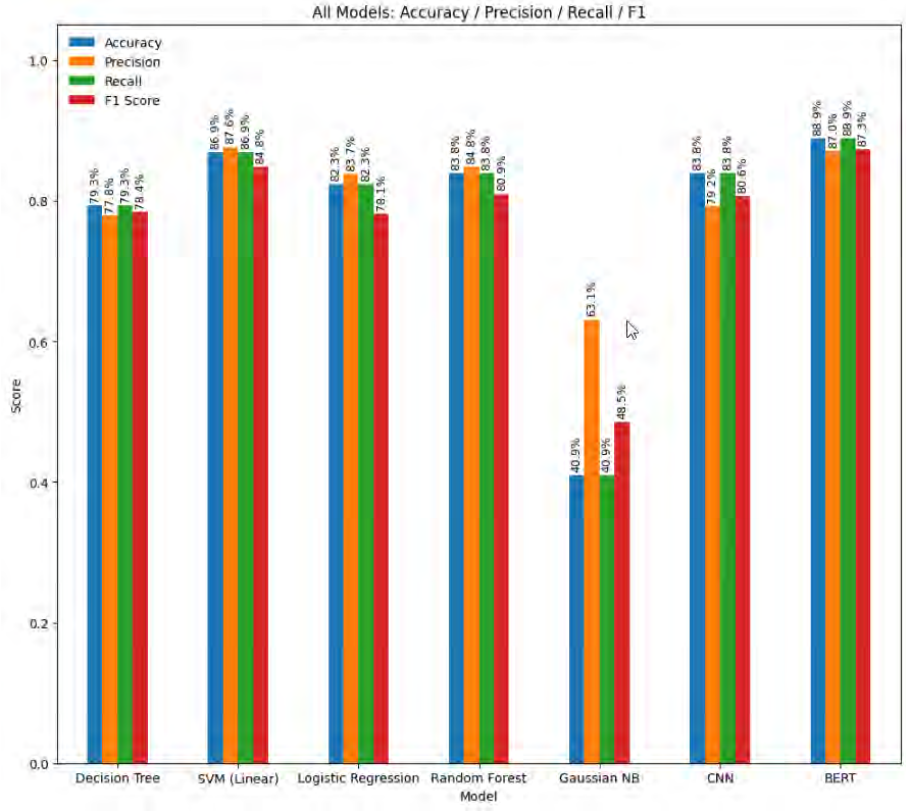

Figure 5.10: Comparison of all models' performance for the Mymensingh division.

Figure 5.11: Negative, Neutral, and Positive word cloud for the Mymensingh division.

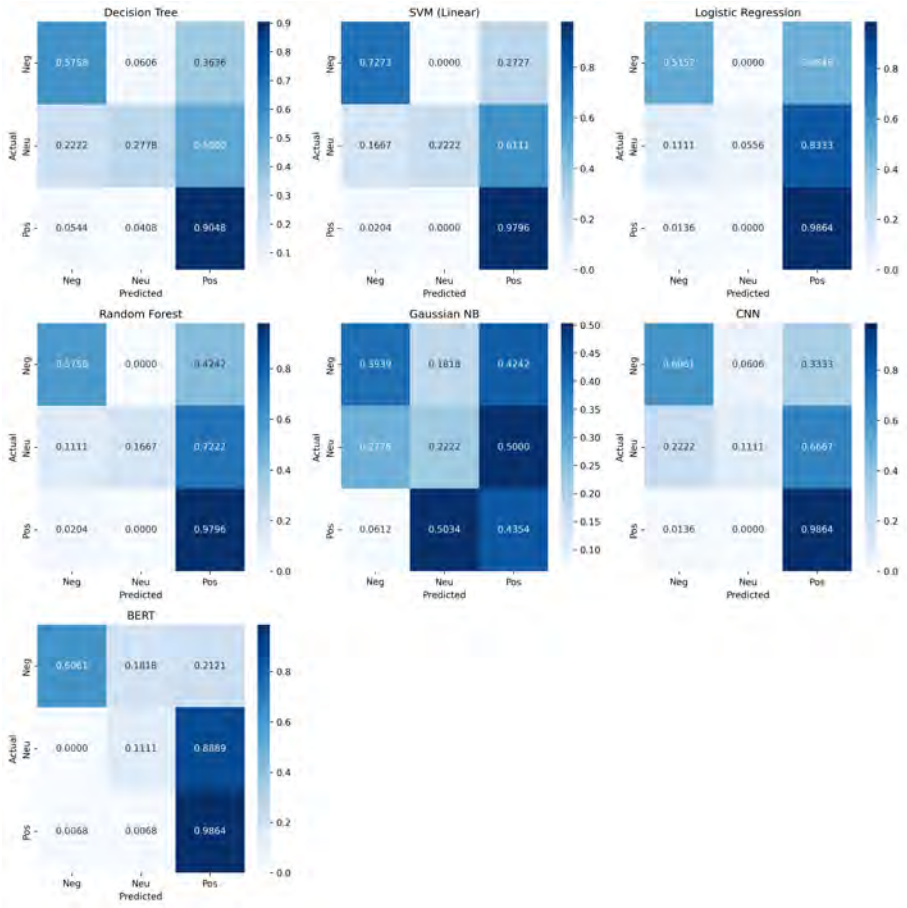

Figure 5.12: Confusion matrices of all models for the Mymensingh division.

5.1.6 Sylhet

For the Sylhet division, BERT produces the highest scores for all the metrics with an accuracy of 94.1%, precision of 94.7%, recall of 94.1%, and F1 score of 94.4%. A comparison among the performance of all the models in figure 5.13 and the confusion matrices is shown in figure 5.15 along with the positive, negative, and neutral word clouds in figure 5.14 for the Sylhet division.

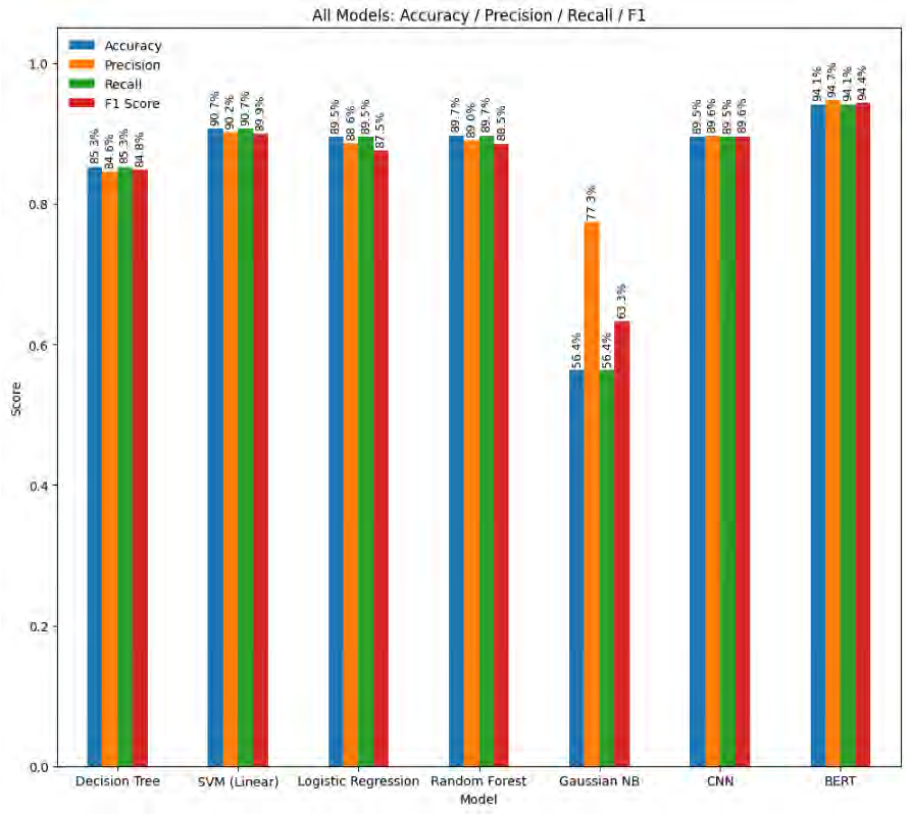

Figure 5.13: Comparison of all models' performance for the Sylhet division.

Figure 5.14: Negative, Neutral, and Positive word cloud for the Sylhet division.

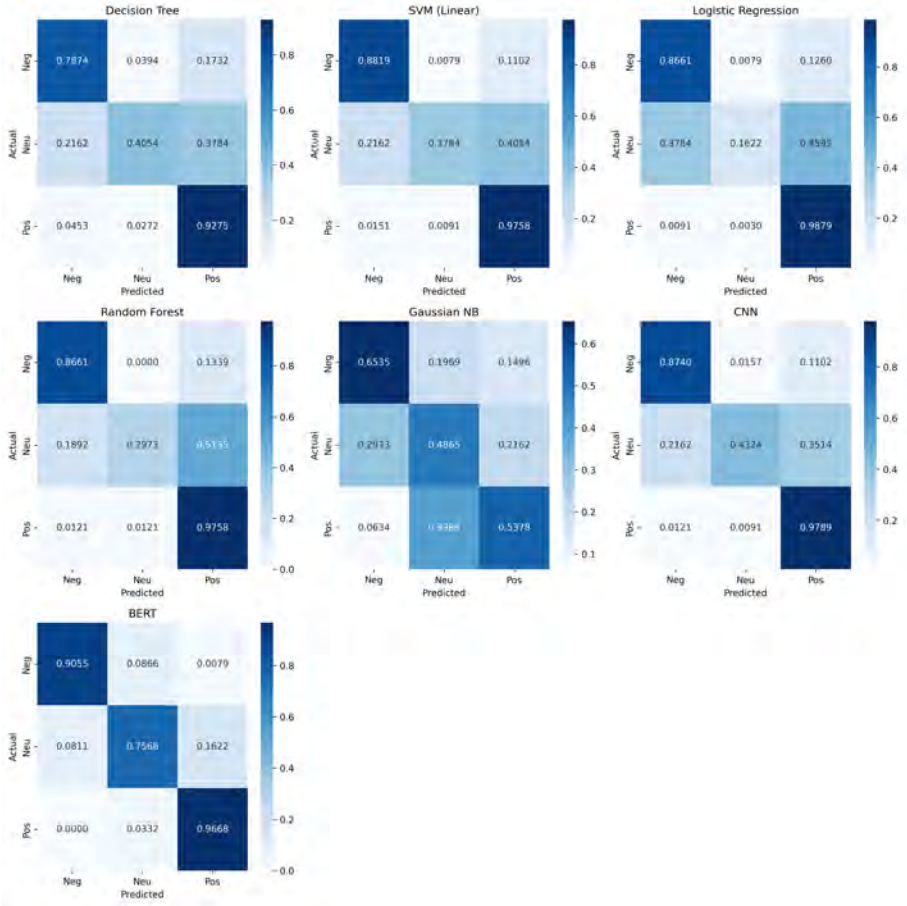

Figure 5.15: Confusion matrices of all models for the Sylhet division.

5.1.7 Barisal

In the case of the Barisal division, BERT computes the highest accuracy, precision, recall, and F1 score of 82.5%, 81.9%, 82.5%, and 81.8%, respectively. A comparison in figure 5.16 among the performance of all the models and the confusion matrices in figure 5.18 is shown along with the positive, negative, and neutral word clouds in figure 5.17 for the Barisal division.

