BANGLADESHI INDIGENOUS LANGUAGE CLASSIFICATION FROM TEXT ANALYSIS

BY

ASRAFUL ISLAM

ID: 201-35-608

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

Supervised By
Dr.Imran Mahmud
Associate Professor and Head
Department of SWE
Daffodil International University

Co-Supervised By

Senior Lecturer
Department of SWE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JULY 2024

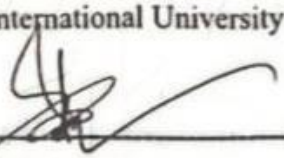
APPROVAL

This thesis titled on **"Bangladeshi Indigenous Language Classification From Text Analysis"**, submitted by **ASRAFUL ISLAM (ID: 201-35-608)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

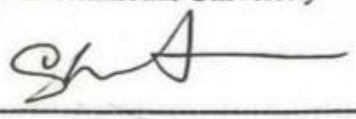
Chairman

Dr. Imran Mahmud
Associate Professor & Head
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University



Internal Examiner 1

Md. Maruf Hasan
Associate Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University



Internal Examiner 2

Md. Shohel Arman
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

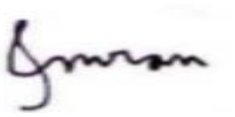
External Examiner

Dr. Md. Sazzadur Rahman
Professor
Institute of Information Technology
Jahangirnagar University

Declaration

We, therefore, declare that this undertaking has been finished by us under the supervision of Dr. Imran Mahmud Associate Professor and Head Department of SWE, Daffodil International University. We further declare that neither and application or an educational grant has been made anywhere for this project or any part of it.

Supervised By



Dr. Imran Mahmud

Associate Professor and Head

Department of Software Engineering

Daffodil International University

Submitted by

Ashraful Islam

ID: 201-35-608

Department of Software Engineering

Daffodil International University

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes it possible to complete the final year project successfully.

We are really grateful and wish our profound indebtedness to Dr.Imran Mahmud Associate Professor and Head, Department of SWE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "Machine Learning" to carry out this project. Her 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.

We would like to express our heartiest gratitude to **Dr.Imran Mahmud** Associate Professor and Head, Department of SWE, for his kind help to finish our project and also to other faculty members and the staff of SWE department of Daffodil International University.

We would like to thank all of our classmates at Daffodil International University who participated in the discussion while also attending the course.

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

ABSTRACT

Bangladesh is a country having cultural and historical linguistic richness and the current pattern of language reflects the sociological characteristic of the society. Of the numerous languages spoken in Bangladesh and around the world, the indigenous languages of the country occupy a special place here as they provide an insight into the distinct culture and identity of various ethnic groups. These languages themselves, which often have their roots in traditions as old as the prehistoric periods, contain details of otherwise traditions, myths and histories of local communities.

The paper names “Bangladeshi Indigenous Language Classification from Text Analysis” addresses the task of classification of indigenous languages in Bangladesh, with the target attributes being ‘Dhaka’ ‘Chattogram’ and ‘Mymensingh’ This work uses an exciting dataset collected through Google Form and employs BNB, SVM, LR Random Forest, Bi-LSTM and CNN to classify the indigenous language A report of the study has shown that CNN is the 66%. This induces more importance and relevance of mature deep learning techniques for proper classification of native languages which are found in the linguistic complex of Bangladesh.

Keywords: Bangladesh, Indigenous Languages, Cultural Diversity, Linguistic Heritage, Text Analysis, Language Preservation, Community Identity, Linguistic Richness

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	i
Declaration	ii
Acknowledgements	iii
Abstract	iv
CHAPTER	
CHAPTER 1: INTRODUCTION	**1-6**
1.1 Introduction	1-2
1.2 Motivation	2
1.3 Rationale of the Study	2-3
1.4 Research Questions	3
1.6 Project Management and Finance	4
1.7 Report Layout	5-6

©Daffodil International University

CHAPTER 2: BACKGROUND **7-17**

2.1 Preliminaries 7

2.2 Related Works 7-15

2.3 Comparative Analysis and Summary 15

2.4 Scope of the Problem 15-16

2.5 Challenges 16-17

CHAPTER 3: RESEARCH METHODOLOGY **18-26**

3.1 Research Subject and Instrumentation 18

3.2 Data Collection Procedure 18-19

3.3 Statistical Analysis 19-20

3.4 Proposed Methodology 20-26

3.5 Implementation Requirements 26

CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION **27-36**

4.1 Experimental Setup 27

©Daffodil International University

4.2 Experimental Results & Analysis 27-29

4.3 Discussion 29-36

CHAPTER 5: IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY **37-38**

5.1 Impact on Society 37

5.2 Impact on Environment 37-38

5.3 Ethical Aspects 38

CHAPTER 6: SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH **39-41**

6.1 Summary of the Study 39

6.2 Conclusions 39-40

6.3 Implication for Further Study 40-41

©Daffodil International University

LIST OF FIGURES

FIGURES	PAGE NO
Figure 3.1: Dataset Distribution	19
Figure 3.2: Methodology Flowchart	21
Figure 3.3: Logistic Regression architecture	23
Figure 3.4: Random Forest model architecture	24
Figure 3.5: Bi-LSTM model architecture	25
Figure 3.6: CNN architecture	26
Figure 4.1 : Accuracy Comparison of Deep Learning And Machine Learning Models	30
Figure 4.2 : Confusion Matrix BNB	31
Figure 4.3 : Confusion Matrix SVM	32
Figure 4.4 : Confusion Matrix LR	34
Figure 4.5 : ROC-Curve	35
Figure 4.6 : Confusion Matrix RF	36

©Daffodil International University

LIST OF TABLES

TABLES	PAGE NO
Table 1.1: PROJECT MANAGEMENT TABLE	4
TABLE 2.1: COMPARISON BETWEEN EXISTING & PROPOSED MODEL	14
TABLE 4.1. PERFORMANCE EVALUATION	29
TABLE 4.1. PERFORMANCE EVALUATION (BNB)	31
Table 4.3. Performance Evaluation(SVM)	32
TABLE 4.3. PERFORMANCE EVALUATION (LR)	33
Table 4.5. Performance Evaluation(RF)	35

