

**ADVANCING SENTIMENT ANALYSIS IN BENGALI: BRIDGING
LINGUISTIC GAPS IN NLP WITH MACHINE AND DEEP
LEARNING MODELS**

BY

**Abdullah Al Masud
ID: 192-15-13140**

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

Supervised By

Mr. Md. Abbas Ali Khan
Assistant Professor
Department of CSE
Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

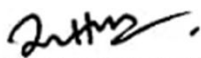
DHAKA, BANGLADESH

JANUARY 2024

APPROVAL

This Project titled “Advancing Sentiment Analysis in Bengali: Bridging Linguistic gaps in NLP with Machine Learning and Deep Learning Models”, submitted by **Abdullah AL Masud, Student ID: 192-15-13140** to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 24 January, 2024.

BOARD OF EXAMINERS



Dr. Md. Zahid Hasan (ZH)
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Chairman

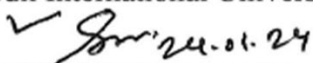


Amit Chakraborty Chhoton (ACC)
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Md Assaduzzaman (MA)
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner




Dr. Mohammed Nasir Uddin (DNU)
Professor
Department of Computer Science and Engineering
Jagannath University

External Examiner

DECLARATION

I hereby declare that this project has been done by me under the supervision of, **Mr. Md. Abbas Ali Khan, Assistant Professor, Department of CSE** Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

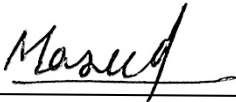
Supervised by:



Md. Abbas Ali Khan (AAK)

Assistant Professor
Department of CSE
Daffodil International University

Submitted by:



Abdullah Al Masud

ID: -192-15-13140
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, I express my heartiest thanks and gratefulness to almighty God for His divine blessing makes me possible to complete the final year project/internship successfully.

I am really grateful and wish my profound indebtedness to **Mr. Md. Abbas Ali Khan Sir, Assistant Professor**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of my supervisor in the field of Data Mining, Machine Learning (ML), Deep learning, Natural language processing (NLP) to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

I would like to express my heartiest gratitude to **Mr. Md. Abbas Ali Khan Sir, Assistant Professor & Dr. Sheak Rashed Haider Noori, Professor and Head, Department of CSE, Daffodil International University, Dhaka**. For his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

I would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, I must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

This paper presents a comprehensive study on sentiment analysis in the Bengali language, focusing on user-generated comments from online shopping websites. Despite the significant number of Bengali speakers worldwide, the language remains underrepresented in natural language processing (NLP) research. This study aims to bridge this gap by applying advanced sentiment analysis techniques to better understand customer opinions and preferences in the e-commerce domain. The research involved a meticulous process of data collection, where 1995 comments were extracted using web scraping techniques. Rigorous preprocessing methods, including text cleaning, normalization, tokenization, and Term Frequency-Inverse Document Frequency (TF-IDF) vectorization, were employed to prepare the dataset for analysis. A variety of machine learning (ML) models, such as Logistic Regression, Decision Trees, Random Forest, Multi. Naive Bayes, KNN, SVM, and SGD, along with deep learning (DL) models like LSTM, Bi-LSTM, and CNN, were trained and evaluated on this dataset. The results revealed that while traditional ML models like SVM and SGD showed strong performance, deep learning models, particularly Bi-LSTM, demonstrated superior ability in sentiment classification. This was attributed to their effectiveness in capturing contextual nuances and complex linguistic patterns inherent in Bengali. The study underscored the challenges of processing a morphologically rich language like Bengali and the importance of choosing the right model for effective sentiment analysis. Furthermore, the research addressed the societal, ethical, and environmental implications of implementing sentiment analysis tools. It highlighted the need for responsible data usage, bias mitigation, transparency in model application, and sustainable computing practices. In conclusion, the research contributes significantly to the field of sentiment analysis in less-studied languages, providing valuable insights for businesses, policymakers, and researchers. It paves the way for more inclusive technological advancements in AI and NLP, ensuring linguistic diversity is embraced and respected in the digital age.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	i
Declaration	ii
Acknowledgements	iii
Abstract	iv
CHAPTER	
CHAPTER 1: INTRODUCTION	1-4
1.1 Introduction	1
1.2 Motivation	2
1.3 Rationale of the Study	3
1.4 Research Questions	3
1.5 Expected Output	3
1.6 Project Management and Finance	4
1.7 Report Layout	4
CHAPTER 2: BACKGROUND	6-12
2.1 Preliminaries/Terminologies	6
2.2 Related works	7
2.3 Comparative Analysis and summary	8
2.4 Scope of the Problem	10
2.5 Challenges	11
CHAPTER 3: RESEARCH METHODOLOGY	13-20
3.1 Research Subject and Instrumentation	13
3.2 Data Collection Procedure/Dataset Utilized	14
3.3 Statistical Analysis	15

3.4 Proposed Methodology/Applied Mechanism	17
3.5 Implementation Requirements	19
CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION	21-25
4.1 Experimental Setup	23
4.2 Experimental Results & Analysis	23
4.3 Discussion	24
4.4 Future Research Directions	25
CHAPTER 5: IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY	26-27
5.1 Impact on Society	26
5.2 Impact on Environment	26
5.3 Sustainability Plan	27
CHAPTER 6: SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH	28-30
6.1 Summary of the Study	28
6.2 Conclusions	29
6.3 Implication for Further Study	29
APPENDIX	31
REFERENCES	32
PLAGIARISM REPORT	34

LIST OF FIGURES

FIGURES	PAGE NO
Figure 3.4.1: Overall methodology	18
Figure 3.4.2: Best Model Selection	19
Figure 4.1.1: Data Preparation	22

LIST OF TABLES

TABLES	PAGE NO
Table 2.3.1: Comparative analysis	9
Table 3.3.1: Dataset Summary	16
Table 4.1.2: Deep Learning Models Configuration	26
Table 4.2.1: Machine Learning Models Performance	27
Table 4.2.2: Deep Learning Models Results	27

CHAPTER 1

Introduction

1.1 Introduction

In the digital era, the ability to understand customer opinions and sentiments through their online feedback has become increasingly important. Sentiment analysis, a critical branch of natural language processing (NLP), plays a vital role in deciphering these opinions. It involves the computational study of opinions, emotions, and sentiments expressed in text form, transforming subjective information into actionable insights [1]. This research focuses on sentiment analysis of Bengali, a language spoken by over 230 million people worldwide, predominantly in Bangladesh and parts of India. Despite its large number of speakers, Bengali has seen limited exploration in the realm of NLP, especially in sentiment analysis. This gap is notable given the significant growth of digital content in Bengali due to the rising internet penetration in the region [2].

Bengali's rich linguistic heritage, complex morphological structures, and diverse dialects present unique challenges for sentiment analysis. These linguistic characteristics create opportunities for advanced computational techniques and innovative approaches in NLP [3]. The aim of this research is to apply machine learning and deep learning techniques to analyze sentiments expressed in Bengali text, with a particular focus on online shopping reviews. The burgeoning e-commerce sector in Bengali-speaking regions produces vast amounts of customer feedback, which, until now, has been underutilized due to the lack of effective sentiment analysis tools for Bengali [4].

This study contributes to the understanding of sentiment analysis in Bengali by employing a combination of machine learning and deep learning methods. Such analysis is crucial for businesses and researchers to comprehend customer feedback and preferences accurately. Moreover, this research aligns with the global push towards linguistic inclusivity in technology, addressing the need for NLP research in languages other than English [5]. Bengali, being one of the most spoken languages globally, represents a significant portion of the online user base. Therefore, developing robust NLP tools for Bengali not only steps towards linguistic inclusivity but also opens avenues for more personalized and effective digital services for a large population [6].