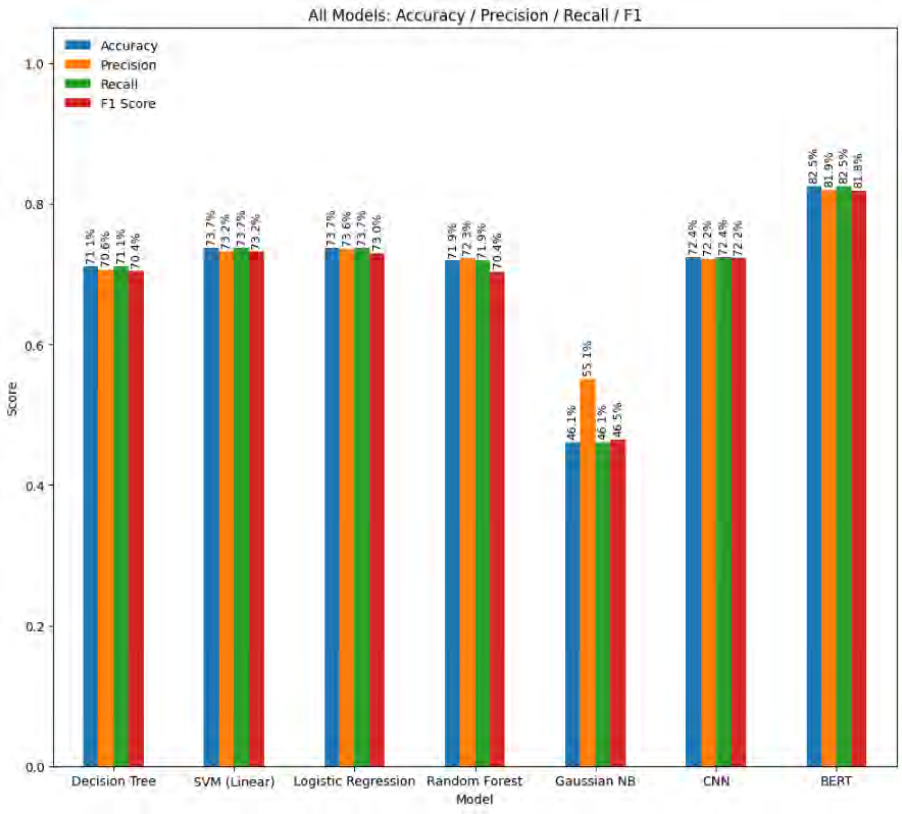

Figure 5.16: Comparison of all models' performance for the Barisal division.

Figure 5.17: Negative, Neutral, and Positive word cloud for the Barisal division.

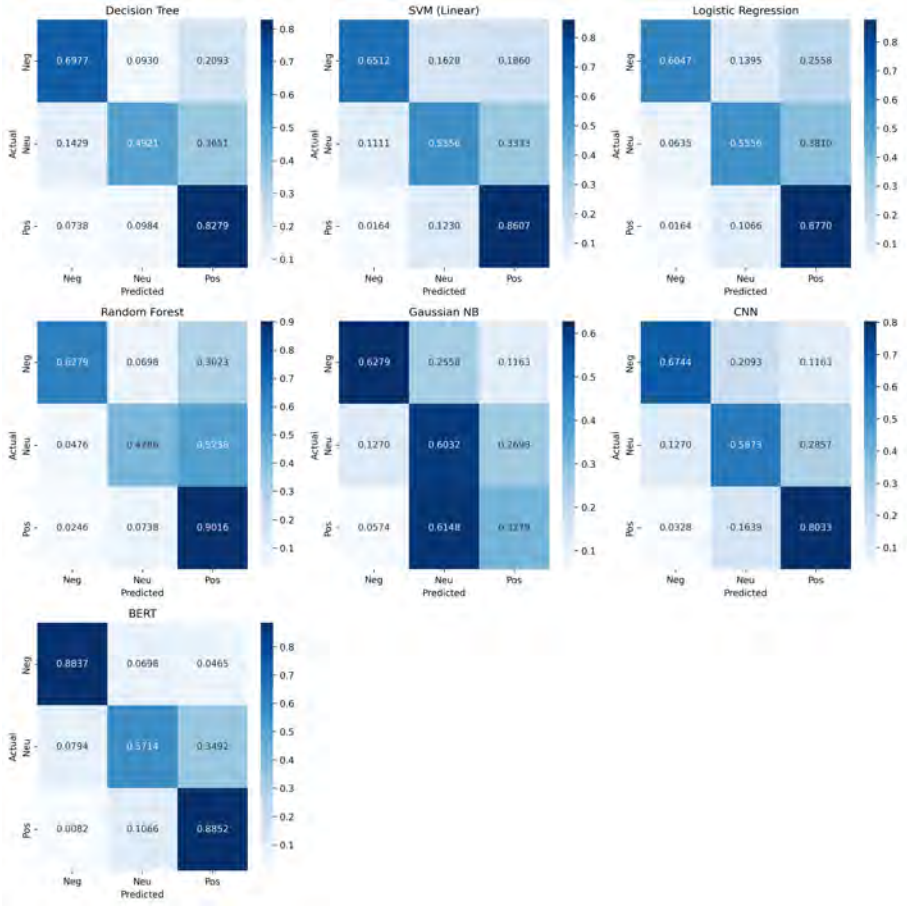

Figure 5.18: Confusion matrices of all models for the Barisal division.

5.1.8 Khulna

For the Khulna division, BERT scores an accuracy of 90.8%, precision of 91.0%, recall of 90.8%, and F1 score of 90.8% with the SVM model following closely with scores that do not differ much from those of BERT. A comparison among the performance of all the models in figure 5.19 and the confusion matrices is shown in figure 5.21 along with the positive, negativ,e and neutral word clouds in figure 5.20 for the Khulna division.

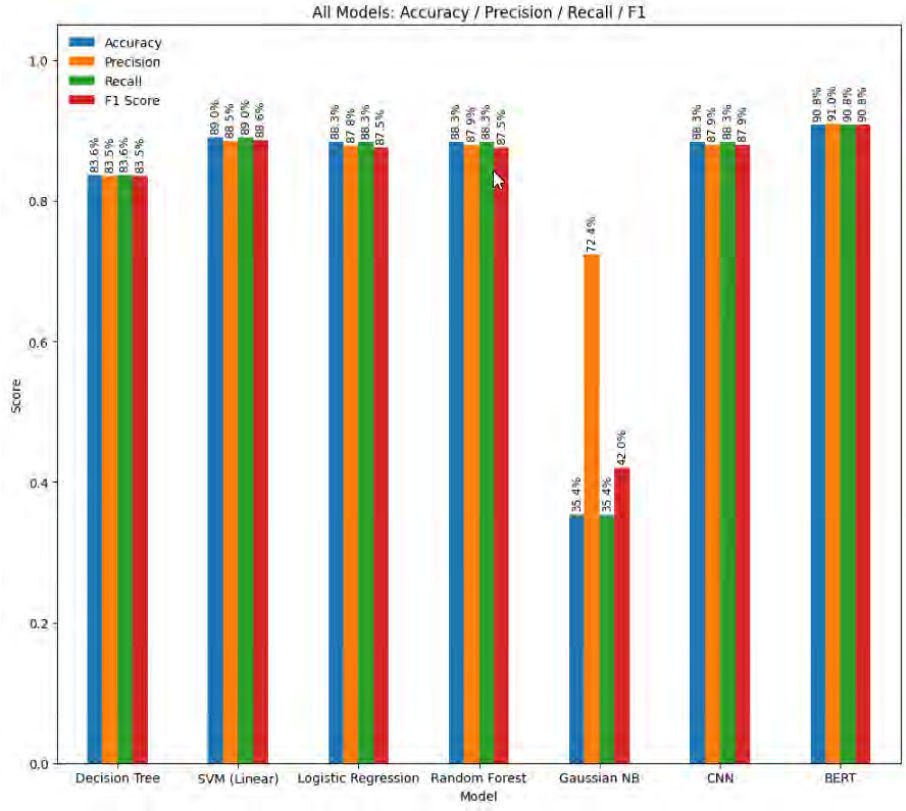

Figure 5.19: Comparison of all models' performance for the Khulna division.

Figure 5.20: Negative, Neutral, and Positive word cloud for the Khulna division.

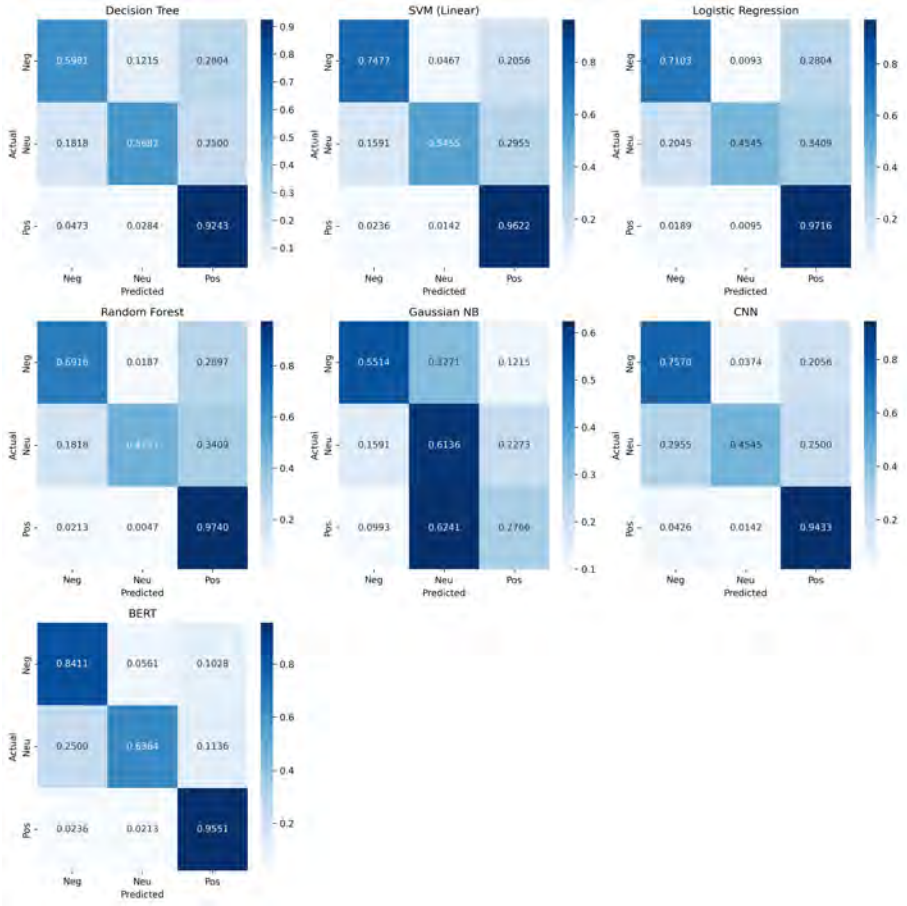

Figure 5.21: Confusion matrices of all models for the Khulna division.

5.1.9 Chittagong

Lastly, for the Chittagong division, BERT computes the highest accuracy, precision, recal,l and F1 score of 84.1%, 82.4%, 84.1%, and 83.0%, respectively. A comparison among the performance of all the models in figure 5.22 and the confusion matrices is shown in figure 5.24 along with the positive, negative, and neutral word clouds in figure 5.23 for the Chittagong division.