©Daffodil International University

CHAPTER 1

INTRODUCTION

1.1 Introduction

Bangladesh is a country of Multiple Languages displays of social cultural Linguistic diversification possessing a good number of indigenous Languages. These languages need to be preserved and acknowledged, mostly in areas like Dhaka, Chittagong, Mymensingh, etc., in order to respect and remember the valuable part of the country's language culture. Therefore, the project "Bangladesh Indigenous Language Classification from Text Analysis," revolves around a more progressive portion of natural language processing techniques and trusted machine learning models valued to fulfill the difficult task of identifying and comprehending the given linguistic spectrum. [1]

The bible corpus of this project is a painstaking amassment of data through this Google Form created for soliciting texts in the native languages of the contributors. The inclusion of submissions from Dhaka, Chittagong, and Mymensingh areas will try to encapsulate such a justification of these subtle differences and peculiarities due to the different lingual tints. The 4 set target ABSTRACT ATTRIBUTE_CLS_INDEX_LABEL has geographical roots in Dhaka, Chittagong and Mymensingh divisions. translate the text to English as it is different from rest of the languages. [2]

These procedures were developed with the purpose of assisting users in avoiding the convoluted issues involved in language classification. Every procedure and stage, starting from data prepossessing and into the NLP data cleaning phase, model selection and proceeding to the model check phase, is designed to ensure maximum precision and coverage. Notably, simple models such as the Bernoulli Naive Bayes, Logistic Regression, and more complex models as RandomForest, BiLstM, and CNN are compared in terms of their performance. [3]

It is not just about classification but the study of what has to be saved, that has been coded in the language of cultures. This effort aids to the general understanding of the native languages in Bangladesh by successfully enabling the enhance of a strong model, strain the

significance and rarity of the languages that commits the nation. As we establish this academic project of investigating and categorizing the vernaculars of the specific areas and cities of Bangladesh, our aim is to unveil the dialectal web that connects Dhaka, Chittagong, Mymensingh people in the spirit of embracing plurality and cultural heritage. [4]

1.2 Motivation

The goal of the "Bangladeshi Indigenous Language Classification from Text Analysis" project stemmed from a recognition of the importance of conserving linguistic cultural heritage in Bangladesh. With a diversity of first nations languages which exists in Bangladesh especially in Dhaka Chittagong Mymensingh this movement aims to categorize and interpret the languages through the help of modern day technologies. The project is part of a bigger goal to support initiatives focused on the preservation and appreciation of different forms of culture through artificial intelligence using techniques such as natural language processing and machine learning. The data was collected from a Google Form and different aspects of the linguistic landscape are depicted in the results The method used for classification has been designed in detail to ensure appropriate classes are formed. The motivation is not only technical awareness accumulated through research but a passion to understand the complex linguistic history of Bangladesh and give the due appreciation to the cultural cosmography inscribed in these languages.

1.3 Rationale of the Study

The idea for the 'Bangladeshi Indigenous Language Classification from Text Analysis' is derived from the proposition that the linguistic diversity of Bangladesh requires attention, particularly the languages of Dhaka,Chittagong , Mymensingh regions. It is imperative that when one talks about indigenous languages that they have an understanding of them and be able to identify them because language is wealth that is a repository of a nation's history that should be preserved. The study acknowledges the losses resulting from globalization and modernization in these languages and therefore points to a need for technological assistances. This work seeks to build a strong classification of model which will be able to

distinguish between languages originating from different location in the use of modern NLP and ML approaches.

The results of this research will not be beneficial only for the language classification, but also in terms of culture served through the comprehension of specific tendencies of Dhaka, Chittagong, Mymensingh . This study can be deemed as a preventive measure to begin appreciation of the Burmese language as well as different cognition and development of the diverse populations in Bangladesh and nurtures the cultural integration and understanding.

1.4 Research Question

i. Which methods like Bernoulli Naive Bayes and logistic regression can help in the classification of indigenous languages of Bangladesh with the text samples gathered by Google Form?.

ii. Which of the preprocessing methods is applicable for cleaning and optimizing text data for its correct categorization into languages, given the heterogeneity of the GPrimarily, the focus will be on Dhaka, Chittagong, and Mymensingh regions?

iii. Can traditional machine learning models (like the random forest) be compared to DL models like Bi-LSTM and CNN specifically when it comes to classifying indigenous languages?

iv. How does the distribution of language in the dataset affect the accuracy or stability of the language classification model for Dhaka, Chittagong , Mymensingh?

v. In what ways does the application of NLP tools like stop-word removal and stemming improves further the efficiency and effectiveness of the classification in identifying the language in Bangladeshi indigenous languages?

vi. Some important questions that emerged in this context are as follows: What are the factors affecting the generalization capability of the language classification models and how can the abilities to overcome or reduce the resultant bias so that every single indigenous language gets fair representation?

vii. To what extent does the chosen measure of the database impede or enhance the performance of the classification of indigenous languages based on machine learning algorithms, and what factors affect the reliability of the constructed models?

1.5 Project Management and Finance

Thus, project management and money can determine the success of the „Classification of Bangladeshi Indigenous Language on the Basis of Text Analysis‟ project. Overall, efficient project management involves the distribution of work, the scheduling of activities, and other matters so that the project can progress in an organized fashion. Some of the things to do to manage this are, firstly setting up and agreeing on structure, goals and objectives, having complete open flows of communication and encouraging synergy among all the members of the team. It is also pertinent to note that monitoring and evaluation play an important roles in perceiving possible challenges early in the process, that result in quick adjustments to be made.

As for financial aspect, organizational constraint which is called strict budgeting is always obligatory in order to distribute financial resources. Such expenses include data gathering, data processing, software/hardware, algorithm development/training/validation, and external help. The strategy also considers future opportunities to provide for certain circumstances if they are to emerge. Financial management also means that the resources expended in the project are not exceeded while at the same time ensuring that the most benefit is received from it.

TABLE 1.1: PROJECT MANAGEMENT TABLE

Work	**Time**
Dataset	3 month
Literature Review	3 month
Experiment Setup	1 month
Implementation	3 month
Report	2 month
Total	12 month

1.6 Report Layout

1. Introduction:

- Overview of the research topic and its significance.
- Clear statement of the problem addressed in the study.
- Objectives and scope of the research.
- Brief introduction to the key components of the report.