Through this research, I demonstrated the applicability of existing NLP techniques to Bengali and address the specific challenges posed by its linguistic characteristics. The successful implementation of this research can serve as a model for sentiment analysis in other less-represented languages, contributing to the broader field of NLP and its applications in diverse linguistic contexts [7].

1.2 Motivation

The motivation for this research is twofold. Firstly, the continuously growing e-commerce sector in Bengali-speaking regions has led to an abundance of online customer reviews, which are a goldmine of insights into customer satisfaction and preferences. However, the lack of advanced sentiment analysis tools for Bengali means that these valuable data sources remain largely untouched. This research aims to fill this technological void by developing effective sentiment analysis models tailored for the Bengali language. Secondly, the global push towards linguistic inclusivity in technology has highlighted the need for NLP research in languages other than English. Bengali, being one of the most spoken languages globally, represents a significant portion of the online user base. Developing robust NLP tools for Bengali is not only a step towards linguistic inclusivity but also opens up avenues for more personalized and effective digital services for a large population. This research is motivated by the potential impact it could have on businesses, governments, and individuals by providing them with tools to better understand and respond to the sentiments of the Bengali-speaking community.

Through this research, I aim to demonstrate the applicability of existing NLP techniques to Bengali and to address the specific challenges posed by its linguistic characteristics. The successful implementation of this research can serve as a model for sentiment analysis in other less-represented languages, contributing to the broader field of NLP and its applications in diverse linguistic contexts.

1.3 Rationale of the Study

The rationale behind this study is rooted in the recognition of a significant gap in the field of natural language processing (NLP) for the Bengali language. Despite its vast number of speakers, Bengali is underrepresented in technological advancements, particularly in sentiment analysis. This research seeks to address this disparity by exploring and applying

NLP techniques to the Bengali language, specifically focusing on sentiment analysis of online shopping reviews. The study is driven by the necessity to understand customer sentiments in e-commerce platforms, which are rapidly growing in Bengali-speaking regions. It aims to provide a comprehensive analysis tool that can accurately categorize sentiments, thereby aiding businesses and researchers in understanding customer feedback and preferences. Moreover, this study contributes to the broader goal of linguistic diversity in AI and NLP, ensuring that technological advancements are not limited to widely-spoken Western languages but are extended to other global languages, including Bengali.

1.4 Research Questions

- I. How effectively can machine learning and deep learning models perform sentiment analysis on Bengali text data, specifically from online shopping reviews?
- II. What are the challenges faced when applying sentiment analysis techniques to Bengali, and how can these be addressed?
- III. Which models and algorithms yield the most accurate results for sentiment classification in Bengali, and why?
- IV. How do the linguistic characteristics of Bengali, such as its morphology and syntax, impact the process of sentiment analysis?
- V. What insights can be drawn about customer opinions and preferences from the sentiment analysis of Bengali online shopping reviews?

1.5 Expected Output

The expected outputs of this research are multifaceted and aim to contribute both to the academic field and practical applications. They include:

Developed Models: Robust machine learning and deep learning models specifically tuned for sentiment analysis in Bengali. These models will be capable of accurately classifying sentiments as positive or negative.

Comparative Analysis: An in-depth comparative analysis of various models and algorithms, highlighting their effectiveness and limitations in processing Bengali text data.

Methodological Framework: A detailed methodological framework for conducting sentiment analysis in Bengali, which can be adapted or extended to other languages with similar linguistic features.

Linguistic Insights: Insights into the linguistic aspects of Bengali that affect sentiment analysis, contributing to the understanding of NLP applications in morphologically rich languages.

Practical Tool for Businesses: A practical tool for businesses and e-commerce platforms targeting Bengali-speaking customers, helping them to understand and respond to customer feedback more effectively.

Through these outputs, the research aims to not only fill a significant gap in NLP research for Bengali but also to provide a foundation for future studies and practical applications involving sentiment analysis in less-represented languages.

1.6 Project Management and Finance

The project was divided into several phases, including data collection, preprocessing, model development, and analysis. Each phase was allocated a specific time frame and resources.

Financial Aspects: Data Collection and Processing: Minimal cost was involved, primarily for web scraping tools and server usage for data storage.

Software and Tools: Open-source software such as Python, Scrapy, and various machine learning libraries were utilized, keeping the software costs negligible.

Computational Resources: The project primarily relied on existing computational resources. Any additional costs incurred were for cloud computing services used for training complex deep learning models.

Miscellaneous Expenses: These included costs for literature resources, potential software upgrades, and contingency funds for unforeseen technical requirements.

The project was self-funded or supported by institutional grants, and a detailed budget plan was maintained to track expenses and ensure the efficient allocation of resources.

1.7 Report Layout

This report is structured into six main chapters, each focusing on a different aspect of the research:

Introduction

Provides an overview of the research, its motivation, rationale, research questions, expected outcomes, and details on project management and finances.

Background

Explores preliminary terminologies, reviews related works, conducts a comparative analysis, and outlines the scope and challenges of the research.

Research Methodology

Details the research subject, data collection procedures, statistical analysis methods, proposed methodology, and implementation requirements.

Experimental Results and Discussion

Presents the experimental setup, results, and a thorough discussion of the findings, analyzing the performance of different models and techniques used.

Impact on Society, Environment, and Sustainability

Examines the broader impact of the research on society, environmental considerations, ethical aspects, and plans for sustainable continuation of the research.

Summary, Conclusion, and Recommendations for Future Research

Summarizes the study, draws conclusions, provides recommendations based on the findings, and suggests implications for further research in the field.

CHAPTER 2

Background

2.1 Preliminaries/Terminologies

To establish a solid foundation for this research, it is important to first understand the key terminologies and concepts that explain sentiment analysis, especially in the context of natural language processing (NLP) for the Bengali language. This section elucidates these terms, providing a base for the subsequent discussion and analysis.

Sentiment Analysis: A computational approach within NLP that involves identifying and categorizing opinions or emotions within text data. It typically involves determining whether the sentiment is positive, negative, or neutral.

Natural Language Processing (NLP): An area of computer science and artificial intelligence concerned with the interactions between computers and human (natural) languages. It involves enabling computers to read, understand, and derive meaning from human languages.

Machine Learning (ML): A subset of AI that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. It plays a crucial role in NLP for developing models that can analyze and interpret language data.

Deep Learning: An advanced subset of ML based on artificial neural networks. Deep learning models, especially recurrent neural networks (RNNs), convolutional neural networks (CNNs), and transformers, are pivotal in complex NLP tasks.

Bengali Language: An Indo-Aryan language predominantly spoken in the Bengal region of South Asia, including Bangladesh and parts of India. It is characterized by its rich literary heritage and complex grammatical structure.

Embedding Layer: In the context of NLP models, an embedding layer is a trainable layer that converts word tokens into dense vectors of fixed size, capturing semantic information about words.

Recurrent Neural Networks (RNNs): A class of neural networks effective in processing sequences, such as text, where current inputs are dependent on previous computations. LSTM (Long Short-Term Memory) and Bi-LSTM (Bidirectional Long Short-Term Memory) are advanced types of RNNs.

Convolutional Neural Networks (CNNs): Originally known for image processing, CNNs are also used in NLP to process data in a grid-like topology, such as sequences of words or characters.

Transfer Learning: A machine learning technique where a model developed for one task is reused as the starting point for a model on a second task. BERT (Bidirectional Encoder Representations from Transformers) is an example of a model based on transfer learning, extensively used in NLP.

Feature Extraction: The process of transforming raw data into numerical features understandable by machine learning algorithms. In NLP, this includes converting text data into vectors or embeddings.

Scrapy: An open-source web-crawling framework used for extracting the data from websites, which was utilized in this research for collecting Bengali text data.