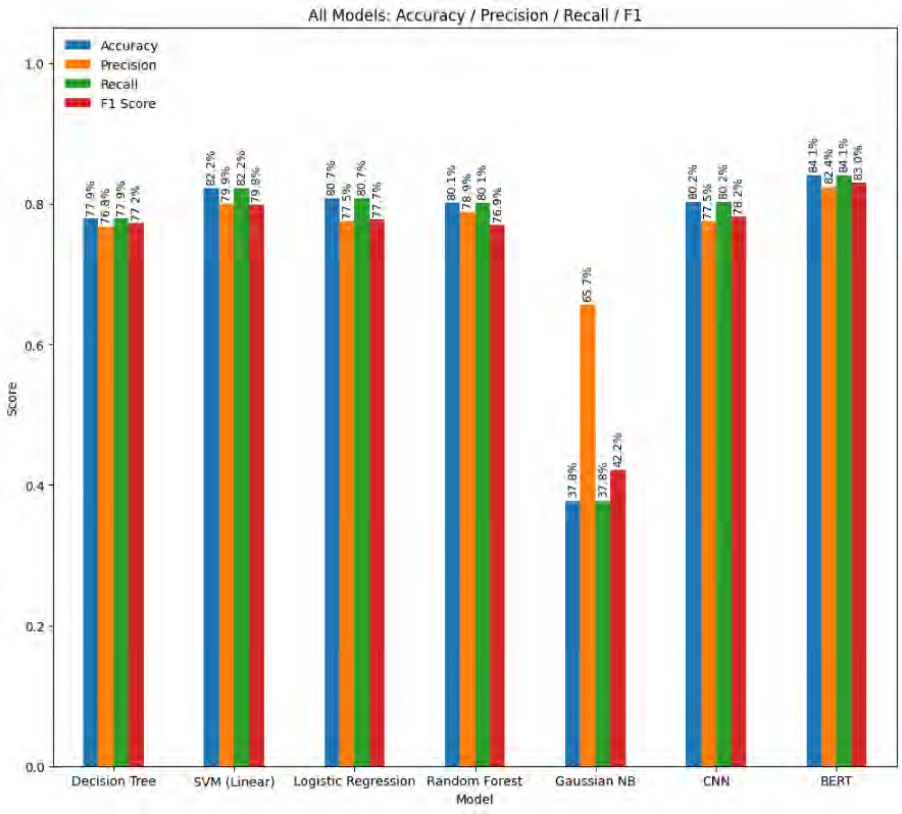

Figure 5.22: Comparison of all models' performance for the Chittagong division.

Figure 5.23: Negative, Neutral, and Positive word cloud for the Chittagong division.

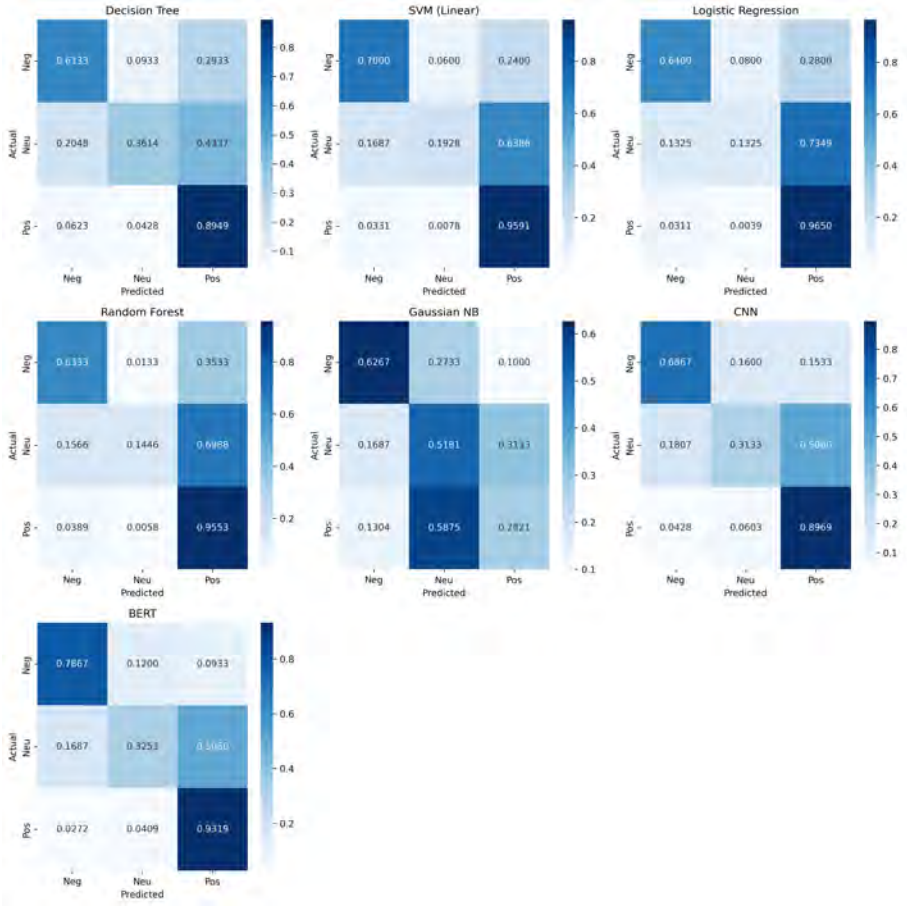

Figure 5.24: Confusion matrices of all models for the Chittagong division.

5.2 Statistical Analysis Graphs - Dhaka Hospital Reviews

5.2.1 Student Model Evaluation

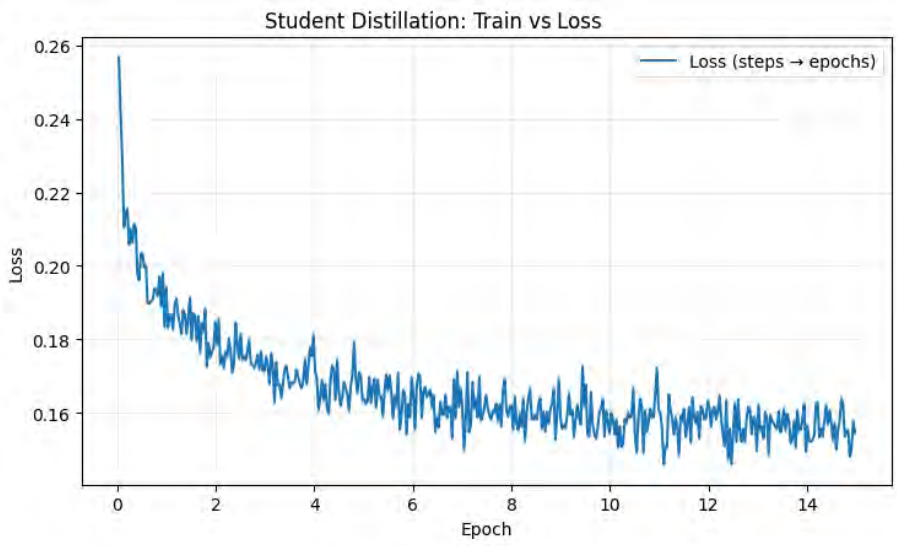

Figure 5.25: Student model training loss over fifteen epochs during distillation.

From figure 5.25, Training loss decreases steadily from ≈ 0.256 at the start to ≈ 0.17 by the end of epoch 15, with only small minibatch-level wiggles. The curve shows the classic "fast early drop, slower later improvement," and there's no upward drift, evidence of stable optimization and no obvious overfitting.

5.2.2 Treatment Quality

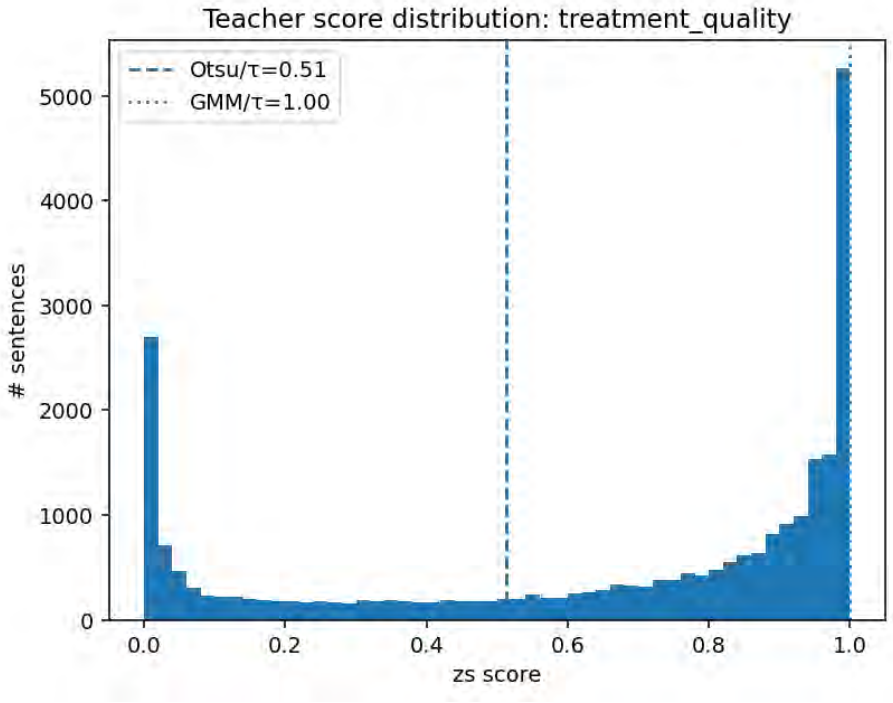

Figure 5.26: Teacher score distribution for treatment quality.

From figure 5.26, we can see a dense right-side cluster and threshold of 0.51 indicate high confidence in identifying treatment-related content, consistent with the frequent clinical discussions in hospital reviews. The histogram is heavily right-skewed, with many high-scoring sentences expressing strong medical care content. The Otsu threshold ($\tau = 0.51$) is stricter here, reflecting that only confident, explicit mentions of treatment quality are classified as positive.

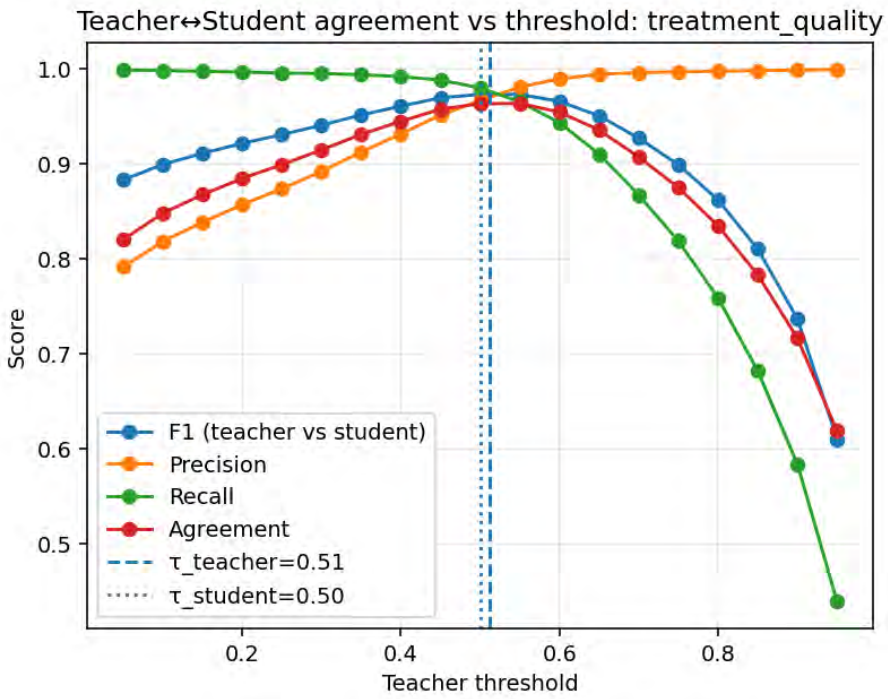

Figure 5.27: Teacher–student agreement vs threshold for treatment quality.