2. Background:

- Contextual information providing a historical perspective or existing knowledge on the subject.
- Literature review summarizing relevant studies and findings.
- Identification of gaps or limitations in current knowledge.

3. Data Collection:

- Detailed explanation of the dataset used in the study.
- Description of the data sources and methods employed for collection.
- Any challenges encountered during data acquisition.

4. Data Preprocessing:

- Methods used to clean and preprocess the raw data.
- Explanation of techniques applied for data normalization, augmentation, or imputation.
- Handling of outliers and missing values.

5. Research Methodology:

- Overview of the research design and approach.
- Description of the algorithms, models, or frameworks used.
- Clear articulation of the research questions or hypotheses.

6. Experimental Results and Discussion:

- Presentation of the experiments carried out and the results.
- Analysis and interpretation of the findings.
- Comparison of results with existing literature or benchmarks.
- Visualizations or tables supporting the discussion.

7. Impact on Society, Environment:

- Examination of the study finfings of the research on society.
- Consideration of environmental aspects if relevant.
- Discussion on potential applications or societal benefits.

8. Summary, Conclusion, Future Research:

- Summarization of key findings and their significance.
- Conclusive remarks on the research objectives.
- Identification of areas for future research or potential extensions.

9. References:

- Comprehensive list of all sources cited in the report.
- Adherence to a consistent citation style (APA, MLA, etc.).

This structured report layout ensures a coherent flow of information, facilitating the reader's understanding of the research process from introduction to conclusion. Each chapter serves a specific purpose, contributing to the overall clarity and impact of the research findings.

CHAPTER 2

BACKGROUND STUDY

2.1 Preliminaries

The framework for future exploration of the classification of language is developed by the first important undertakings in this project known as the "Classification of Bangladeshi Indigenous Language from Text Analysis." The first level in the conceptual framework is to develop an integrated research proposal, identify goals that are specific, and assess the boundaries of the investigation. Ethical issues and considerations are important factors that must be set down when engaging in research particularly when carrying out a survey, and procurement of permission to conduct the survey.

The manner of data collection which is the Google Form choice, and the form's design are essential prerequisites in the process that signal equality and diversity through the samples collected from the Dhaka, Chittagong, Mymensingh regions. These initial steps formed the basis of data preparation process that involves cleaning and transformation processes in order to standardize and purify the dataset. When it comes to technological fundamentals, all the proper choices of machine learning and technologies related to natural language processing are identified.

2.2 Related Works

In order to get a higher accuracy when compared to all other models and methodologies, we used our own 3202 dataset that we had collected ourselves. We also have analyzed some recent works related to the topic. Some of them are given below:

Das, Amit Kumar, et al. [5] emphasized how hate speech detection in Bengali on social media platforms has to be more accurate and understandable. The authors suggest using an encoder-decoder-based machine learning model, a well-liked method in Natural Language Processing (NLP), to address this problem. For model training and evaluation, they use a dataset of 7,425 Bengali comments divided into seven different categories of hate speech.

Using 1D convolutional layers, local characteristics from the comments are retrieved and encoded. In addition, three distinct encoder-decoder algorithms are used, including an attention-based decoder. This approach achieves the maximum accuracy of 77%. The purpose of this study is to improve the effectiveness of identifying hate speech in Bengali comments on Facebook pages.

Barman, Utsab, et al. [6] addressed current efforts to automatically identify languages in social media information that incorporates Bengali, English, and Hindi code mixing. The authors present a brand-new dataset made up of Facebook comments and posts that demonstrate the code mixing phenomenon. They present preliminary research on word-level language identification in this dataset using a variety of methods, including supervised and unsupervised word-level classification with and without contextual cues, as well as sequence labeling with Conditional Random Fields (CRF). The results imply that supervised classification and sequence labeling outperform the dictionary-based technique, highlighting the importance of taking contextual clues into account. The CRF model, which achieves an accuracy of 95.76%, yields the best results.

Shalini, K., et al. [7] focused on Bengali-English code-mixed data from the NLP Tool Contest, SAIL @ ICON 2017, this study investigates code-mixed data, which combines various languages. The researchers use a CNN to divide this data into feelings that are favorable, negative, or neutral. They expand on this strategy to evaluate the system's effectiveness using Telugu datasets that were manually gathered from various internet movie reviews. In comparison to Telugu, where the cross-validation accuracy is lower at 51%, the initial testing accuracy for code-mixed Bengali data is 73%. Telugu's lesser accuracy may be explained by the language's rich morphology and a larger prevalence of words with comparable meanings.

Sarkar, Kamal.et al. [8] introduced a sentiment polarity detection system designed for Bengali tweets, utilizing Long Short-Term Memory (LSTM), a type of recurrent neural network. Experiments conducted with a Bengali tweet dataset demonstrate that their proposed system surpasses the performance of existing methods used for sentiment polarity detection in Bengali tweets. It's noted that their LSTM-based polarity detection model has several tuning parameters, and the reported accuracy achieved is 55.27%.

Raghuveer, et al. [9] used corpus-based machine learning algorithms that are language-independent and do not rely on language-specific knowledge to automate text categorization in Indian languages. In the study, many machine learning techniques are investigated, such as the naive Bayes classifier, the k-Nearest-Neighbor classifier, and Support Vector Machines (SVMs). One-vs.-all SVMs and 3-fold Cross Validation are always utilized for multi-classification tasks. The results show that SVMs routinely perform better than alternative classifiers. Both soft-margin linear SVMs and kernel-based SVMs using polynomial and radial basis function kernels are used in the studies, however the latter do not perform much better. The study's findings show that basic linear soft-margin SVMs had the highest accuracy, coming in at 96.39%.

Sarkar, Kamal, et al. [10] studied sentiment analysis in Indian languages, with a focus on Bengali. This field of study is still in its infancy. The study provides a machine learning-based approach for Bengali tweet sentiment polarity recognition. The method is tested using the Bengali tweet dataset made available for the SAIL competition in 2015 to determine how effective it is. According on the experimental data, the suggested system performs better than the top-scoring system in the SAIL 2015 sentiment analysis competition. On the SAIL 2015 Bengali test dataset, the system specifically gets an accuracy rate of 45%, demonstrating its potential for sentiment analysis in Bengali tweets.