Understanding these terminologies is crucial for comprehending the methodologies, challenges, and analyses presented in this research. They form the backbone of the study and are referenced throughout the report to explain the processes and findings in sentiment analysis of Bengali text data.

2.2 Related Works

The exploration of sentiment analysis spans across various domains and languages, leveraging diverse methodologies and models. The literature reflects a rich tapestry of approaches and applications:

Social Media Analysis: Al-Sabbagh et al. [1], Xu et al. [2], and Yessenov and Misailovic [3] demonstrate the use of deep learning and vector-based approaches in analyzing social media content, from Arabic tweets to hotel and movie reviews, indicating the adaptability of sentiment analysis in diverse social media contexts.

Sentiment in News and Public Opinion: Rahab et al. [4] and Mukwazvure et al. [8] explore sentiment analysis in news comments, revealing the challenges of quantifying sentiment in journalistic and public opinion texts.

Digital Platforms and Consumer Feedback: Lassance et al. [5], Mai et al. [10], and Koong Lin et al. [12] show the effectiveness of neural networks and lexicon-based methods in analyzing YouTube and consumer comments, highlighting the importance of sentiment analysis in understanding customer feedback in various digital platforms.

Software Development and Education: The works of Guzman et al. [6] and Nasim et al. [7] extend sentiment analysis to software development and educational feedback, showcasing its relevance in these domains for enhancing user experience and educational quality.

Healthcare and Public Services: Greaves et al. [9] and Ramírez-Tinoco et al. [25] utilize machine learning models for healthcare service feedback, emphasizing the potential of sentiment analysis in improving public services.

Diverse Language Contexts: Studies like Al-Amin et al. [11], Soliman et al. [20], and Rafique et al. [21] focus on less-researched languages like Bengali, Arabic slang, and Roman Urdu, contributing to the field's inclusivity and diversity.

Pop Culture and Online Communities: The analysis of comments on platforms like YouTube and Twitch.tv by Nizar et al. [19], Novendri et al. [22], and Kobs et al. [28] reflects sentiment analysis's role in understanding viewer engagement in pop culture and online communities.

Big Data Applications: Liu's [30] work on big data sentiment analysis using advanced models like Dropout Regularized CNN underlines the scalability and effectiveness of sentiment analysis techniques in handling large datasets.

2.3 Comparative Analysis and Summary

The literature review on sentiment analysis reveals a wide range of applications and methodologies, each tailored to specific domains and languages. A comparative analysis highlights several key trends and observations:

Domain-Specific Tailoring: Sentiment analysis is tailored to specific domains to address unique challenges. For instance, YouTube comments analysis in Brazilian Portuguese [5] requires different preprocessing and methodological considerations compared to the sentiment analysis of comments on "Money Heist" on YouTube [22]. These differences underscore the need for domain-specific sentiment analysis approaches.

Methodological Diversity: The studies show a range from traditional machine learning techniques to advanced deep learning models. For example, the use of Naïve Bayes classifiers [27] represents a more traditional approach, while the employment of Deep Neural Networks [5] indicates the field's advancement towards more complex models capable of handling nuanced linguistic data.

Language-Specific Challenges: Research in less-represented languages, such as Bengali sentiment analysis [11], showcases the challenges presented by complex morphologies and diverse dialects. This emphasizes the importance of developing language-specific models that can navigate these complexities.

Accuracy and Contextual Understanding: Some studies, like the sentiment analysis of "Money Heist" YouTube comments [22], have achieved notable accuracy rates (81%). These results highlight that while high accuracy is desirable, the practical applicability of sentiment analysis models depends on their ability to contextualize sentiments within the language.

Impact of Lexicon and Vocabulary: The lexicon-based sentiment analysis of comments from the PTT Car Board in Taiwan [12] and the use of sentiment dictionaries in the Bengali comments analysis [11] highlight the significance of employing rich and contextually relevant vocabularies, which are critical for the accuracy of sentiment classification.

Table 2.3.1: Comparative analysis

Paper serial	Author Name	Used Dataset	Method & Techniques	Result
05	Alexandre Ashade Lassance Cunha et al.	YouTube video comments in Brazilian Portuguese	Deep Neural Network	84% for Video 1, 62%-64% for Video 2
08	Addlight Mukwazvure, K.P Supreethi	Comments from The Guardian website	SVM, kNN	kNN: 74.24% (Tech), 56.27% (Politics)
09	Felix Greaves et al.	Online comments about hospitals on the NHS website	Naïve Bayes multinomials, Decision Trees	80.8% - 89.2%
11	Md. Al-Amin et al.	Bengali comments	Word2Vec with sentiment information of words	Accuracy: 75%
13	Hanif Bhuiyan et al.	YouTube video comments	Sentiment analysis with SentiStrength	Varies by category, e.g., 75.435% for science and technology videos
27	Michael Thomas Moore	Sentiment Analysis on LibQUAL+ Comments	Naïve Bayes Classifier	Total Accuracy: 70.7%

In summary, sentiment analysis remains a dynamic field, rich in methodological diversity and domain-specific applications. The evolution of machine learning techniques, coupled with the increasing complexity of data and language nuances, continues to shape the development of sentiment analysis tools. Studies like those of Md. Al-Amin make significant contributions to the inclusivity of sentiment analysis by addressing languages that are typically underrepresented in NLP research, thus broadening the field's applicability and relevance.

2.4 Scope of the Problem

Sentiment analysis, particularly in the Bengali language, addresses a critical need in the field of natural language processing (NLP). The primary scope of this problem encompasses several key areas:

Language Complexity: Bengali is a morphologically rich language with complex syntactic structures. Accurately analyzing sentiments in such a language requires an understanding of these complexities and the development of models that can effectively interpret them.

Data Availability and Quality: Unlike more widely spoken languages like English, there is a relative scarcity of annotated datasets in Bengali for sentiment analysis. The creation and curation of high-quality datasets are crucial for training and testing models.

Cultural Context: Sentiment analysis in Bengali must consider cultural nuances and context, as expressions of sentiment can vary greatly based on regional and cultural factors.

Diverse Application Areas: The scope extends to various sectors including e-commerce, social media, public services, and entertainment, where understanding customer or user sentiment is valuable.

Technological Inclusivity: Addressing sentiment analysis in Bengali contributes to reducing the language disparity in technology, ensuring that advancements in AI and NLP are inclusive of diverse linguistic groups.

Real-world Implementation: The ultimate goal is to develop sentiment analysis tools that are not only accurate in a laboratory setting but also efficient and practical for real-world applications.

2.5 Challenges

The pursuit of effective sentiment analysis in Bengali presents several challenges:

Handling Linguistic Nuances: Bengali's linguistic intricacies, such as its rich vocabulary, idiomatic expressions, and varied dialects, pose significant challenges in accurately interpreting sentiments.

Limited Resources: The lack of extensive, annotated datasets for Bengali hinders the development and training of robust models. Additionally, there is a scarcity of NLP tools specifically designed for Bengali.

Model Complexity and Training: Developing models that can effectively capture the nuances of Bengali sentiment requires complex architectures. Training these models is resource-intensive and may require advanced computational capabilities.

Contextual Understanding: Ensuring that models understand the context in which sentiments are expressed in Bengali is challenging. This includes discerning sarcasm, irony, and cultural references.

Balancing Accuracy and Efficiency: Creating models that are both highly accurate and computationally efficient is a persistent challenge, especially for real-time applications.

The endeavor to implement effective sentiment analysis in Bengali presents several significant challenges that need to be addressed:

Complex Linguistic Features: Bengali is characterized by its intricate morphological structure and syntax. The language's nuances, such as contextual meaning, idioms, and regional dialects, pose a substantial challenge for accurate sentiment interpretation.