In figure 5.27, the student mirrors the teacher almost perfectly on clinical content: precision ∼0.98, recall ∼0.98, F1 ∼0.97, and overall agreement ∼0.96 peak close to the Otsu threshold of 0.51.

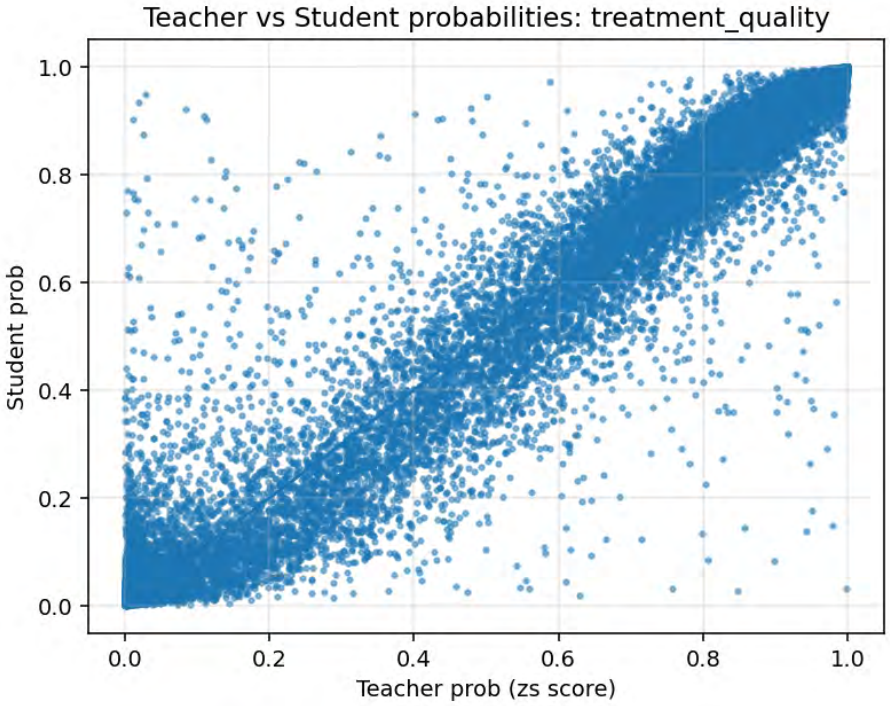

Figure 5.28: Teacher vs Student probabilities for treatment quality. (Pearson $r = 0.977$)

From figure 5.28, we can conclude that this aspect displays the highest consistency,

with almost all points following the diagonal, indicating that medical-related statements are highly learnable and unambiguous. Treatment Quality (r = 0.977) demonstrates high consistency, suggesting that the student model learned subtle distinctions between affordability cues effectively.

5.2.3 Cleanliness

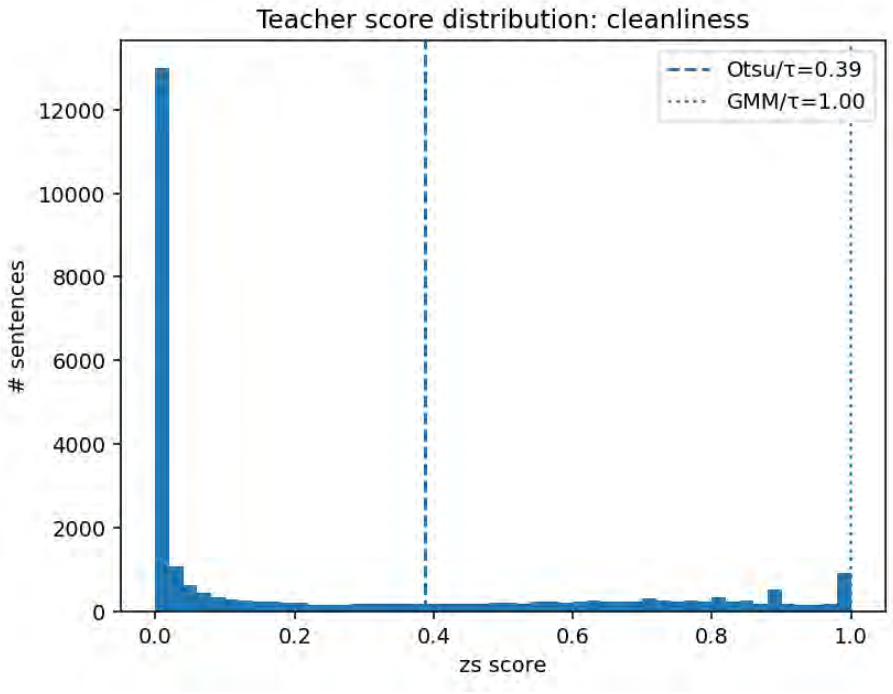

Figure 5.29: Teacher score distribution for cleanliness.

In figure 5.29, It can be seen that most sentences score near zero, confirming that explicit cleanliness mentions are sparse. The Otsu threshold at 0.39 efficiently separates generic sentences from those focused on hygiene. The histogram shows a strong left-side peak near 0, indicating that most sentences do not mention cleanliness. The Otsu threshold ($\tau = 0.39$) falls in the valley between near-zero and high-score regions, correctly separating neutral from cleanliness-related sentences.

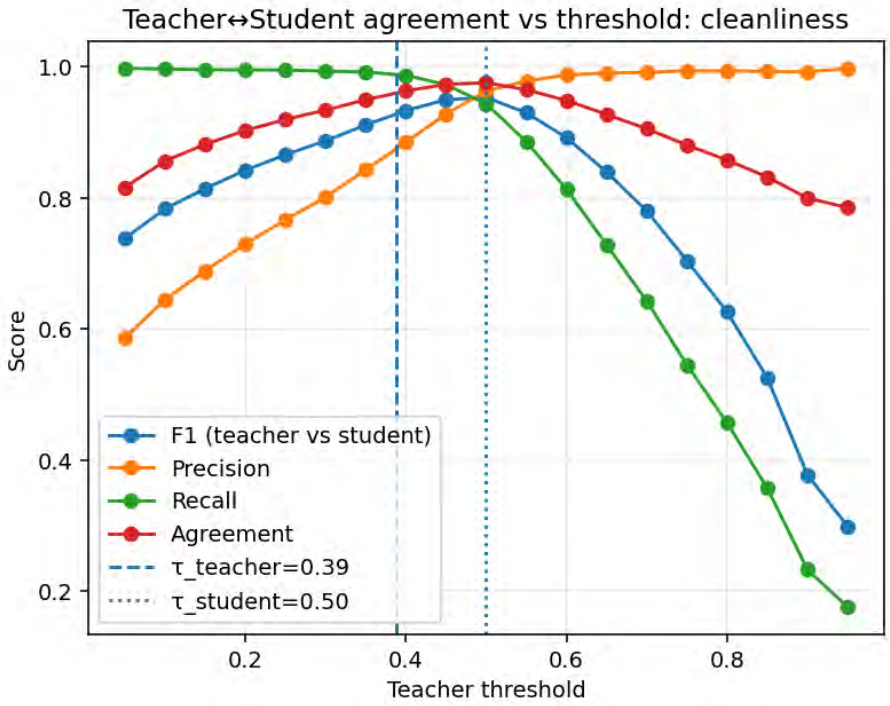

Figure 5.30: Teacher–student agreement vs threshold for cleanliness.

From figure 5.30, we observe precision ∼0.88, recall ∼0.98, F1 ∼0.92, and agreement ∼0.94.

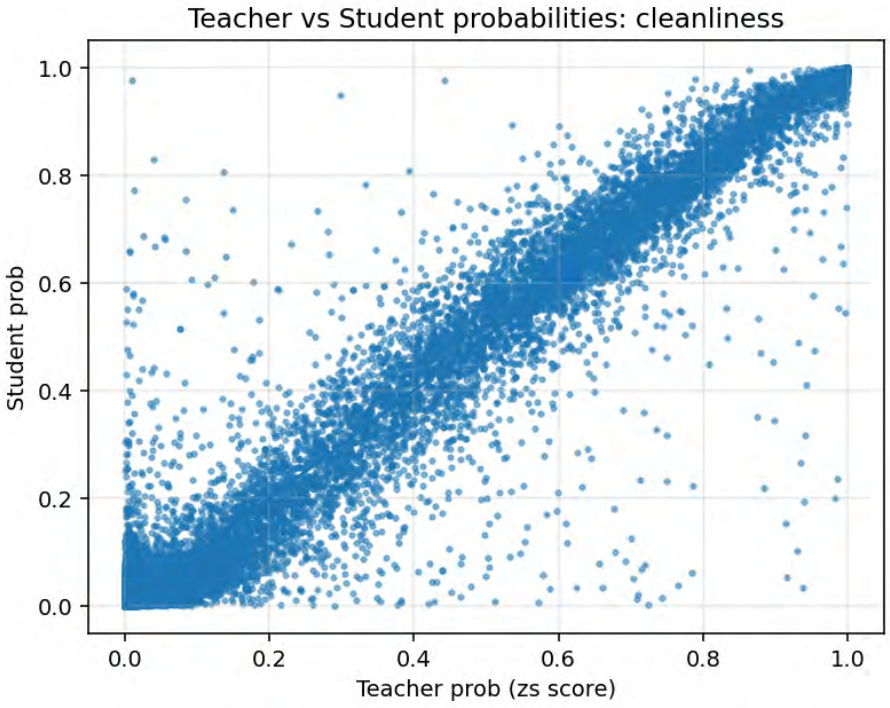

Figure 5.31: Teacher vs Student probabilities for cleanliness. (Pearson $r = 0.984$)

In figure 5.31, the scatter points cluster tightly along the diagonal, confirming that the student model successfully mirrors the teacher's predictions for cleanliness-related sentences. Cleanliness (r = 0.984) shows the highest alignment, with almost

all points hugging the diagonal, indicating near-perfect transfer of the teacher's probability structure.

5.2.4 Affordability

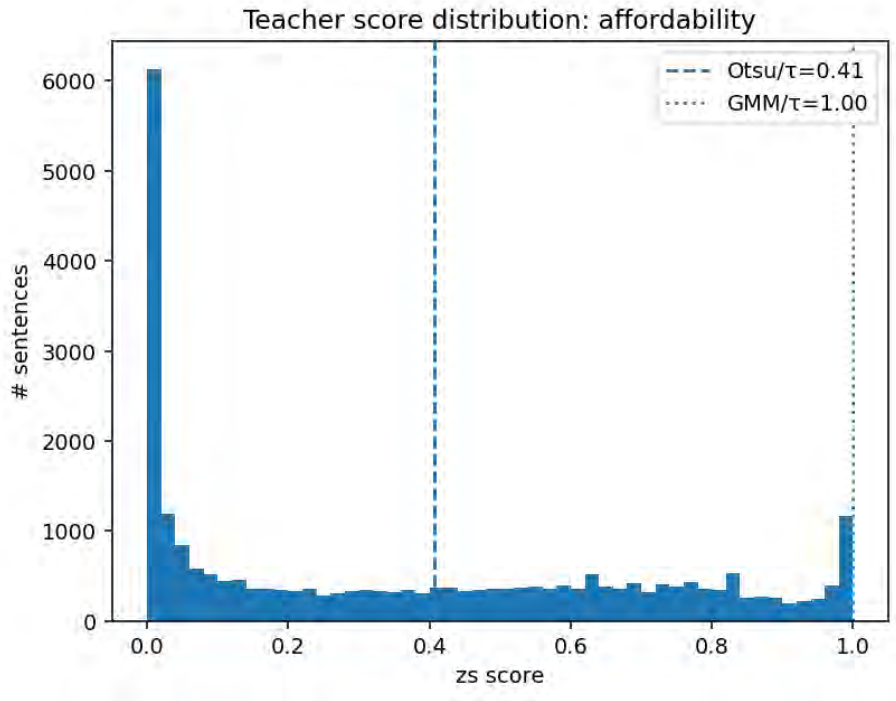

Figure 5.32: Teacher score distribution for affordability.