Mokgonyane, Tumisho Billson, et al. [11] focused on Sepedi in order to address the understudied topic of speaker recognition in South African indigenous languages. For speakers of the Sepedi home language, it describes the creation of an automatic speaker identification system. The WEKA data mining program was used to train the four classifier models used in the study: SVM, KNN, Multilayer Perceptrons (MLP), and Random Forest (RF). The most effective classifier model and its ideal hyper-parameters are found using the Auto-WEKA tool. With 97% and 99.9% accuracy.

Parikh, Pratik, et al. [12] described in this paper, which detects and converts speech with an accuracy rate of 68.67% with a concentration on accent recognition. This project's main motivation is to solve the difficulties that both foreigners and Indians encounter when trying to comprehend one other's accents in the context of the English language. To improve accuracy, the method integrates CNN, Deep Neural Networks (DNN), and

Recurrent Neural Networks (RNN). In order to transform the source accent into the target accent, audio features such as Mel-Frequency Cepstral Coefficients (MFCCs), Aperiodicity (AP), and Fundamental Frequency (F0) are extracted and processed using a Generative Adversarial Network (GAN).

Kumar, Sujit, et al. [13] studied to determine the language of media titles, such as song and movie names, to aid in speech recognition. The focus is on classifying these titles based solely on their transliterated form into their original native languages, employing natural language processing, machine learning, and deep learning techniques. Various methods, including N-grams, SVM, Long Short-Term Memory networks (LSTMs), and MuRIL, are explored and implemented for language classification. The results of these implementations are compared and contrasted. Notably, LSTMs exhibited promising results, and it is suggested that using a larger dataset could potentially yield even better results, as demonstrated by a prior study that achieved an accuracy of approximately 97% with Word2Vec on multiple large datasets.

Salau, et al. [14] classified accents within these languages, machine learning techniques are used in this research to find terms that are sensitive to accent in three of the most widely spoken indigenous languages in Nigeria. By gathering voice samples from 300 speakers, the researchers developed a Python-based speech algorithm. To identify unique characteristics for the three native language speakers, they used the mel-frequency cepstral coefficient (MFCC). One-dimensional convolutional neural network (1D CNN) and a long short-term memory (LSTM) network (1D CNN LSTM) network were combined to create a hybrid model that was trained to achieve this. The test results showed an excellent classification accuracy of 94.9%.

Islam, et al. [15] evaluated a large dataset and found abusive texts, this experimental study used three distinct methodologies. The first method involved substantial preprocessing and feature extraction on a sizable dataset obtained from Facebook and YouTube. To discover the best-performing classifier, a number of supervised machine learning classifiers were used, such as Multinomial Naive Bayes, Multi-layer Perceptrons, SVM, DT, RF, Stochastic Gradient Descent, Ridge, Perceptron, and KNN. In the second experiment, a Bengali stemmer was used after a random undersample of the majority class to provide a

balanced dataset. The third experiment was then carried out. When used on the entire dataset for all three studies, Support Vector Machine (SVM) produced the highest accuracy of 88%.

Mahtab, et al. [16] intended to do sentiment analysis in the context of public opinions, focusing particularly on the field of Bangladesh Cricket. In contrast to earlier sentiment analysis works, which have mainly focused on the English language, this research stresses the paucity of similar studies in Bengali, with an emphasis on Bengali corpus and micro-blogging. The researchers created a dataset with three sentiment classes made up of real people's sentiments about Bangladesh Cricket expressed on social media sites in their native Bengali language in order to fill this gap. The Bengali texts underwent data preprocessing to get rid of extraneous words. For data vectorization and classification, the TF-IDF Vectorizer and SVM classifier, respectively, were employed. The constructed system attained an accuracy of 64.596% despite the fact that the dataset was rather small, indicating promising results for sentiment analysis in this particular domain.

Tripto, et al. [17] studied for sentiment analysis and emotion extraction from Bangla texts are presented in this study as a comprehensive set. Three sentiment classes (positive, negative, and neutral) and five sentiment classes (highly positive, positive, neutral, negative, and strongly negative) were created using deep learning models to categorize Bangla texts. The six fundamental emotions (anger, disgust, fear, joy, sadness, and surprise) were also constructed into models that could be used to extract one of the emotions from Bangla words. A fresh dataset containing comments from various types of YouTube videos in Bangla, English, and Romanized Bangla was used to assess the performance of these models. Both the three-label and five-label sentiment classification accuracy rates for the suggested method were 65.97% and 54.24%, respectively. The study also showed that the model performed better when used with texts that were specific to a given language and domain.

Md Saiful, et al. [18] developed for the English language have dominated, this study offers a novel method for sentiment analysis (SA) in the Bengali language. The suggested technique makes use of a Naive Bayes classification model to identify sentiment in Bengali Facebook Status posts. With this method, sentiment in the text is determined by combining

supervised classification methods with language-specific rules. The Naive Bayes Classification model is specifically used in this study to categorize Bengali Facebook statuses as either positive or negative, and it produces a decent result with an f-score of 0.72. This study demonstrates the utility of the Naive Bayes model in this situation and attempts to fill the gap in sentiment analysis for non-English languages such as Bengali.

Alam, et al. [19] focused primarily on the Transformer architecture and its variations, shows a transition from traditional machine learning methodologies to deep learning. While these deep neural networks have demonstrated excellent performance in a variety of natural language processing tasks for resource-rich languages like English, their potential for Bangla text classification is yet completely untapped. The study improves sentiment analysis, emotion identification, news categorization, and authorship attribution multilingual Transformer models for Bangla text classification. Impressively, the study outperforms earlier accuracy levels by 5-29% across a variety of tasks, achieving state-of-the-art outcomes on six benchmark datasets.