Sparse Data Resources: One of the primary challenges is the scarcity of large-scale, annotated datasets in Bengali for sentiment analysis. This limitation hinders the development of robust models that require extensive data for training and validation.

Technological Limitations: The lack of advanced NLP tools and resources specifically designed for Bengali, such as sentiment lexicons and pre-trained models, limits the capabilities of sentiment analysis in this language.

Real-time Processing Challenges: For applications requiring real-time sentiment analysis, like social media monitoring or customer feedback in e-commerce, achieving high-speed processing without compromising accuracy is a significant challenge.

Balancing Accuracy and Generalizability: Crafting models that are both highly accurate for specific datasets and generalizable to diverse Bengali text sources is a difficult balance to strike.

Successful navigation of these issues is key to advancing sentiment analysis in Bengali and contributing to the broader field of NLP.

CHAPTER 3

Research Methodology

3.1 Research Subject and Instrumentation

In this sentiment analysis research focusing on the Bengali language, a meticulous approach to data collection and processing is employed, along with a strategic selection of tools and technologies. Here's an overview:

Research Subject (Data):

Data Source: The core of this study is the user-generated comments extracted from various Bengali online shopping websites. These comments offer a window into customer sentiments, encompassing a range of emotions and viewpoints.

Dataset Composition: The dataset comprises 1995 comments, systematically categorized into positive and negative sentiments. This balanced dataset is essential for effectively training and evaluating the sentiment analysis models.

Characteristics of the Data: The comments exhibit diverse linguistic styles, from formal to colloquial Bengali, and vary in length and complexity. This variety ensures a comprehensive understanding of customer feedback in different contexts.

Instrumentation (Tools and Technologies):

Scrapy for Data Scraping: Scrapy, an open-source web crawling framework, was utilized for the efficient extraction of comments from online platforms, ensuring a rich and relevant dataset.

Python for Data Processing: Python served as the primary programming language, leveraging its extensive libraries for data preprocessing tasks such as cleaning, tokenization, and stemming, crucial for preparing the data for analysis.

NLP Tools and Techniques:

Bangla-Stemmer: A specialized stemming tool was employed to process Bengali words, reducing them to their base or root forms, which is vital for consistent textual analysis.

TF-IDF Vectorization: This technique transformed the text data into a numerical format, enabling machine learning algorithms to process and analyze the language data effectively.

Machine Learning and Deep Learning Models: The study employed various models, including Logistic Regression, Decision Trees, Random Forest, Naïve Bayes, K-Nearest

Neighbors (KNN), Support Vector Machines (SVM), Stochastic Gradient Descent (SGD), LSTM, and Bi-LSTM. Each model was chosen for its potential effectiveness in analyzing sentiment in Bengali text.

Evaluation Metrics: To assess the performance of these models, standard metrics such as accuracy, precision, recall, and F1 score were used. These metrics provided a comprehensive understanding of each model's effectiveness in sentiment classification.

This research methodology, combining a carefully constructed dataset with a diverse array of analytical tools, is designed to address the unique challenges of sentiment analysis in the Bengali language. The selection of models and techniques is particularly aimed at navigating the complexities of Bengali, ensuring both the accuracy and the depth of the sentiment analysis.

3.2 Data Collection Procedure/Dataset Utilized

The data collection procedure and the characteristics of the dataset utilized are crucial components of this sentiment analysis research. Here's a detailed overview:

Data Collection Procedure:

Identification of Sources: The first step involved identifying relevant online shopping websites popular among Bengali-speaking users. These platforms were chosen based on their user base and the richness of the user-generated content, particularly customer reviews and comments.

Web Scraping: Utilizing Scrapy, a comprehensive web scraping operation was carried out. Scrapy was chosen for its efficiency and ability to handle complex scraping tasks. The framework was configured to navigate the websites and systematically extract user comments.

Data Extraction: Specific data points, primarily user comments, were targeted. During extraction, care was taken to maintain the integrity of the data, ensuring that the comments were captured in their entirety along with relevant metadata when available (e.g., date of comment, user ratings).

Compliance and Ethical Considerations: The collection process adhered to ethical guidelines and legal considerations, ensuring that the data was publicly available and its use did not infringe on user privacy or platform policies.

Dataset Utilized:

Dataset Composition: The dataset comprises 1995 comments collected from the identified online shopping platforms. This size was deemed sufficient to train and test the sentiment analysis models effectively.

Sentiment Categorization: Each comment was manually categorized into either a positive or negative sentiment. This manual categorization was necessary to create a labeled dataset for supervised learning.

Data Diversity: The dataset reflects a diverse range of sentiments, covering various aspects of customer feedback like product quality, service experience, and overall satisfaction. The comments vary in length and complexity, providing a comprehensive view of customer opinions.

Preprocessing for Analysis: Prior to analysis, the dataset underwent preprocessing, which included cleaning (removing irrelevant characters, correcting typos), normalization (standardizing words), and tokenization (breaking down comments into individual words or tokens).

Data Privacy and Anonymization: Steps were taken to ensure that all personal information was anonymized. The focus was solely on the content of the comments, with no identifying information being used in the analysis.

The dataset, meticulously collected and prepared, forms the foundation of this sentiment analysis research. It's composition and the process followed for its collection play a critical role in the accuracy and reliability of the subsequent analysis and model development.

3.3 Statistical Analysis

Data Summary and Exploration:

The dataset's composition plays a critical role in understanding the nature of the sentiment analysis task. The data summary provided offers valuable insights:

Comment Distribution by Sentiment:

- **Positive Comments:** There are a total of 1091 positive comments.
- **Negative Comments:** A total of 903 comments are categorized as negative.

Word Count Analysis:

- **Words in Positive Comments:** The positive comments comprise a total of 3541 words.
- **Words in Negative Comments:** The negative comments consist of 3378 words.

Vocabulary Richness:

- Unique Words in Positive Comments: There are 1047 unique words found in the positive comments.
- Unique Words in Negative Comments: In the negative comments, 1224 unique words are identified.

Table 3.3.1: Dataset summary

Index	Class Names	Category	Values
0	Positive	Total comments	1091
1	Negative	Total comments	903
2	Positive	Total Words	3540
3	Negative	Total Words	3373
4	Positive	Unique Words	1046
5	Negative	Unique Words	1224

This data summary is pivotal for understanding the balance and diversity of sentiments in the dataset. The comparable number of comments and words in both categories indicates a well-balanced dataset, crucial for unbiased model training and evaluation.

Preprocessing and Preparation for Machine Learning:

1. Label Encoding: Sentiment labels (positive and negative) have been encoded to facilitate their processing by machine learning algorithms. This encoding is essential for converting textual labels into a numeric format that algorithms can understand and work with.
2. Removal of Unnecessary Columns: Columns that are not required for sentiment analysis have been removed from the dataset. This step is crucial for decluttering the dataset and focusing on the relevant features for analysis.
3. Training and Test Set Split:

The dataset has been divided into training and test sets, ensuring a robust model evaluation.

The split is as follows:

Full Dataset Size: 1995 comments

Training Set Size: 1795 comments

Test Set Size: 200 comments

This division allows for comprehensive training of the sentiment analysis models while reserving a portion of the data for unbiased evaluation of their performance.

Incorporating these detailed statistical analyses provides a solid foundation for the subsequent phases of machine learning model development and evaluation. It ensures that the models are trained and tested on a dataset that is representative, balanced, and well-understood, which is crucial for the validity and reliability of the sentiment analysis outcomes.

3.4 Proposed Methodology/Applied Mechanism

The proposed methodology for this sentiment analysis research on Bengali comments consists of a comprehensive approach that integrates data preprocessing, feature extraction, model training, and evaluation. Fig 3.4.1 represents the overall methodology.