The model clearly distinguishes cost-related language which can be seen in figure 5.32, with a bimodal distribution and Otsu threshold at 0.41 marking the valley between unrelated and strongly price-focused sentences. Two clear modes appear near 0 and 1, reflecting sentences that either do not mention or strongly emphasize cost. The Otsu cutoff ($\tau = 0.41$) lies at the intersection between these groups, providing a balanced separation between affordability-related and unrelated text.

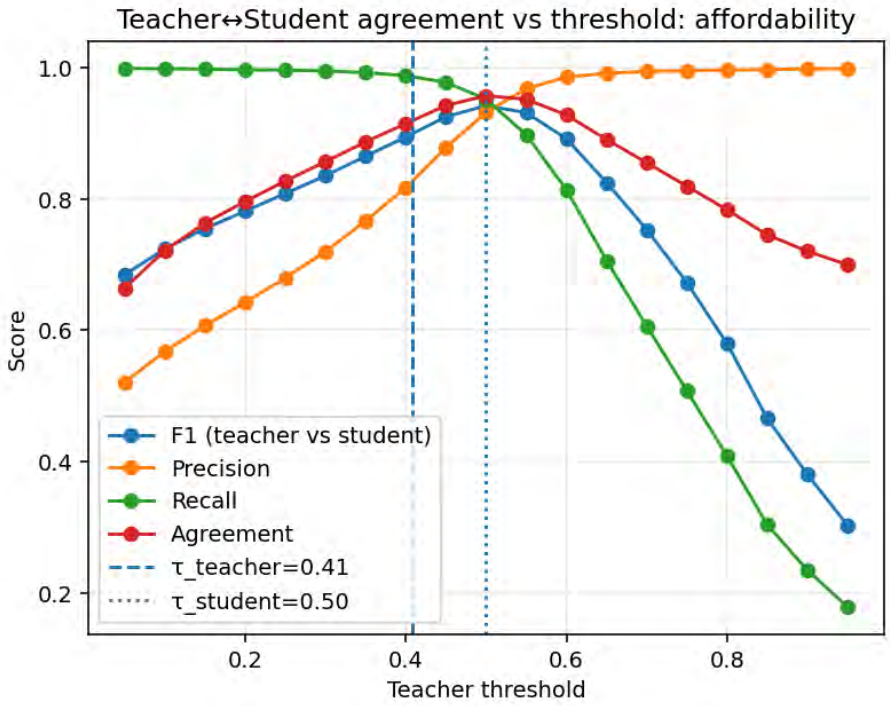

Figure 5.33: Teacher–student agreement vs threshold for affordability.

From figure 5.33, we can see that at the teacher's ($\tau = 0.41$) for affordability, the student (fixed ($\tau = 0.50$)) shows very high recall ($\sim$.99) but moderate precision ($\sim$0.82), giving F1 $\sim$0.90 and overall agreement $\sim$0.92. This indicates the student catches almost all teacher-positives but also includes more false positives at this operating point. As you move the teacher threshold right toward $\sim$0.5, precision rises while recall falls, with F1 peaking near the ($\tau = 0.50$),suggesting the student is calibrated a bit "stricter" than the teacher.

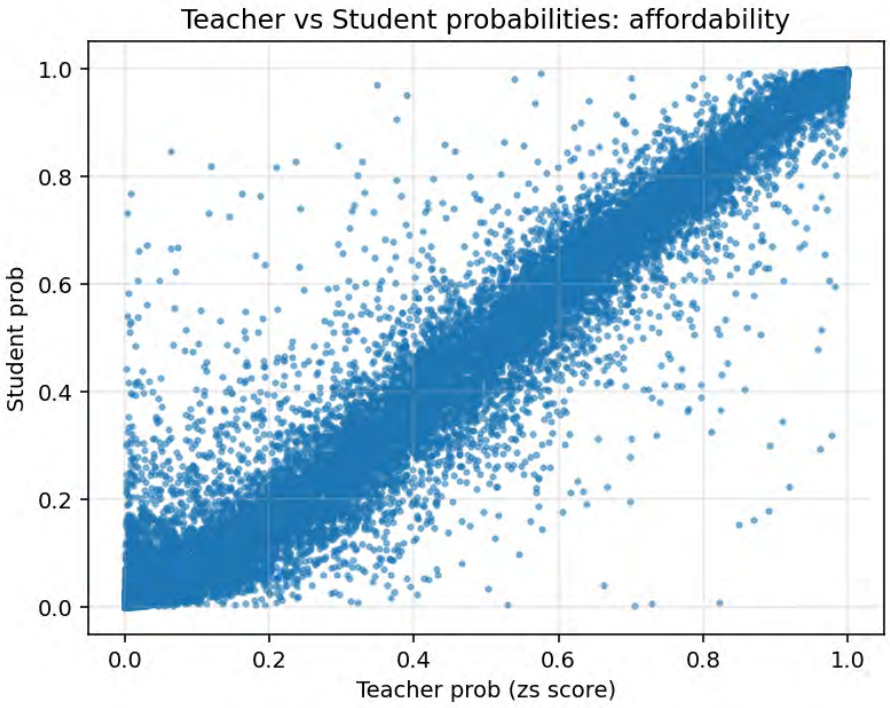

Figure 5.34: Teacher vs Student probabilities for affordability. (Pearson $r = 0.978$)

From figure 5.34, Affordability (r = 0.978) demonstrates high consistency, suggesting that the student model learned subtle distinctions between affordability cues effectively.

5.2.5 Service Quality

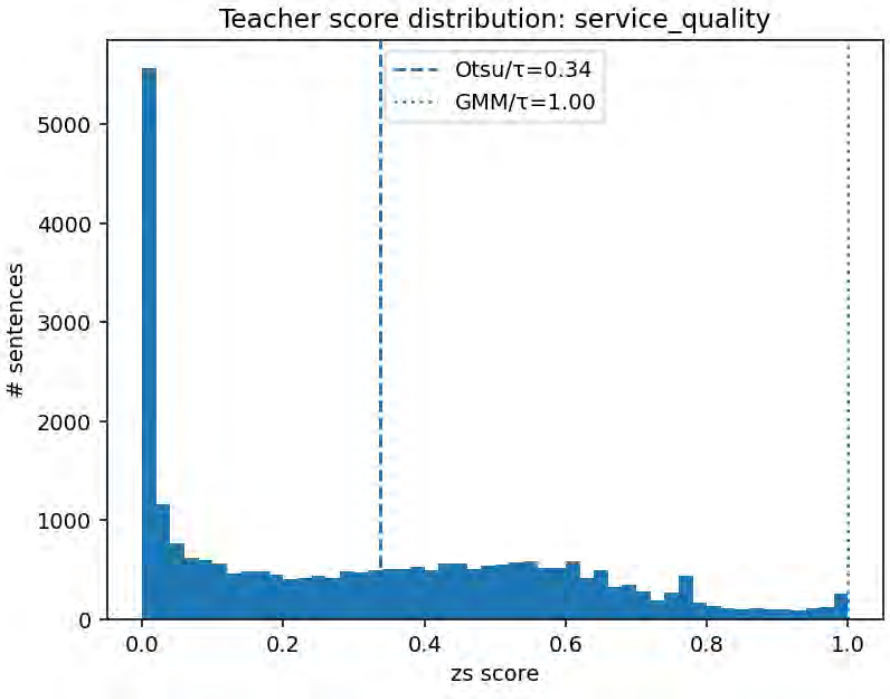

Figure 5.35: Teacher score distribution for service quality.

For figure 5.35; the broader, more even spread and lower cutoff at 0.34 reflect the subtle and context-dependent nature of service-related remarks such as staff behavior and waiting time. Compared to other aspects, this distribution is more diffuse. The Otsu threshold ($\tau = 0.34$) is relatively lower, capturing a broader range of moderate service-related mentions and compensating for the noisier linguistic variability in staff and behavior descriptions.

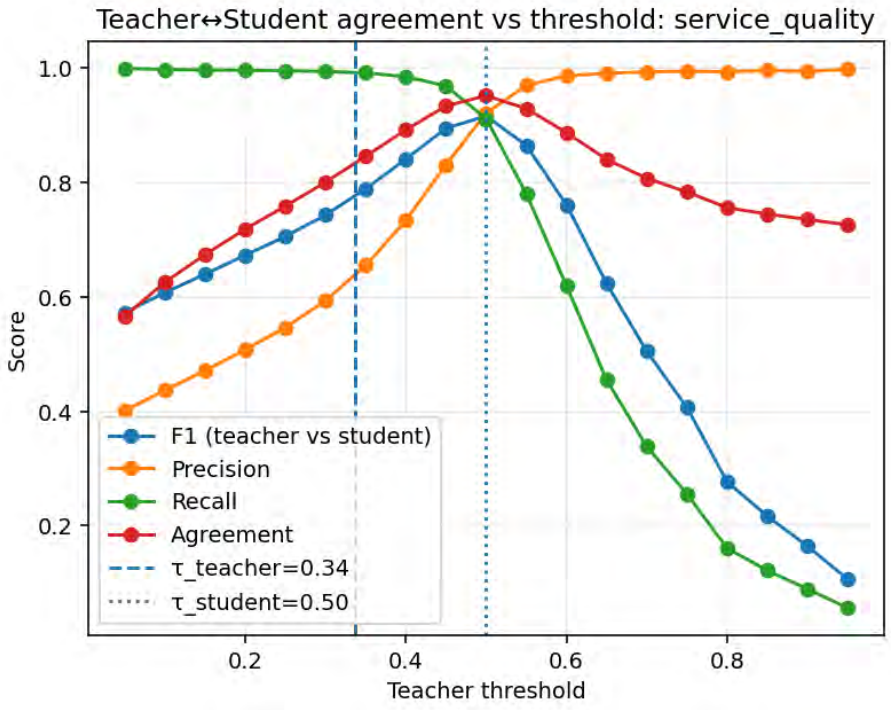

Figure 5.36: Teacher–student agreement vs threshold for service quality.

In figure 5.36, the student (fixed τ=0.50) is highly recall-heavy at the teacher's lower τ, catching nearly all teacher positives but with many extra positives, so precision lags(precision $\sim$ .64) and F1 is $\sim$0.78s. As the teacher τ moves right toward $\sim$0.5, precision rises and F1 peaks, suggesting the student runs "stricter" than the teacher.