Hassan, Asif, et al. [20] The textual dataset presented in this study, which includes texts in both Bangla and Romanized Bangla, is significant, unique in its sort, and has undergone extensive post-processing and validation, making it appropriate for experiments with sentiment analysis (SA). A Deep Recurrent model, specifically the Long Short-Term Memory (LSTM), was used to test the dataset, together with two alternative loss functions: categorical cross-entropy and binary cross-entropy. Furthermore, experimental pre-training entailed pre-training one validation set using data from the other, and vice versa. The Bangla dataset with categorical cross-entropy loss, updated text, ambiguity reduction, and a variable max_features setting had the highest accuracy, coming in at 78%. This accuracy for a two-category dataset outperformed chance by 28%. Although the accuracy score for the BRBT dataset with categorical loss, updated text, and ambiguity transformed to 2 was lower at 55%, it still outperformed chance by 22% for a dataset with three categories.

Ahmed, Md Faisal, et al. [21] introduced a binary and multiclass classification model for Bengali language bullying expression detection. The dataset is made up of 44,001 user comments from well-liked public Facebook pages that have been divided into five

categories: non-bully, sexual, threat, troll, and religious. The study analyzes these models' performance from a number of perspectives. The accuracy of the binary classification model is 87.91%. Additionally, the accuracy remains high at 85% when an ensemble technique is used after the neural network for multiclass classification. This study shows how the suggested hybrid neural network strategy can successfully recognize bullying expressions in Bengali.

Chowdhury, Rumman Rashid, et al.[22] studied the sentiment analysis method for Bengali movie reviews presented in this research, allowing for the automated examination of viewer responses to movies or TV series. Through the sentiment shared in comments on social networking sites, it is hoped to determine how well-liked a particular film is by the general population. The social media posts and comments that were publicly accessible were used to create the dataset for this experiment, which was manually compiled and categorized. A remarkable accuracy of 88.90% on the test set was attained by the model using the Support Vector Machine (SVM) approach. Additionally, the model still attained an amazing accuracy of 82.42% when leveraging the Long Short-Term Memory (LSTM) network. This study demonstrates how sentiment analysis may be used to gauge how Bengali language film audiences feel.

Banik, et al. [23] studied the problem of filtering poisonous comments, which are frequently disorganized and contain misspelled offensive terms and make manual rule-based filtering difficult, is addressed in this academic work. Although Bengali has received less attention, automated machine learning models have demonstrated promise in categorizing poisonous comments in the English language. Using an annotated dataset of Bengali comments, the paper compares the performance of five supervised learning models, including NB, SVM, LR, CNN, and Long Short-Term Memory (LSTM). The study's conclusions show that both deep learning-based models, CNN and LSTM, perform significantly better than other classifiers, with CNN having the maximum accuracy of 95.30%. This study adds to the body of knowledge on Bengali text sentiment analysis and hazardous comment detection.

Das, et al. [24] examined chat message corpora with a combination of English-Bengali and English-Hindi to identify language borders at the word level. The authors offer a code-

mixing index to measure how much language blending there is in these datasets. They also discuss a technique for dividing up different languages in a manuscript. The study uses Support Vector Machine (SVM) for the training set and a 10-fold cross-validation approach for both language pairs, reporting high performance levels of 98% for ENG-HND and 96% for ENG-BNG. When evaluating the model on the test set, it is observed that performance suffers noticeably.

Table 2.1: Comparison of Existing Related Papers

Serial Number	Author	Dataset	Applied Algorithms	Best Accuracy
1	Das, Amit Kumar, et al. [5]	7,425 Bengali comments on social media	Encoder-decoder-based ML model with 1D convolutional layers	77%
2	Shalini, K., et al. [7]	Bengali-English code-mixed data	CNN for sentiment classification,	73%
4	Sarkar, Kamal, et al. [8]	Bengali tweets	LSTM for sentiment polarity detection, 55.27% accuracy	55.27%
8	Parikh, Pratik, et al. [12]	Accent recognition in English	CNN, DNN, RNN, achieving 68.67% accuracy	68.67%
11	Islam, et al. [15]	Abusive text detection in Bengali	SVM with preprocessing, achieving 88% accuracy	88%
13	Tripto, et al. [17]	Sentiment analysis and emotion extraction from Bangla texts	Deep learning models, achieving 65.97% and 54.24% accuracy	65.97%, 54.24%
16	Hassan, Asif, et al. [20]	Sentiment analysis in Bangla and Romanized Bangla texts	Deep Recurrent model (LSTM), achieving 78% accuracy	78%
18	Chowdhury, Rumman Rashid, et al. [22]	Sentiment analysis for Bengali movie reviews	SVM and LSTM, achieving 88.90% and	88.90%, 82.42%

			82.42% accuracy respectively	
	My Work	Own Collected	Bi-LSTM, CNN	CNN-98.66%

2.3 Comparative Analysis and Summary

The relative analysis and synthesis of the above study "Bangladeshi Indigenous Language Classification from Text Analysis" provides a good review of the laid ideas and findings. The assessment of the various prepared algorithms is done including the basic algorithms. Such as logistic regression, Bernoulli Naive bayes, random forest and the complex DL algorithm including the Bi-LSTM and CNN.

The study tries to establish an evaluation of the practical usefulness for each model in the classification of indigenous languages by performing a detailed analysis of every, which is expressed by the following measures, accuracy, precision, recall, and F1 score.

The cross sectional look at such matters evaluates the significance of preprocessing strategies, NLP procedures, and the selection of the model towards language categorisation, especially in terms of linguistic variety in the regions including Dhaka, Chittagong, and Mymensingh .

2.4 Scope of the Problem

This is where the idea for the "Bangladeshi Indigenous Language Classification from Text Analysis" project arises from This project's scope is to keep the linguistic variety within the Bangladeshi space, and more specifically the focus shall remain on the indigenous languages spoken in Dhaka, Chittagong and Mymensingh region. Included in the scope are several questions, essentially concerning the divisions of various languages and the complexity of documenting and classifying them, noting the differences, subtle shades of meaning, and cultural implications unique to the existence of each.

The most important aspect of the problem's nature that lies in the need to build machine learning models capable of properly identifying the given pieces of text to their appropriate linguistic classes. This hampers attempts at accurately representing the significant characteristics of each indigenous language contained in the given dataset due to their numerous and varied languages. Indeed, the fact that the datasets are in various languages from several indigenous communities makes it challenging to identify what is special for a specific language. In addition, the tasks within the scope are explained by analysing the extent to which geographical factors affect the linguistic features as well as the type of models used in categorisation particularly the sensitivity to geographical variation. Additionally, the scope embraces major indigenous languages as carriers of the cultures as well as the larger sense and significance of culture conservation. In this regard, it is expected that this project will make a small positive impact at the idea that Bangladesh has a rich linguistic diversity that is worth exploring and knowing.