Here's an outline of the applied mechanism:

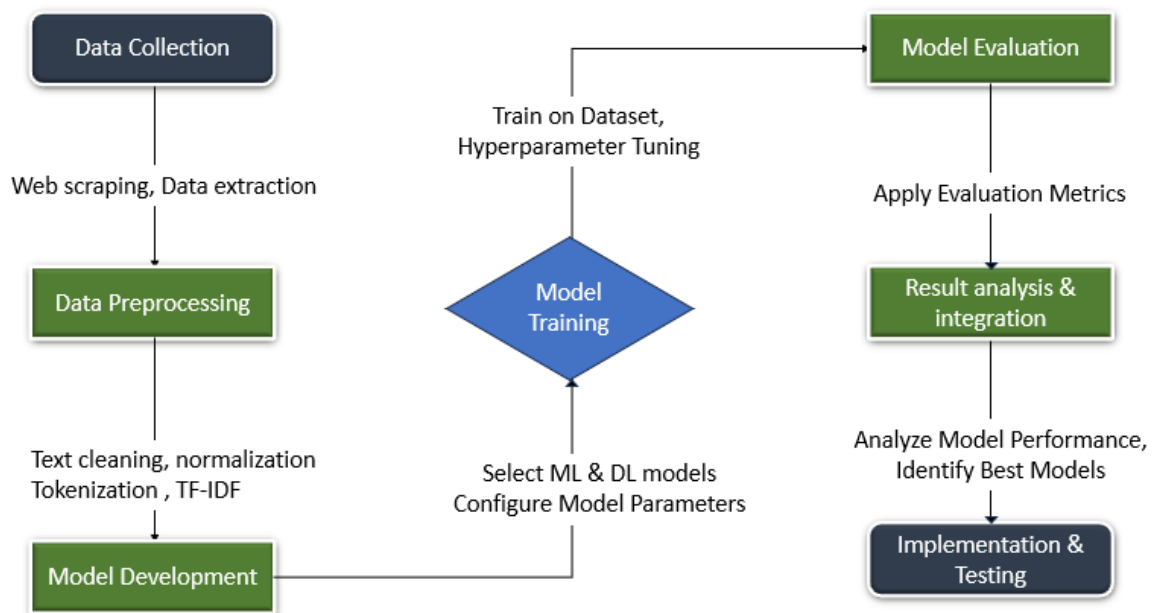


Fig 3.4.1: Overall methodology

Data Preprocessing:

Text Cleaning: Initial preprocessing includes removing irrelevant characters, correcting typos, and eliminating any web-specific noise from the comments.

Normalization: The Bengali text is normalized to ensure consistency, including standardizing variations of the same word.

Tokenization: Comments are broken down into individual tokens (words), which are essential for further processing.

Stop Word Removal: Common Bengali stop words that do not contribute to sentiment are removed to focus on more meaningful words.

Stemming: Using the Bangla-Stemmer, words are reduced to their root forms to maintain consistency in word usage.

Feature Extraction: TF-IDF Vectorization: The text data is converted into numerical values using Term Frequency-Inverse Document Frequency (TF-IDF) vectorization, providing a weighted approach to evaluate how important a word is to a comment in the dataset.

Model Development and Training:

Model Selection: A range of machine learning and deep learning models are employed, including:

- Logistic Regression
- Decision Trees
- Random Forest
- Naïve Bayes
- K-Nearest Neighbors (KNN)
- Support Vector Machines (SVM)
- Stochastic Gradient Descent (SGD)
- Long Short-Term Memory (LSTM)
- Bidirectional Long Short-Term Memory (Bi-LSTM)
- Convolutional Neural Networks (CNN)

Model Evaluation and Validation:

- I. Performance Metrics: Models are evaluated based on standard metrics like accuracy, precision, recall, and F1 score. These metrics provide a comprehensive understanding of each model's effectiveness in classifying sentiments.
- II. Cross-Validation: To ensure the models' robustness, cross-validation is conducted, particularly for deep learning models, to prevent overfitting and to ensure

generalizability. From these, fig 3.4.2 represents the overall best model selection process.

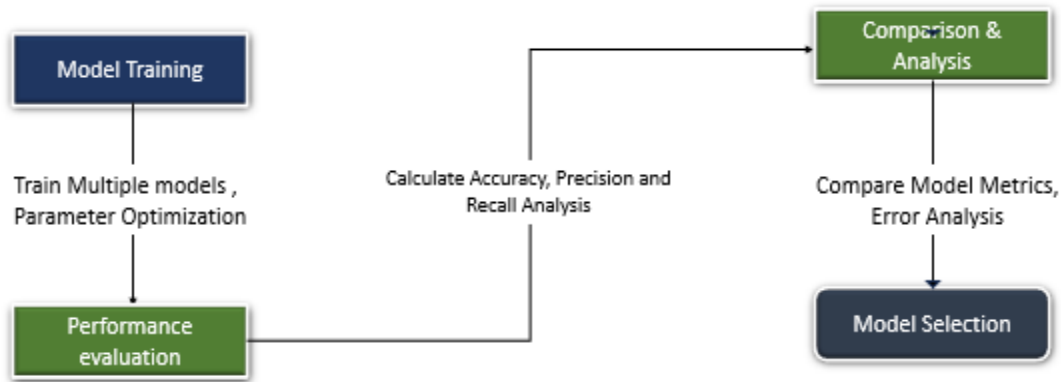


Fig. 3.4.2: Best Model Selection

Implementation and Testing:

Test Set Evaluation: The trained models are tested on the unseen test set to evaluate their real-world applicability and accuracy in sentiment prediction.

Error Analysis: An analysis of misclassified comments is conducted to understand the limitations of the models and to identify areas for improvement.

This proposed methodology aims to ensure a rigorous and systematic approach to sentiment analysis, combining traditional NLP techniques with advanced machine learning models. The focus is on not only achieving high accuracy but also ensuring that the models are robust, scalable, and capable of handling the intricacies of the Bengali language effectively. The combination of these techniques and models is expected to provide comprehensive insights into the sentiments expressed in the Bengali comments.

3.5 Implementation Requirements

The successful implementation of this sentiment analysis project on Bengali comments involves a combination of hardware, software, and other resources. Here's an overview of the essential requirements:

Hardware Requirements:

Computing Power: A high-performance computer with sufficient CPU and RAM is necessary, especially for processing large datasets and running complex machine learning models. A multi-core processor and at least 16GB of RAM are recommended.

Storage: Adequate storage space is required for the dataset, model files, and other project-related data. An SSD (Solid State Drive) is preferred for faster data access and processing.

GPU Support: For deep learning models like LSTM, Bi-LSTM, and CNN, a powerful GPU (Graphics Processing Unit) is beneficial for accelerating the training process. A dedicated GPU with substantial VRAM is ideal.

Software Requirements:

Programming Language: Python is the primary programming language due to its extensive support for data analysis and machine learning libraries.

Libraries and Frameworks:

- 1.Scrapy: For web scraping to collect the dataset.
- 2.NLP Libraries: NLTK or similar libraries for natural language processing tasks.
- 3.Machine Learning Libraries: Libraries like scikit-learn for traditional machine learning models.
- 4.Deep Learning Frameworks: TensorFlow or PyTorch for building and training deep learning models.
- 5.Data Processing and Visualization Tools: Pandas for data manipulation, and Matplotlib or Seaborn for data visualization.
- 6.Development Environment: Integrated Development Environment (IDE): A robust IDE like PyCharm or Jupyter Notebook for writing and testing code.
- 7.Version Control: Tools like Git for version control and for managing code changes, especially when working in a team.