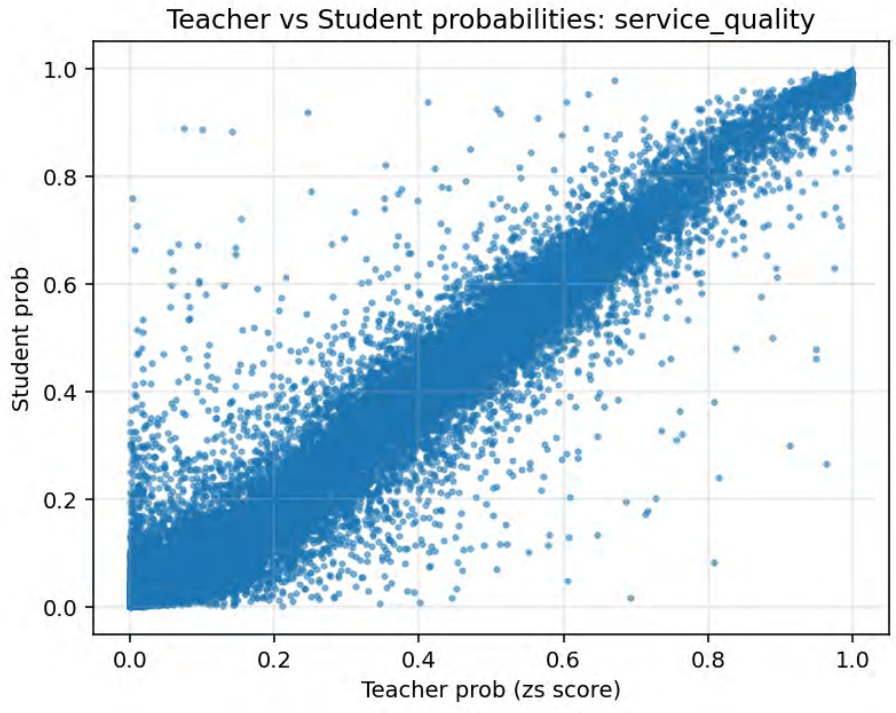

Figure 5.37: Teacher vs Student probabilities for service quality. (Pearson $r = 0.970$)

From figure 5.37, Service Quality (r = 0.970), while slightly lower, still reflects a very strong linear correlation. The slightly wider spread aligns with this aspect's subjective linguistic variability in user feedback.

5.3 Performance Evaluation

The proposed system demonstrates excellent accuracy, with the BERT model achieving up to 94.1% accuracy in the Sylhet division and consistently strong performance above 86% across most divisions. The key efficiency breakthrough lies in the zero-shot learning approach for aspect detection. Instead of requiring thousands of manually labeled examples, the system intelligently reuses a pretrained multilingual model (XLM-RoBERTa) to automatically identify four critical aspects in patient reviews: cleanliness, affordability, treatment quality, and service quality. The model evaluates each review by asking "Is this review about [aspect]?" and returns confidence scores between 0 and 1, marking aspects present when confidence exceeds 0.40. This eliminates months of data preparation work while maintaining high detection accuracy. The system efficiently handles both English and Banglish text through subword tokenization, breaking words into meaningful components and using contextual clues to understand meaning. The multi-label classification approach lets simultaneous detections of multiple aspects of a single review, reflecting how real patients naturally discuss their experience.
The consistency and reliability of the system are demonstrated by various meticulously crafted mechanisms. The zero-shot aspect detection allows for consistent classification by an unsupervised per aspect threshold, which ensures accurate aspect determination in all reviews for different aspects. The multi-label approach individually accounts for each sectence for different concerns, allowing the model to

detect multiple aspect in single sentence without forcing artificial decisions among aspects. The confidence scoring model increases reliability by incorporating review count through logarithmic scaling, with appropriate weighting of those hospitals with very large numbers of feedback without undermining statistical validity. The model is particularly reliable in divisions such as Sylhet (94.1%), Rajshahi (93.0%), and Khulna (90.8%), where BERT's bidirectional contextual comprehension works well to pick up nuanced sentiment. Aspect-based aggregation correctly computes scores based on both mention frequency and sentiment polarity, providing stable measures that reflect real patient experience. Normalized scoring across hospitals allows for fair comparisons, preventing the possibility of regional variation in review patterns from distorting ranking.

The design strongly meets user needs by providing rich, aspect-based hospital ratings much more effective than simple star systems. Patients provide detailed feedback on four specific dimensions that matter the most: affordability and billing, treatment quality and medical care, cleanliness and hygiene, and quality of service, staff behavior, and waiting time. Its intelligent scoring mechanism calculates every component by combining mention frequency with strength of sentiment, and then weighted aggregation in which treatment quality and overall sentiment are given 25% weightage each, cleanliness 20%, and affordability and service quality 15% each. This produces sensible overall and aspect-wise top-three hospital rankings to enable patients to make informed decisions based on their personal priorities. The review-volume-based confidence multiplier ensures reliable rankings by appropriately weighting hospitals with extensive patient reviews. Users are given transparency through brief metrics of mention frequencies and average sentiment scores, with the exact knowledge of why hospitals are ranked as they are.

The system successfully addresses Bangladesh's $4 billion healthcare tourism problem by helping patients identify quality local facilities across multiple performance dimensions. The zero-shot approach allows flexible threshold tuning and customizable aspect weights, making the system adaptable to evolving user needs. Most importantly, the aspect-based framework captures nuanced patient experiences that generic star ratings completely miss, providing actionable insights for patients, healthcare administrators, and policymakers alike.

5.4 Analysis of Design Solutions

Our goal was to mine aspects from hospital reviews without any manual labels and still keep the system fast enough for Colab-scale hardware. We first explored a lexicon approach, but it was too brittle: reviewers use slang and Banglish/code-mixing, so fixed wordlists missed many true mentions. We then tried review-level zero-shot classification (one decision for the whole review). It was faster but too coarse, long reviews often mix praise and complaints across different aspects, so a single label per review blurred the signal. We therefore moved to sentence-level zero-shot detection: split each review into sentences and ask a multilingual NLI model whether each sentence is about our four aspects (cleanliness, affordability, treatment quality, service quality). This solved the "topic dilution" problem but increased compute. To make it practical, we used teacher→student distillation: the heavy zero-shot model (teacher) generates soft scores per sentence, and a lighter XLM-RoBERTa-base student is trained to mimic them. That gives us sentence-level detail

with much faster inference. Throughout, the final ranking weights (how we combine S, C, A, TQ, SQ) were kept fixed for transparency; there is no ground-truth "best hospital" to optimize them against, so we treat them as a clear, defensible policy choice. We used the earlier ML/DL baselines and the district-wise results to get our Sentiment scores. We evaluated five classic ML models, Decision Tree, SVM, Logistic Regression, Random Forest, Gaussian Naive Bayes, plus two DL models, CNN and BERT, after standard preprocessing (emoji and blank-row removal, tokenization, vectorization, lower-casing). Those ML models were paired with TF-IDF features. Their role is to provide a clear, explainable baseline before moving to transformers. Across divisions, BERT consistently ranked at or near the top on the four metrics (accuracy, precision, recall, F1), with Random Forest and SVM competitive in some regions. For example, in Rangpur, Random Forest had the highest accuracy and recall at 81.9%, while BERT achieved the best precision (82.2%) and F1 (80.1%). In Rajshahi, BERT led all metrics, accuracy 93.1%, precision 92.8%, recall 93.1%, F1 92.9% . In Dhaka, BERT again topped the board, accuracy 86.6%, precision 86.5%, recall 86.6%, F1 86.4% . In Mymensingh, SVM had the highest precision (87.6%), but BERT produced the best accuracy 88.9%, recall 88.9%, F1 87.3% . Sylhet showed BERT dominating with accuracy 94.1%, precision 94.7%, recall 94.1%, F1 94.4% . In Barisal, BERT again led, accuracy 82.5%, precision 81.9%, recall 82.5%, F1 81.8% . For Khulna, BERT scored accuracy 90.8%, precision 91.0%, recall 90.8%, F1 90.8% (SVM was close) . And in Chittagong, BERT posted accuracy 84.1%, precision 82.4%, recall 84.1%, F1 83.0% . These results, together with your metric definitions (§4.3.1), established a strong review-level sentiment baseline across the country before we introduced aspect-level modeling. Even though BERT (review-level) was very strong, it predicts one sentiment per whole review; it does not tell us which aspect is being discussed or where in the text it occurs. For ranking hospitals by aspect-specific experience (e.g., cleanliness vs treatment quality), we needed sentence-level evidence and aspect tags, hence the shift to zero-shot NLI at the sentence level and the distillation step for speed. In short, the previous baselines proved that transformer models are reliable on your data; the new pipeline adds the granularity and efficiency required for aspect-aware hospital ranking at scale.

5.5 Final Design Adjustments

We finalized the sentence-level aspect detection and sentiment. The teacher is *joeddav/xlm-roberta-large-xnli* (multi-label NLI) with the hypothesis "This sentence is about {}."; the sentiment model is cardiffnlp/twitter-xlm-roberta-base-sentiment. Batches: 8 (teacher) and 16 (sentiment). To convert teacher scores into yes/no per aspect, we used unsupervised thresholds (Otsu) learned per aspect from the data of Dhaka hospital Reviews; cleanliness 0.3875, affordability 0.4075, treatment quality 0.5125, service quality 0.3375 . For diagnostics, we show student decisions at a simple 0.50 probability. We aggregate sentences→reviews (an aspect is present if any sentence crosses its τ; aspect sentiment is the mean of those sentences), then reviews→hospitals by combining mention frequency × sentiment and min–max normalizing to $C, A, TQ, SQ \in [0, 1]$. We keep hospitals with ≥ 5 reviews to avoid sparsity and blend in a confidence term based on log-volume with $\alpha = 0.85$. The final score uses fixed weights that sum to 1: $S = 0.25$, $C = 0.20$, $A = 0.15$, $TQ = 0.25$, $SQ = 0.15$.

For distillation, the student model is trained on xlm-roberta-base for 3 epochs with a learning rate of 2×10^{-5}, batch size of 16 for training and 32 for evaluation, weight decay of 0.01, logging every 50 steps, and evaluation at the end of each epoch. We use a weighted BCE-with-logits loss against the teacher's soft targets.
Similarly, for the dataset consisting of Ranjshahi Hospital Reviews, the weight values and the confidence score remain the same, while the unsupervised threshold learnt is different; cleanliness 0.41, affordability 0.395, treatment quality 0.5325, service quality 0.345. For Khulna Hospital Reviews; cleanliness 0.41, affordability 0.41, treatment quality 0.535, service quality 0.36. For Barishal Hospital Reviews; cleanliness 0.40, affordability 0.40, treatment quality 0.5175, service quality 0.3375
.