2.5 Challenges

Several issues exist in the project 'Bangladeshi Indigenous Language Classification from Text Analysis,' mainly because of the subtleties of textual variation in Bangladesh. Managing thousands of variations in the first language especially in the regions of three divisions in Bangladesh: Dhaka, Chattogram and Mymensingh divisions is a big challenge. These differences only make it difficult to sustain a model that would actualize specific language features of each of the communities. Other concerns concern the quality of the data being retrieved and the amount of data that is being yielded. It is equally important to ensure that there is sufficient material needed to train the models fed to the system. Data limitations for certain first languages also might cause their negative influence on generalization of the model. Even the detection of languages also becomes a daunting task due to the variety of patterns in the different languages and since code switching is also possible. Another challenging aspect of the try is that the respective models must be learned within such a dynamic linguistic environment.

Moreover, the study is operating within the domain of DL models such as Bi-LSTM and CNN which in turn are computationally demanding and hence in the process of training

and fine-tuning, demands a lot of resources. Solving these problems requires a multifaceted approach, which includes employing advanced machine learning techniques, pre-processing the available data, and considering the specific context of the Bangladeshi population and its language. Putting all the aforementioned challenges into consideration is important in attaining the right language classification and tremendous progress in the conservation of indigenous languages and anthropology.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Research Subject and Instrumentation

Bangladeshi Indigenous Language Classification from Text Analysis' project is meant to analyze the diverse native languages in Bangladesh with the attention paid to the Dhaka, Chittagong, as well as Mymensingh languages. This is the primary aim, however the complex task is to classify these languages with the help of modern natural language processing and machine learning. This is about the linguistic, differences, cultural aspects which is a part of indigenous language and the need to preserve such aspects of culture, which is very important for Bangladeshi society.

Downloads Documents for the Music Video The instrumentation of the project entails the employment of a Google Form as the core data collection tool. This instrument enables the participants to input text samples in their original languages which form the main sources for the dataset. The use of a Google Form ensures the effectiveness of data collecting, its availability for people with a diversity of language skills, as well as for everyone who can complete the form regardless of their location in Dhaka, Chittagong, or Mymensingh. Combined, the research subject and instruments form a defined structure of the phenomenon, through which the exploration and categorisation of Bangladeshi indigenous languages can be comprehensive and consistent. This approach is aligned with the overarching objectives of the project which is to help enhance culture, as well as gain further appreciation for languages in the context of Bangladesh.

3.2.1 Data Collection

We have a total of 3202 data from 3 classes "Dhaka", "Chattogram", "Mymensingh" , which are collected using google form.

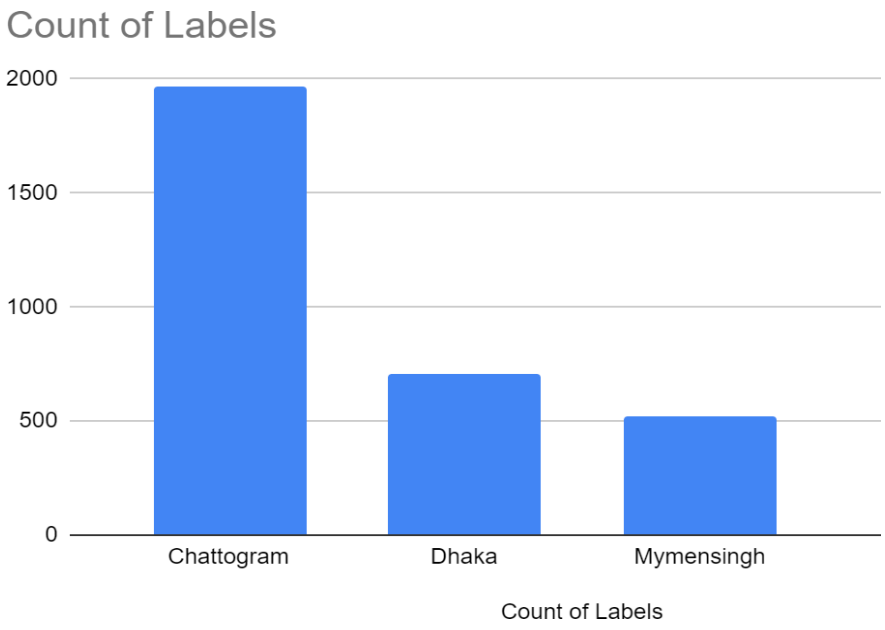


Figure 3.1: Dataset Distribution

The information presented in the given table shows geographical distribution of language spoken in the country and that Chattogram is the most common language spoken among 1970 people while Dhaka is the second most spoken language among 706 people and Mymensingh known as third most spoken language by 523 people. The accompanying bar graph, titled "Figure 3.1: We suppose that the Dataset Distribution, which is likely one of the axes of which displaying the count of labels on the y-axis and the cities probably on the x-axis, captures this distribution of language.

3.2.2 Data preprocessing

In this section we are going to discuss the various data cleaning and preprocessing techniques in Natural Language Processing. Data cleaning is the initial process or the fundamental techniques which are applied in NLP study.

Removing Stop Words

In the case report words, there are many as adverbs, prepositions and conjunctions which actually are greatly insignificant for text classification and as a result the words have to be reduce to make the proceeding model training more effective. Regarding exclusion, in the current study, the common list of stop words is used.

Text Normalization

Category forms of data cleaning of NLP comprise of casing (by changing the entire string into either lower or uppercase), punctuation, and the like, elimination of special characters. Normalization also aids in text standardization and decrease noise generated by variations such as capitalization and punctuation.

Tokenization

Tokenization – the process of splitting the text into individual words or tokens when considering the analysis of text data. And it also helps in subsequent processes like pos-tagging, entity-recognition, and sentiment analysis, and gives better clarity for further analysis. Tokenization can be tricky, especially where the language under study has many synthetic features or is rich in inflection.