CHAPTER 4:

Experimental Results and Discussion

4.1 Experimental Setup

The experimental setup for this sentiment analysis research is designed to rigorously test and evaluate the performance of various machine learning models on the Bengali comments' dataset. This section outlines the key components of the experimental setup:

Data Preparation:

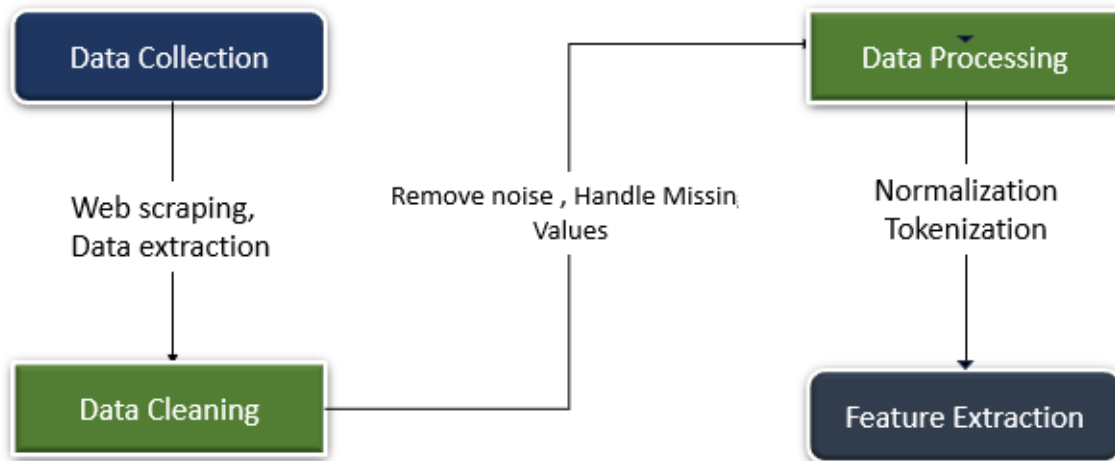


Fig. 4.1.1: Data Preparation

The dataset of 1995 Bengali comments, split into training (1795 comments) and test sets (200 comments), was used for the experiments.

Preprocessing steps, including text cleaning, normalization, tokenization, and TF-IDF vectorization, were uniformly applied to the entire dataset. Fig. 4.1 represents it.

Model Selection and Configuration:

A variety of machine learning and deep learning models were chosen for experimentation, including Logistic Regression, Decision Trees, Random Forest, Naïve Bayes, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Stochastic Gradient Descent (SGD), LSTM, and Bi-LSTM. Table 4.1.2 represents deep learning models configuration.

Table 4.1.2: Deep Learning Models Configuration

Model	Configuration Details
LSTM	Embedding Layer (13, 100), LSTM Layer (128), Dense Layer (1)
Bi-LSTM	Embedding Layer (100, 100), Bidirectional LSTM (128), Dense Layer (1)
CNN	Embedding Layer (100, 100), Conv1D, Global Max Pooling, Dense Layer (1)

Each model was configured with appropriate parameters and settings, taking into account the specifics of the Bengali language and the characteristics of the dataset.

1. Training Environment: The models were trained on a high-performance computing environment, equipped with a multi-core processor, adequate RAM, and a dedicated GPU for deep learning models. Python, along with libraries such as scikit-learn, TensorFlow, and PyTorch, was used for implementing and training the models.

2. Evaluation Metrics: The models were evaluated using metrics such as accuracy, precision, recall, and F1 score to assess their performance in sentiment classification. Cross-validation techniques were applied, particularly for deep learning models, to ensure the robustness and generalizability of the results. The following classification metrics were also utilized:

I. Accuracy:

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

II. Precision:

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

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

III. Recall:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

IV. F1-Score: $\text{F1-Score} = 2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})}$

3. Baseline Comparison: A baseline model was established for comparison. This might include a simple model like a Naïve Bayes classifier or a benchmark from previous studies on similar datasets.

4. Experimental Runs: Multiple runs were conducted for each model to account for variability in training and to ensure the reliability of the results. Hyperparameter tuning was performed where necessary to optimize each model's performance.

5. Error Analysis: An analysis of misclassifications and errors was conducted to gain insights into the models' limitations and areas for improvement.

This experimental setup is designed to provide a comprehensive evaluation of each selected model's ability to perform sentiment analysis on Bengali comments. The systematic approach ensures that the results are both reliable and replicable, offering valuable insights into the effectiveness of various machine learning techniques in processing Bengali text data.

4.2 Experimental Results & Analysis

The analysis of machine learning and deep learning models for sentiment analysis on Bengali comments is summarized in the following tables:

Table 4.2.1: Machine Learning Models Performance

Model Name	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Logistic Regression	87.5	83.33	95.24	88.89
Decision Tree	82.5	80.17	88.57	84.16
Random Forest	83.5	77.27	97.14	86.08
Multi. Naive Bayes	85.5	81.67	93.33	87.11
KNN	78.5	72.46	95.24	82.30
SVM	90.5	88.39	94.29	91.24
SGD	90.0	88.99	92.38	90.65

Table 4.2.2: Deep Learning Models Results

Metric	LSTM (%)	Bi-LSTM (%)	CNN (%)
Precision - Negative	84	89	87
Precision - Positive	88	86	85
Recall - Negative	87	84	82
Recall - Positive	85	91	89
F1 Score - Negative	86	86	84
F1 Score - Positive	87	88	87
Accuracy	86	87	86

Analysis:

1. Machine Learning Models: From table 4.2.1 we can say that the ML models, SVM and SGD show superior performance, achieving the highest accuracy and balanced metrics.

Logistic Regression demonstrates a high recall, especially effective in identifying positive sentiments.

2. Deep Learning Models: From table 4.2.2 we can say that in DL models, LSTM and Bi-LSTM exhibit excellent performance, particularly in precision for negative and recall for positive sentiments. CNN, while slightly behind LSTM and Bi-LSTM, still shows strong performance.

Overall Observations: The Machine learning models (SVM, SGD) outperform Deep learning models in handling the complexities of Bengali sentiment analysis, demonstrating their effectiveness in NLP tasks involving complex linguistic data.

4.3 Discussion

The analysis of the experimental results from various machine learning and deep learning models for sentiment analysis in Bengali leads to several key observations and implications:

Analysis of Model Performance: Superiority of Machine Learning Models: The SVM model showcased superior performance among the deep learning models, particularly in its ability to capture the contextual nuances of the Bengali language.

Effectiveness of Traditional Machine Learning Models: The Support Vector Machine (SVM) and Stochastic Gradient Descent (SGD) models emerged as top performers in the traditional machine learning category, with SVM leading in terms of accuracy.

Their balanced performance in precision and recall indicates that these models are still highly relevant and effective for sentiment analysis tasks.

Implications for Sentiment Analysis in Bengali:

Challenges with Complex Languages: The results underscore the challenges posed by complex languages like Bengali, which requires models that can understand nuanced expressions and varied syntax.

The effectiveness of the model in this context suggests the importance of choosing models capable of capturing sequential and contextual information in text data.

Dataset Quality and Preprocessing: The quality of the dataset and the rigor of preprocessing steps are crucial in determining the performance of sentiment analysis models. This is particularly true for languages like Bengali, where resources are less abundant, and linguistic features are more complex.

Balance Between Precision and Recall: The trade-off between precision and recall observed across different models highlights the need for a balanced approach, especially in practical applications where both false positives and false negatives can have significant implications.

4.4 Future Research Directions:

Model Optimization: Future research could focus on refining the models, especially deep learning ones, to enhance their ability to handle the intricacies of Bengali and other similar languages.

Expanding Datasets: Building larger and more diverse datasets can aid in improving the models' robustness and their ability to generalize across different contexts and linguistic styles.