5.6 Statistical Analysis

Our dataset was collected from Google Reciews and the Sentiment was labeled by us in a multi-step process. Firstly, each member were assigned two divisions in order to scrape the reviews. After the scrapping, we, first, labelled the sentiment individually; To further eliminate bias, we relabelled the whole dataset four times anonymously. The label column contained four values. Each value either 1,0, or -1 for positive, neutral, or negative. Afterwards, the majority of the four sentiment was selected as the label for a review. Lastly, for any review constesting between two labels, we repeated this anonymous voting process again to reach a decision.
Our working corpus across the eight districts is roughly 27,000 reviews and 67,000 sentences. Using the Otsu-calibrated thresholds on the Dhaka corpus, sentence-level aspect prevalence is: cleanliness 29.4%, affordability 43.6%, treatment quality 68.4%, and service quality 45.1% of sentences. For context, a single fixed cutoff ($\tau = 0.40$) would mark cleanliness 29.0%, affordability 44.2%, treatment quality 72.4%, and service quality 39.0%. In other words, Otsu is stricter for treatment quality (which had a clearer high-score mode) and looser for service quality (which was more diffuse), matching the observed score distributions. The distilled student converged smoothly: training loss drops from ~ 0.256 early to ~ 0.169 by the end of epoch 3, indicating it learned a close approximation of the teacher's probability distribution without overfitting.
At the hospital level, we examined how review volume relates to scores using Spearman correlations with N_{reviews}. The relationships are modestly negative for Affordability ($\rho \approx -0.307$) and Service Quality ($\rho \approx -0.241$), and weak/non-significant for the final score ($\rho \approx -0.118$). Practically, popular hospitals tend to be pricier and face tougher service scrutiny, but "more reviews" does not automatically translate into a higher overall score. Inter-aspect relationships across hospitals are strong and intuitive: treatment quality tracks overall sentiment most closely (TQ$\leftrightarrow$S ≈ 0.747), and cleanliness co-moves with service quality (C$\leftrightarrow$SQ ≈ 0.650). We also see meaningful links between treatment quality and service (≈ 0.655) and between affordability and cleanliness (≈ 0.632), suggesting that operational hygiene and front-of-house experience often rise together, while clinical experience drives how patients feel overall.
These new findings sit on top of the previous ML/DL baselines. Classical ML models (Decision Tree, SVM, Logistic Regression, Random Forest, Gaussian NB) and two deep models (CNN, BERT) were evaluated per division using standard metrics

(accuracy, precision, recall, F1). BERT consistently ranked at or near the top across divisions, with SVM or Random Forest occasionally competitive on specific metrics. That result statistically justifies our use of transformer backbones here. However, those P2 models, and their review-level metrics, could not tell us which aspect drove a given sentiment. Sentence-level analysis resolves that limitation by quantifying aspect prevalence, aspect-specific sentiment, and the correlations between aspects and overall satisfaction.

To keep the statistics robust, we filter to hospitals with at least five reviews and apply a volume-aware confidence multiplier to shrink extreme scores when evidence is thin while preserving high-volume estimates. Our thresholds are data-driven (per-aspect Otsu) rather than one-size-fits-all, which improves the precision/recall trade-off without human labels. Taken together, these choices produce district-level results that are both granular, pinpointing where cleanliness, affordability, treatment quality, and service quality are discussed, and credible, stabilized against small-sample noise and aligned with strong transformer performance seen in the baselines.

5.7 Comparisons and Relationships

5.7.1 Lexicon-Based Models vs Zero-Shot Aspect Detection

Lexicon-based models rely on predefined dictionaries of words associated with specific sentiments and aspects. For example, a lexicon might contain lists like "clean," "spotless," and "hygienic" tagged as cleanliness-related positive words, or "expensive," "costly," and "overpriced" as affordability-related negative words. To analyze a review, these models search for matching words from their dictionaries and count occurrences to determine which aspects are present and whether sentiment is positive or negative. While this approach is simple and rapid, it has significant disadvantages. The construction of full lexicons is time-consuming, especially in specialized domains like medicine where vocabulary constantly evolves. Meaning is challenging to lexicon-based models with depending on context since words convey dissimilar sentiments in different situations. They completely lack the words outside of their dictionary and fail to handle language variations like Banglish, slang, or spelling errors unless they have been specifically crafted. Above all, they require standalone lexicons for each language and must be manually updated each time a new word is added.

The zero-shot aspect detection model, on the other hand, exploits an existing pre-trained multilingual model (XLM-RoBERTa) that handles language contextually rather than matching words. Instead of storing word lists, it classifies whether a review semantically has some connection to an aspect by posing it as a natural language inference task. The model asks the question "Does this review ever refer to cleanliness and hygiene?" and draws on its deep language pattern learning from massive text corpora to answer. This approach has several significant advantages. It doesn't require any human dictionary build and update, handles synonyms, paraphrasing, and context variations automatically, and works between English and Banglish without separate resources. The model represents semantic meaning rather than surface keywords and knows that "the place wasn't properly maintained" is associated with cleanliness even if it does not contain the word "clean." It adapts to

new terminology naturally since it relies on contextual understanding rather than fixed word lists. The zero-shot method is also more flexible, allowing aspect definitions to be adjusted simply by changing the prompt phrases rather than rebuilding entire lexicons. While lexicon-based models might identify 60-70% of aspect mentions through direct keyword matching, the zero-shot approach achieves higher coverage by understanding implicit references and contextual cues. For the hospital rating system, this means more accurate aspect detection, better handling of diverse patient language, and significantly reduced maintenance overhead compared to maintaining multiple sentiment and aspect lexicons for healthcare terminology.

5.7.2 Sectence-level vs Review-level

Comparing sentence-level (student) to our earlier review-level ranking (zero-shot), we find high but not identical agreement: Spearman's $\rho \approx 0.846$ and Kendall's $\tau \approx 0.674$ across shared hospitals, but only 4/10 overlap in the Top-10. In other words, sentence-level modeling reorders the very top: hospitals that looked strong from overall reviews sometimes drop when we require explicit sentence-level evidence per aspect, and others rise when targeted praise (e.g., for treatment quality) is no longer diluted by unrelated text in the same review. The biggest positive movers under sentence-level evidence include *Manikganj United Hospital* (Δ rank $\approx +119$), *Pain Care* ($+115$), and *Dr. Shojib's Diabetes Asthma & Medical Center* ($+109$). The largest declines are *Dhaka Mohanagar General Hospital* (-122), *Di Khidmah Jonata Hospital* (-111), and *Dr. Mr. Khan Shishu Hospital & Institute of Child Health–Mirpur 2* (-106). These shifts are exactly what we expect when moving from broad review labels to granular sentence-level detection tied to each aspect.

5.8 Discussions

We grounded our design choices on two recalcitrant truths about the data: we lack human labels, and the reviews are noisy, code-mixed, and inclined to talk about many things at once. In review-level baselines previously, conventional ML models (SVM, Random Forest, Logistic Regression, etc.) and deep models (CNN, BERT) gave solid global sentiment accuracy across divisions, and BERT consistently ranked among the top. Those conclusions taught us that transformer encoders are powerful on this space, but they uncovered also a blind spot: a single label per review can't inform us which part of the text is talking about cleanliness, price, treatment quality, or service. That observation motivated the core shift in this work from review-level to sentence-level modeling. By asking a multilingual NLI model whether each sentence relates to a target aspect, we avoid "topic dilution," represent mixed opinions in one review, and handle Banglish code-mixing more gracefully than lexicons or bag-of-words features.

The distillation makes that possible at scale. The slow but consistent zero-shot teacher (XLM-RoBERTa-Large, XNLI) learns to estimate the teacher's soft scores and runs at a fraction of the expense as a distilled student (XLM-RoBERTa-Base). In practical terms, this meant that we could keep sentence-level granularity without bloating inference time, which is desirable when working with tens of thousands of sentences on Colab-class hardware. The student's smooth loss plots and strong

teacher–student agreement plots indicate that we did not trade away too much signal for speed.

We kept the last ranking weights fixed rather than learned intentionally. Without ground-truth labels showing us which hospitals are "best," any weighting learned would be numerically arbitrary and harder to justify. Adjustment for weight makes the policy of scoring reproducible and auditable: treatment quality and overall sentiment are paramount; cleanliness is secondary; affordability and service quality also count but not as much. This openness is complemented by two data defense strategies against noise and imbalance: (i) we condition on hospitals having at least five reviews and (ii) we apply a volume-aware confidence modifier to gently shrink small-sample scores towards the mean while high-volume hospitals remain intact.

Empirically, the findings are coherent and clinically sound. The quality of care is the biggest determinant of overall satisfaction, and service moves in tandem with cleanliness, perhaps suggesting that front-of-house experience and operational hygiene are interconnected. While this is happening, the analogy of sentence vs review level tells us why granularity matters: headline concurrence is fine, but Top-10 reorganizations follow once we insist on explicit sentence-level evidence for every argument and others rise as concentrated praise is no longer diluted by irrelevant text. This is precisely what we wished to have: a ranking by specifically what people say something about, not what they are generally feeling.

There are limitations. Thresholds are data-driven heuristics (Otsu for teacher marks) and not gold-standard thresholds; the teacher itself is itself a pseudo-reference, not an oracle human; and very informal Banglish slang, sarcasm, or field-specific acronyms might still catch the models off guard. Aggregation choices (e.g., min–max normalization by hospital, the particular confidence function, and the application of fixed weights) include value judgments even when kept transparently traced. Finally, district-level data can be susceptible to looking at volume and population composition; we protect against this with the

*ge*5 condition and the mixture of confidence. This is not, however, the final solution to the massive data unbalance we were facing on district-level data, but they are no panaceas for sampling bias in general. A small annotated chunk, even a few hundred sentences in each dimension, would enable us to calibrate thresholds, compute actual precision/recall, and validate the teacher's choices against human instincts. We could explore probability calibration of the student (e.g., temperature scaling) to narrow agreement at decision boundaries. At the ranking level, we might conduct weight-sensitivity tests or use stakeholder-driven schemes (e.g., Analytic Hierarchy Process) to ensure the order stays stable for probable shifts in preferences. Finally, moderate active learning, presenting uncertain sentences to quick human judgment, would increasingly improve thresholds and the student while keeping annotation cost low.

In short, the sentence-level, distilled, threshold-calibrated pipeline improves over past review-level systems by offering actionable granularity without being unfeasible. It adds the strengths demonstrated by the past BERT baselines, reinforces them with aspect awareness which institutions can act upon, and maintains transparency as to where the numbers are obtaining their sources and how sensitive they are to these sources, key ingredients in achieving a credible, deployable analysis of patient feedback.

Chapter 6

Conclusion

6.1 Contributions to the Field

Our major contribution of this work is the manner in which we constructed the dataset. We gathered over 27,000 reviews comprised of nearly 67,000 sentences from hospitals in all eight divisions of Bangladesh. These reviews came with numerous challenges such as being unstructured, a combination of English and Bangla, spelling errors, slang, and sarcasm. In order to address these, we conducted extensive preprocessing operations. These included tokenization, translation, typo correction, and manual filtering of texts. The result was a clean, analysis-ready corpus, the first in the country. Previously, the only available dataset contained only the reviews of hospitals in Dhaka. We extended on that work and created a dataset containing all the Google reviews from all the hospitals in Bangladesh. This can now be used for other futures research work. Furthermore, we manually labeled each review with the proper sentiment; doing so, this work contributes a practical, auditable, and scalable framework for turning noisy, code-mixed patient reviews into aspect-aware hospital ratings at a national scale, which is perfect for supervised training and unsupervised clustering.

Building on this, we developed an aspect-aware sentiment analysis pipeline using zero-shot sentence-level classification with XLM-RoBERTa-Large as the teacher model. We transferred this knowledge to a lighter XLM-RoBERTa-Base student model through multi-label training, achieving high agreement (F1 ≈ 0.9) with improved efficiency, even though simplifying the model may lead to some loss of details. Methodologically, we also show how to achieve aspect-based sentiment analysis without any manual labels by combining a sentence-level zero-shot "teacher" (XLM-RoBERTa-Large, XNLI) with a compact, distilled "student" (XLM-RoBERTa-Base). The paper adds further evidence of what patients talk about hospitals in Bangladesh. We quantify correlations between domains (e.g., how monitoring treatment quality forecasts satisfaction overall; cleanliness co-varying with service) and show how affordability and service scores behave when review volume grows. These results are not better scores; they are narratives that enable hospital managers to see what to fix.