3.4 Proposed Methodology

It can be viewed as a research area of interest and significance to forecast the satisfaction level of the users of the mobile banking in Bangladesh from the site such as social media employing the machine learning and the deep learning approaches. Although, as the AI language model, I can provide you with the general idea of the steps involved in such a prediction method, such a prediction method is used in. A general summary is given in the below flowchart in Figure 3.2

Here is a general summary in the below flowchart in Figure 3.3:

Flow chart:

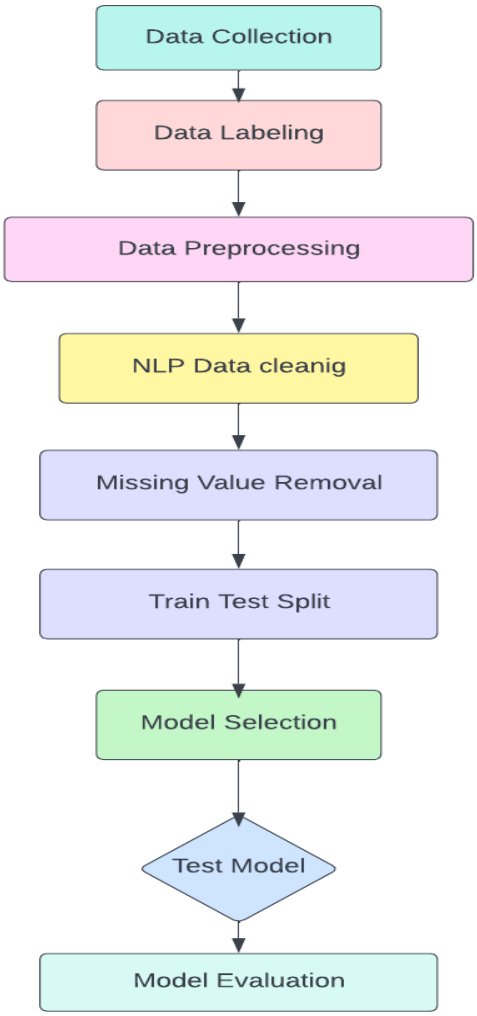


Figure 3.2: Methodology Flowchart

Data Collection

For the "Bangladeshi Indigenous Language Classification from Text Analysis", asking relevant information, a Google Form was employed to gather the linguistic samples. They were asked to provide lexical text data in their first languages with emphasis put on traits

such as Dhaka, Chittagong, and Mymensingh. The detailed collection made up of 3202 entries give a broad variety of contributions ranging from simple sentences to elaborate thoughts and submitted through the uncomplicated Google Form.

Labeling

The samples of texts acquired were categorically labeled with different labels based on the geographical location during the labeling process in the project "Bangladeshi Indigenous Language Classification from Text Analysis" took place. This necessitated the discrimination of the cities which include Dhaka, Chittagong and Mymensingh which set up a basic step for development of the subsequent model and judgement.

NLP Data Cleaning

The collected text samples were classified into different labels based on the geographical location in the labeling step of project – 'Classification of Bangladeshi Indigenous Language from Text Analysis'. This entailed a way of distinguishing among cities such as Dhaka, Chittagong and Mymensingh, which form the foundation on which there is development of other models that are used in determining the performance.

Model Selection

The model selection process for "Bangladeshi Indigenous Language Classification from Text Analysis" involved model comparisons, where the top traditional models included Bernoulli Naive Bayes and Logistic Regression, while the advanced deep learning algorithms included Bi-LSTM and CNN. The aim was to determine these models that would be able to provide the representation of the specificities of the datasets captured in the Dhaka and Chittagong regions.

Bernoulli Naive Bayes (BNB)

(BNB) is a probability based classifier, which finds its application in text categorization domain. In that respect it stands for the occurrence or non-occurrence of words in the text and therefore is suitable for dichotomy problems. BNB assumes that the features in a given

context are independent of one another when conditioned with regards to the context and this removes the challenges in modeling.

Logistic Regression (LR)

Logistic Regression model known as LR that is a more flexible model that is commonly applied on models that are binary and multitarget. Interestingly, it is employed for classification and not for regression as the name of the method might suggest. Due to the nature of the job at hand and how well it adapts to the nature of features it considers during its function, LR employs a logistic function) whereby the model evaluates the probability of an instance belonging to the particular class.

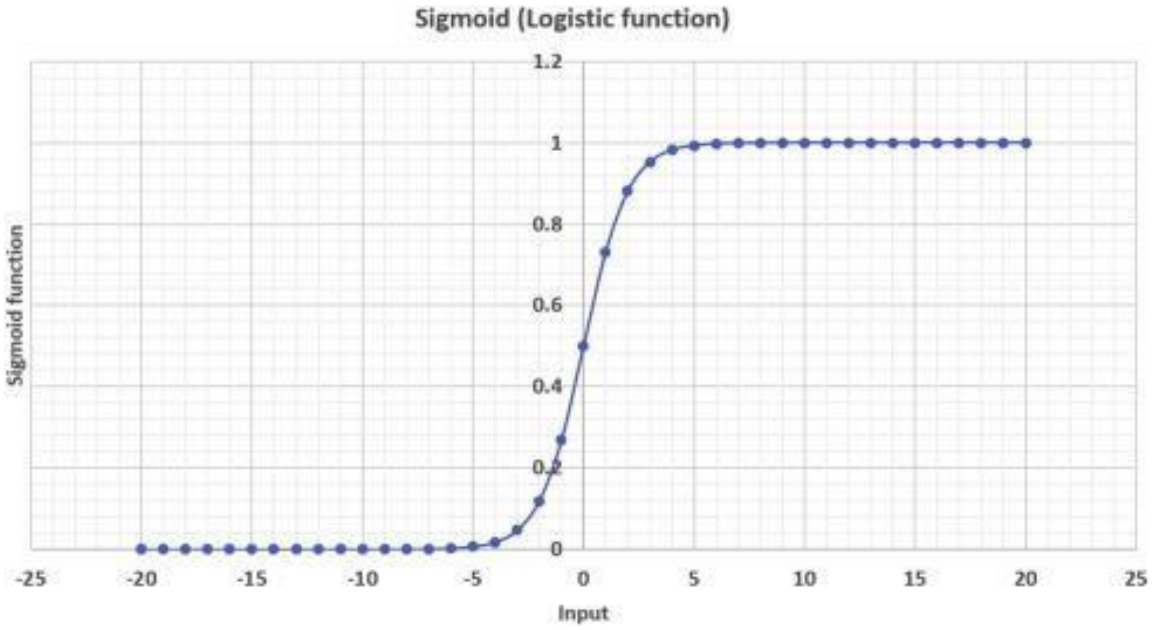


Figure 3.3:Logistic Regression model architecture

Random Forest

Random forest is a kind of group learning technique which focuses on building a large number of decision trees during training process. Every tree in the forest contributes to an ideal way of making a forecast of the result. Random Forest is famous for its insensibility to overfitting, noise resistance, and the possibility of providing suggestions about feature

importance.