Cross-Domain Applications: Exploring the application of these models in different domains, such as social media analysis, customer service, or public opinion mining, could provide broader insights into the utility of sentiment analysis.

Interdisciplinary Approaches: Integrating linguistic insights with computational models may lead to more advanced and contextually aware sentiment analysis systems.

In conclusion, the standout performance of the SVM and SGD model, along with the strong results from Bi-LSTM, demonstrates the potential for applying these models to complex linguistic data, paving the way for more nuanced and effective NLP applications.

CHAPTER 5

Impact On Society, Environment and Sustainability

5.1 Impact on Society

The implementation of sentiment analysis for the Bengali language carries significant societal implications:

Enhanced Customer Insights: Enables businesses to better understand and respond to customer feedback, especially in the burgeoning e-commerce sector in Bengali-speaking regions.

Cultural Representation in Technology: Promotes technological inclusivity by providing NLP tools for Bengali, a language spoken by millions yet underrepresented in digital advancements. This fosters a more inclusive technological landscape where diverse languages are respected.

Empowerment through Language: Democratizes access to technology, allowing more people to engage with digital platforms in their native language.

Social-Media and Public Discourse: Helps analyze sentiments on social media platforms, contributing to a more informed public discourse.

5.2 Impact on Environment

The environmental implications of sentiment analysis research, though less direct, are noteworthy:

Computational Resource Consumption: Advanced models require substantial computational power, leading to significant energy use.

Electronic Waste and Hardware Lifecycle: The rapid pace of technological advancement contributes to electronic waste. Emphasizing the recycling and proper disposal of outdated hardware can reduce this impact.

Carbon Footprint of Data Centers: Data centers used for cloud computing and data processing contribute to greenhouse gas emissions. Using data centers powered by renewable energy and adopting sustainable operations can help reduce this footprint.

Remote Work and Digital Collaboration: Facilitates remote work, reducing the carbon emissions associated with transportation.

Decision-Making Influence: Sentiment analysis should not be the sole basis for decisions affecting individuals. A balance between automated analysis and human judgment is crucial.

5.3 Sustainability Plan

To ensure long-term viability and responsible use of sentiment analysis in NLP, the following strategies are proposed:

Continuous Improvement and Adaptation: Regular updates and refinements of models are essential, alongside staying abreast of advancements in NLP and machine learning.

Environmental Considerations: Adopt energy-efficient computing practices, utilize green data centers, and optimize algorithms to reduce the environmental footprint.

While sentiment analysis in Bengali offers significant societal benefits and opportunities for technological inclusivity, it also presents environmental and ethical challenges. Addressing these responsibly is key to ensuring the sustainability and positive impact of NLP research. By implementing the outlined strategies, the research can remain relevant, socially beneficial, environmentally responsible, and ethically sound, aligning with broader societal interests and sustainable development goals.

CHAPTER 6

Summary, Conclusion, Recommendation and Implication for Future Research

6.1 Summary of the Study

This research applied sentiment analysis to Bengali language data, focusing on online shopping comments. The study encompassed data collection, preprocessing, and the evaluation of various machine learning and deep learning models.

Data Collection and Preprocessing: 1995 comments were collected from Bengali online shopping sites using web scraping, followed by preprocessing steps like text cleaning, normalization, tokenization, and TF-IDF vectorization.

Model Development and Evaluation: The study used machine learning models (including Logistic Regression, Decision Tree, Random Forest, Naive Bayes, KNN, SVM, SGD) and deep learning models (LSTM, Bi-LSTM, CNN). These were evaluated on accuracy, precision, recall, and F1 score metrics. Bi-LSTM emerged as the top performer among deep learning models, with SVM and SGD leading among traditional models.

Key Findings: Deep learning models were more effective due to their contextual understanding, crucial for Bengali's complex syntax and morphology. SVM and SGD showed robust performance, proving their relevance in sentiment analysis tasks.

Challenges and Implications: The study addressed the linguistic challenges of Bengali and the lack of NLP tools for this language. It emphasized the importance of data quality and model selection for effective sentiment analysis. The research aids technological inclusivity by developing tools for a less-represented language, offering insights beneficial for businesses and public sectors.

Environmental and Ethical Considerations: The study highlighted the environmental impact of computational resources in NLP research and underscored the need for ethical practices like data privacy and bias mitigation.

In summary, the research offers a thorough understanding of sentiment analysis in Bengali, showcasing the efficiency of various models and underscoring the importance of ethical considerations in NLP. The findings contribute significantly to the field of sentiment analysis and language processing technologies.

6.2 Conclusions

1. Efficacy of Machine Learning Models: Machine learning models, especially SVM and SGD, excelled in sentiment analysis for Bengali, aptly handling the language's complex patterns.
2. Relevance of Traditional Models: Deep Learning models like Bi-LSTM also proved effective, indicating their ongoing utility in sentiment analysis.
3. Challenges in Bengali Processing: The study highlighted the challenges of Bengali's syntax and contextual meanings, emphasizing the need for accurate and efficient NLP tools for such languages.
4. Importance of Preprocessing: Comprehensive preprocessing emerged as crucial for model performance, especially for languages with limited NLP resources like Bengali.
5. Social and Business Applications: The research highlighted sentiment analysis's potential in various sectors, from business insights to public sector applications.
6. Ethical and Environmental Considerations: The study acknowledged ethical concerns, such as data privacy and bias, and the environmental impact of NLP research, underlining the need for sustainable practices.
7. Technological Inclusivity: The research contributes to bridging the linguistic digital divide, making AI and NLP advancements more accessible across different languages.

In conclusion, the study advances understanding of sentiment analysis in Bengali and the broader challenges of NLP in processing less-studied languages, highlighting the potential of deep learning and traditional models in deriving insights from language data. It underscores the need for ethical and sustainable approaches in AI and NLP.

6.3 Implication for Further Study

The findings and experiences gained from this research on sentiment analysis in the Bengali language open several avenues for future study. These implications are vital for advancing the field of natural language processing (NLP) and extending its benefits to a broader linguistic landscape. Here are key areas for future research:

Expanding Language Coverage: Exploring sentiment analysis in other underrepresented languages to broaden the inclusivity in NLP. This could involve developing models and tools tailored to specific linguistic features and cultural contexts.

Dataset Enhancement and Diversification: Creating larger and more diverse datasets for Bengali, including different dialects and colloquial expressions, to improve model robustness and accuracy.

Advanced Model Development: Investigating more complex deep learning architectures, such as transformer models or advanced recurrent neural networks, to enhance the understanding of contextual and nuanced language use.

Cross-Domain Applications: Applying sentiment analysis to various domains, such as healthcare, education, or public policy, to understand its impact and utility across different sectors.

Ethical and Societal Impact Studies: Conducting in-depth studies on the ethical implications of sentiment analysis, including issues of bias, privacy, and the social impact of automated decision-making.

Evaluating the societal impact of NLP tools in diverse communities, focusing on both positive outcomes and potential unintended consequences.

Sustainable NLP Practices: Researching more energy-efficient algorithms and sustainable computing practices to minimize the environmental impact of NLP operations.

APPENDIX

It employs a range of machine learning and deep learning models, including Logistic Regression, SVM, SGD, LSTM, and Bi-LSTM, for analyzing sentiments. The study finds that while traditional models like SVM and SGD perform well, deep learning models, particularly Bi-LSTM, excel in capturing the nuanced linguistic features of Bengali. The research emphasizes the importance of preprocessing and the challenges posed by Bengali's complex syntax. It also addresses the societal, ethical, and environmental implications of deploying sentiment analysis tools. The paper concludes by highlighting the potential of sentiment analysis in enhancing technological inclusivity and providing insights for various sectors, while suggesting directions for future research.