From a systems perspective, the pipeline is multilingual, lightweight, and cost-effective. It runs on commodity cloud GPUs, uses wholly open-source tooling, and does not need an annotation campaign to get started, lowering the bar for public health teams that require dynamic, district-level monitoring.

Finally, this project bridges two communities that are likely to be separate in reality: (i) applied health-informatic stakeholders who need clear, cost-effective scoring schemes, and (ii) modern NLP able to handle noisy, under-resourced languages. By generating both a strong transformer-based backbone and policy-actionable scoring recipe, we provide an example that can be replicated: start with zero-shot sentence-level detection, distill for efficiency, calibrate per aspect, and release fixed, documented aggregation rules. In short, our work is a complete end-to-end playbook for building robust, interpretable, and scalable ratings from real patient stories.

6.2 Limitations and Future Work

The proposed sentiment-based hospital rating system demonstrates strong performance, with BERT achieving up to 94.1% accuracy and effective zero-shot aspect detection across *cleanliness*, *affordability*, *treatment quality*, and *service quality*. The system addresses data imbalance through a confidence scoring mechanism based on logarithmic scaling:

$$\text{conf} = \frac{\log(1+N)}{\max_h \log\big(1+N_h\big)},$$

where N is the number of reviews (and N_h the count for hospital h). We then apply a confidence multiplier to stabilize rankings:

$$\text{final\_score} = \alpha \times \text{base\_score} + (1-\alpha) \times \text{conf}, \qquad \alpha = 0.85.$$

This guarantees that low-volume hospitals retain at least 85% of their base score, preventing undue penalization while still acknowledging volume differences. In practice, the scheme prevents individual reviews from disproportionately affecting smaller hospitals while giving proper credit to facilities with substantial patient feedback.

Despite this robust foundation, one limitation remains: the current review-level sentiment attribution approach assigns the entire review's sentiment to each detected aspect, which can misrepresent cases where patients express different opinions about different services within the same review. In addition, the system presently processes reviews in batches rather than providing real-time updates.

A high-impact improvement is *sentence-level* sentiment analysis. This change would evaluate the sentiment associated with each aspect mention at the sentence level, accurately handling cases where, for example, treatment quality is praised while waiting times or billing are criticized. Our documents already show that the sentence-level variant improves precision by separating mixed opinions instead of blending them into a single review-level label. Building a real-time, dynamic platform where new reviews automatically refresh rankings would keep the information current for time-sensitive decisions.

Another important direction is the integration of *static hospital information*, which the current system deliberately excludes. Such data include bed capacity, specialties, accreditation, equipment availability, physician credentials, operating hours, and infrastructure (e.g., ICU beds, diagnostic facilities). The present work focuses on a fully dynamic, sentiment-driven layer to capture evolving patient perspectives. Future iterations can combine this with static metrics to form a *hybrid* model that balances subjective experience with objective infrastructure. This would enable comprehensive profiles such as "50-bed facility with cardiac specialization and 24/7

emergency services," shown alongside sentiment-based quality scores, thereby supporting more informed patient choices.

Additional enhancements include broadening data sources beyond Google Maps to multiple platforms, social media, and offline surveys for better coverage; tuning the system's parameters (e.g., the zero-shot threshold $\tau = 0.40$, aspect weights, and the confidence floor α) as more data accrue; and adding bias-detection mechanisms to filter promotional content or coordinated fake reviews. Finally, adapting this framework to other domains, education, retail, or public services, would illustrate its generality and help promote broader, data-driven quality assessment in Bangladesh.

Bibliography

[1] Crowdflower, *Twitter us airline sentiment dataset*, Accessed: 2025-01-31, 2015. [Online]. Available: https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment.

[2] K. Dashtipour et al., "Multilingual sentiment analysis: State of the art and independent comparison of techniques," *Cognitive Computation*, vol. 8, pp. 757–771, 2016. DOI: 10.1007/s12559-016-9415-7. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC4981629/pdf/12559_2016_Article_9415.pdf.

[3] R. A. Tuhin, B. K. Paul, F. Nawrine, M. Akter, and A. K. Das, "An automated system of sentiment analysis from bangla text using supervised learning techniques," in *2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS)*, 2019, pp. 360–364. DOI: 10.1109/CCOMS.2019.8821658.

[4] G. Xu, Z. Yu, H. Yao, F. Li, Y. Meng, and X. Wu, "Chinese text sentiment analysis based on extended sentiment dictionary," *IEEE Access*, vol. 7, pp. 43 749–43 762, 2019. DOI: 10.1109/ACCESS.2019.2907772.

[5] N. C. Dang, M. N. Moreno-García, and F. De la Prieta, "Sentiment analysis based on deep learning: A comparative study," *Electronics*, vol. 9, no. 3, 2020, ISSN: 2079-9292. DOI: 10.3390/electronics9030483. [Online]. Available: https://www.mdpi.com/2079-9292/9/3/483.

[6] K. Nawab, G. Ramsey, and R. Schreiber, "Natural language processing to extract meaningful information from patient experience feedback," *Applied Clinical Informatics*, vol. 11, no. 2, pp. 242–252, Mar. 2020. DOI: 10.1055/s-0040-1708049. [Online]. Available: https://doi.org/10.1055/s-0040-1708049.

[7] N. R. Bhowmik, M. Arifuzzaman, M. R. H. Mondal, and M. S. Islam, "Bangla text sentiment analysis using supervised machine learning with extended lexicon dictionary," *Natural Language Processing Research*, vol. 1, pp. 34–45, 3-4 2021, ISSN: 2666-0512. DOI: 10.2991/nlpr.d.210316.001. [Online]. Available: https://doi.org/10.2991/nlpr.d.210316.001.

[8] L. Khan, A. Amjad, N. Ashraf, H.-T. Chang, and A. Gelbukh, "Urdu sentiment analysis with deep learning methods," *IEEE Access*, vol. PP, no. 99, pp. 1–1, 2021. DOI: 10.1109/ACCESS.2021.3093078. [Online]. Available: https://www.researchgate.net/publication/352810896_Urdu_Sentiment_Analysis_With_Deep_Learning_Methods.

[9] L. N. Pathi, *Imdb dataset of 50k movie reviews*, Accessed: 2025-01-31, 2021. [Online]. Available: https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews.

[10] M. T. H. K. Tusar and M. T. Islam, "A comparative study of sentiment analysis using NLP and different machine learning techniques on US airline twitter data," *CoRR*, vol. abs/2110.00859, 2021. arXiv: 2110.00859. [Online]. Available: https://arxiv.org/abs/2110.00859.

[11] L. Khan, A. Amjad, and N. e. a. Ashraf, "Multi-class sentiment analysis of urdu text using multilingual bert," *Scientific Reports*, vol. 12, p. 5436, 2022. DOI: 10.1038/s41598-022-09381-9. [Online]. Available: https://doi.org/10.1038/s41598-022-09381-9.

[12] W. Ccoya and E. Pinto, *Comparative analysis of libraries for the sentimental analysis*, 2023. arXiv: 2307.14311 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2307.14311.

[13] K. P. Gunasekaran, "Exploring sentiment analysis techniques in natural language processing: A comprehensive review," *arXiv preprint arXiv:2305.14842*, 2023, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 8, Issue 1, January 2019. [Online]. Available: https://arxiv.org/abs/2305.14842.

[14] M. Hasan, L. Islam, I. Jahan, S. M. Meem, and R. M. Rahman, "Natural language processing and sentiment analysis on bangla social media comments on russia–ukraine war using transformers," *Vietnam Journal of Computer Science*, vol. 10, no. 03, pp. 329–356, 2023. DOI: 10.1142/S2196888823500021. eprint: https://doi.org/10.1142/S2196888823500021. [Online]. Available: https://doi.org/10.1142/S2196888823500021.

[15] K. R. Mabokela, T. Celik, and M. Raborife, "Multilingual sentiment analysis for under-resourced languages: A systematic review of the landscape," *IEEE Access*, vol. 11, pp. 15 996–16 020, 2023. DOI: 10.1109/ACCESS.2022.3224136.

[16] U. Sehar, S. Kanwal, and N. e. a. Allheeib, "A hybrid dependency-based approach for urdu sentiment analysis," *Scientific Reports*, vol. 13, p. 22 075, 2023. DOI: 10.1038/s41598-023-48817-8. [Online]. Available: https://doi.org/10.1038/s41598-023-48817-8.

[17] P. Seth, R. Goel, K. Mathur, and S. Vemulapalli, "RSM-NLP at BLP-2023 task 2: Bangla sentiment analysis using weighted and majority voted fine-tuned transformers," in *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, F. Alam, S. Kar, S. A. Chowdhury, F. Sadeque, and R. Amin, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 305–311. DOI: 10.18653/v1/2023.banglalp-1.40. [Online]. Available: https://aclanthology.org/2023.banglalp-1.40/.

[18] A. A. Tamjid, G. P. Galpo, K. B. Urmi, F. S. Chitto, and S. Annafi, "An advanced hospital rating system using machine learning and natural language processing," BRAC University, May 2023. [Online]. Available: http://hdl.handle.net/10361/21912.

[19] N. E. Zekaouiu, S. Yousfi, M. Rhanoui, and M. Mikram, "Analysis of the evolution of advanced transformer-based language models: Experiments on opinion mining," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 12, no. 4, p. 1995, Dec. 2023, ISSN: 2089-4872. DOI: 10.11591/ijai.v12.i4.pp1995-2010. [Online]. Available: http://dx.doi.org/10.11591/ijai.v12.i4.pp1995-2010.

[20] B. Zhang, "A bert-cnn based approach on movie review sentiment analysis," *SHS Web of Conferences*, vol. 163, p. 04 007, 2023. DOI: 10.1051/shsconf/202316304007. [Online]. Available: https://www.shs-conferences.org/articles/shsconf/abs/2023/12/shsconf_icssed2023_04007/shsconf_icssed2023_04007.html.

[21] JP797498e, *Twitter entity sentiment analysis dataset*, Accessed: 2025-01-31, 2024. [Online]. Available: https://www.kaggle.com/datasets/jp797498e/twitter-entity-sentiment-analysis.

[22] M. Miah, M. Kabir, T. Sarwar, et al., "A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and llm," *Scientific Reports*, vol. 14, p. 9603, 2024. DOI: 10.1038/s41598-024-60210-7. [Online]. Available: https://doi.org/10.1038/s41598-024-60210-7.

[23] The Business Standard, "India saw 48% jump in medical tourists from bangladeshis in 2023," *The Business Standard*, 2024, Accessed: 2025-01-31. [Online]. Available: https://www.tbsnews.net/bangladesh/health/india-saw-48-jump-medical-tourists-bangladeshis-2023-890526.

[24] The Financial Express, "Bangladeshis spend $4.0b for medical tourism every year," *The Financial Express*, 2024, Accessed: 2025-01-31. [Online]. Available: https://thefinancialexpress.com.bd/national/bangladeshis-spend-40b-for-medical-tourism-every-year.

[25] Statista R, *Statista r: Research & insights*, https://r.statista.com/en/, Accessed: 18 June 2025, 2025.

[26] R. Zhang, L. Nie, C. Zhao, and Q. Chen, *Achieving semantic consistency: Contextualized word representations for political text analysis*, 2025. arXiv: 2412.04505 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2412.04505.