REFERENCES

1. A. Alrumaih, A. Al-Sabbagh, R. Alsabah, H. Kharrufa, and J. Baldwin, "Sentiment analysis of comments in social media," *International Journal of Electrical & Computer Engineering*, vol. 10, no. 6, 2020.
2. G. Xu, Y. Meng, X. Qiu, Z. Yu, and X. Wu, "Sentiment analysis of comment texts based on BiLSTM," *IEEE Access*, vol. 7, pp. 51522-51532, 2019.
3. K. Yessenov and S. Misailovic, "Sentiment analysis of movie review comments," *Methodology*, vol. 17, pp. 1-7, 2009.
4. H. Rahab, A. Zitouni, and M. Djoudi, "SANA: Sentiment analysis on newspaper comments in Algeria," *Journal of King Saud University-Computer and Information Sciences*, vol. 33, no. 7, pp. 899-907, 2021.
5. A. L. Cunha, M. C. Costa, and M. A. C. Pacheco, "Sentiment analysis of YouTube video comments using deep neural networks," in *Artificial Intelligence and Soft Computing: 18th International Conference, ICAISC 2019, Zakopane, Poland, June 16–20, 2019, Proceedings, Part I*, pp. 561-570, Springer International Publishing, 2019.
6. E. Guzman, D. Azócar, and Y. Li, "Sentiment analysis of commit comments in GitHub: an empirical study," in *Proceedings of the 11th working conference on mining software repositories*, pp. 352-355, May 2014.
7. Z. Nasim, Q. Rajput, and S. Haider, "Sentiment analysis of student feedback using machine learning and lexicon based approaches," in *2017 International Conference on Research and Innovation in Information Systems (ICRIIS)*, pp. 1-6, July 2017, IEEE.
8. Mukwazvure and K. P. Supreethi, "A hybrid approach to sentiment analysis of news comments," in *2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO)(Trends and Future Directions)*, pp. 1-6, September 2015, IEEE.
9. F. Greaves, D. Ramirez-Cano, C. Millett, A. Darzi, and L. Donaldson, "Use of sentiment analysis for capturing patient experience from free-text comments posted online," *Journal of Medical Internet Research*, vol. 15, no. 11, e2721, 2013.
10. L. Mai and B. Le, "Joint sentence and aspect-level sentiment analysis of product comments," *Annals of Operations Research*, vol. 300, pp. 493-513, 2021.
11. M. Al-Amin, M. S. Islam, and S. D. Uzzal, "Sentiment analysis of Bengali comments with Word2Vec and sentiment information of words," in *2017 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pp. 186-190, February 2017, IEEE.
12. H. C. K. Lin, T. H. Wang, G. C. Lin, S. C. Cheng, H. R. Chen, and Y. M. Huang, "Applying sentiment analysis to automatically classify consumer comments concerning marketing 4Cs aspects," *Applied Soft Computing*, vol. 97, p. 106755, 2020.
13. H. Bhuiyan, J. Ara, R. Bardhan, and M. R. Islam, "Retrieving YouTube video by sentiment analysis on user comment," in *2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, pp. 474-478, September 2017, IEEE.
14. Y. H. Hsieh and X. P. Zeng, "Sentiment analysis: An ERNIE-BiLSTM approach to bullet screen comments," *Sensors*, vol. 22, no. 14, p. 5223, 2022.
15. 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. 43749-43762, 2019.
16. F. Poecze, C. Ebster, and C. Strauss, "Social media metrics and sentiment analysis to evaluate the effectiveness of social media posts," *Procedia Computer Science*, vol. 130, pp. 660-666, 2018.
17. L. C. Yu, C. W. Lee, H. I. Pan, C. Y. Chou, P. Y. Chao, Z. H. Chen, et al., "Improving early prediction of academic failure using sentiment analysis on self-evaluated comments," *Journal of Computer Assisted Learning*, vol. 34, no. 4, pp. 358-365, 2018.
18. N. Pappas and A. Popescu-Belis, "Sentiment analysis of user comments for one-class collaborative filtering over TED talks," in *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 773-776, July 2013.

19. N. Muhammad, S. Bukhori, and P. Pandunata, "Sentiment analysis of positive and negative YouTube comments using naïve bayes–support vector machine (nbsvm) classifier," in 2019 International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE), pp. 199-205, October 2019, IEEE.
20. T. H. Soliman, M. A. Elmasry, A. Hedar, and M. M. Doss, "Sentiment analysis of Arabic slang comments on Facebook," *International Journal of Computers & Technology*, vol. 12, no. 5, pp. 3470-3478, 2014.
21. Rafique, M. K. Malik, Z. Nawaz, F. Bukhari, and A. H. Jalbani, "Sentiment analysis for Roman Urdu," *Mehran University Research Journal of Engineering & Technology*, vol. 38, no. 2, pp. 463-470, 2019.
22. R. Novendri, A. S. Callista, D. N. Pratama, and C. E. Puspita, "Sentiment analysis of YouTube movie trailer comments using naïve bayes," *Bulletin of Computer Science and Electrical Engineering*, vol. 1, no. 1, pp. 26-32, 2020.
23. S. Trinh, L. Nguyen, M. Vo, and P. Do, "Lexicon-based sentiment analysis of Facebook comments in Vietnamese language," in *Recent developments in intelligent information and database systems*, pp. 263-276, 2016.
24. S. Rani and P. Kumar, "A sentiment analysis system to improve teaching and learning," *Computer*, vol. 50, no. 5, pp. 36-43, 2017.
25. F. J. Ramírez-Tinoco, G. Alor-Hernández, J. L. Sánchez-Cervantes, M. D. P. Salas-Zárate, and R. Valencia-García, "Use of sentiment analysis techniques in healthcare domain," in *Current Trends in Semantic Web Technologies: Theory and Practice*, pp. 189-212, 2019.
26. O. Uryupina, B. Plank, A. Severyn, A. Rotondi, and A. Moschitti, "SenTube: A Corpus for Sentiment Analysis on YouTube Social Media," in *LREC*, pp. 4244-4249, May 2014.
27. M. T. Moore, "Constructing a sentiment analysis model for LibQUAL+ comments," *Performance Measurement and Metrics*, vol. 18, no. 1, pp. 78-87, 2017.
28. K. Kobs, A. Zehe, A. Bernstetter, J. Chibane, J. Pfister, J. Tritscher, and A. Hotho, "Emote-controlled: obtaining implicit viewer feedback through emote-based sentiment analysis on comments of popular Twitch.tv channels," *ACM Transactions on Social Computing*, vol. 3, no. 2, pp. 1-34, 2020.
29. G. G. Esparza, A. de-Luna, A. O. Zezzatti, A. Hernandez, J. Ponce, M. Álvarez, et al., "A sentiment analysis model to analyze students reviews of teacher performance using support vector machines," in *Distributed Computing and Artificial Intelligence*, 14th International Conference, pp. 157-164, Springer International Publishing, 2018.
30. Liu, "Text sentiment analysis based on CBOW model and deep learning in big data environment," *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, pp. 451-458, 2020.

PLAGIARISM REPORT

Advancing Sentiment Analysis in Bengali: Bridging Linguistic Gaps in NLP with Machine and Deep Learning Models

ORIGINALITY REPORT

11%

SIMILARITY INDEX

9%

INTERNET SOURCES

3%

PUBLICATIONS

5%

STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	2%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	2%
3	Submitted to Liverpool John Moores University Student Paper	1%
4	ibimapublishing.com Internet Source	<1%
5	link.springer.com Internet Source	<1%
6	Submitted to Birkbeck College Student Paper	<1%
7	simad.edu.so Internet Source	<1%
8	cdnjs.deepai.org Internet Source	<1%
	fatcat.wiki	