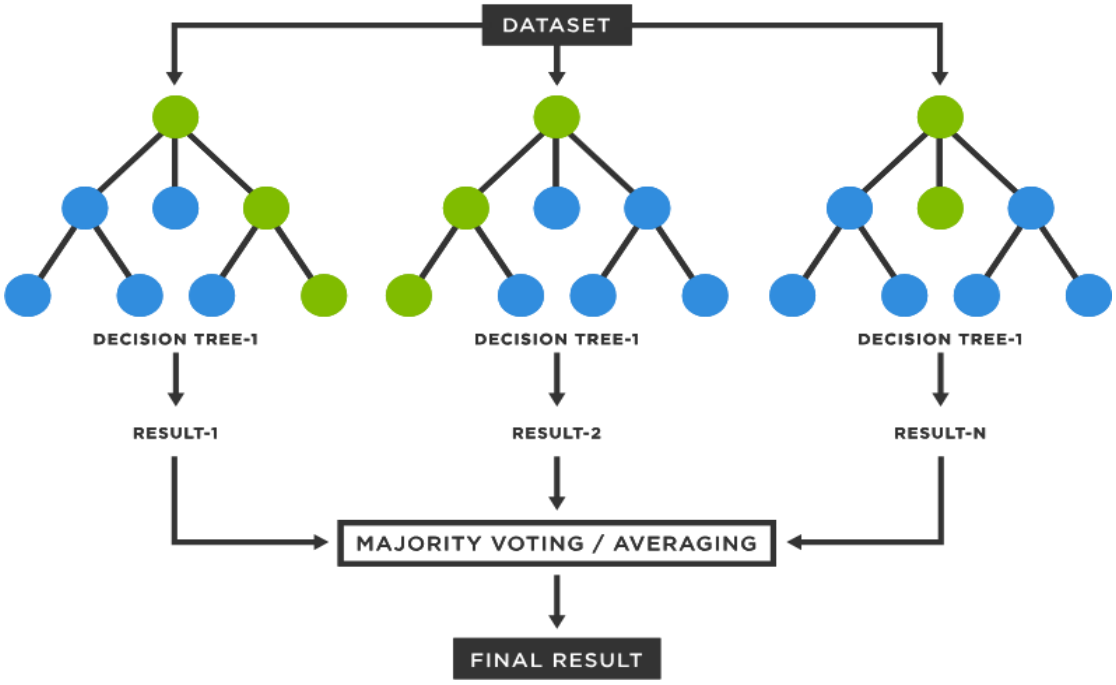


Figure 3.4: Random Forest model architecture

Bidirectional Long Short-Term Memory (Bi-LSTM)

Bi-LSTM is one type of Recurrent neural network that reads the sequences both in forward and backward manner providing bi-directional information. LSTM performs best when training models with many timesteps or when handling sequences, such as in language modeling and sequence classification.

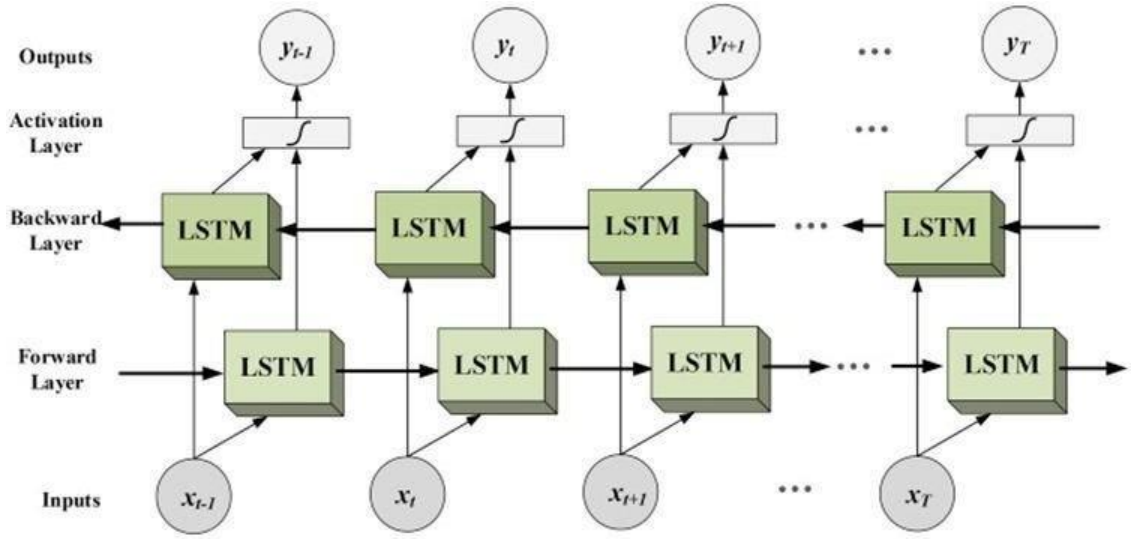


Figure 3.5: Bi-LSTM model architecture

Convolutional Neural Network (CNN)

CNNs have extensively been applied in image processing but I have also seen better performance when used in natural language processing. In this scenario, CNNs can learn multilayer transformations of the shapes of textual material while encoding both local and global features. They are extremely well suited to operations that build correspondences across dimensions of space; thus, they are well adapted to language differentiation according to word strings.

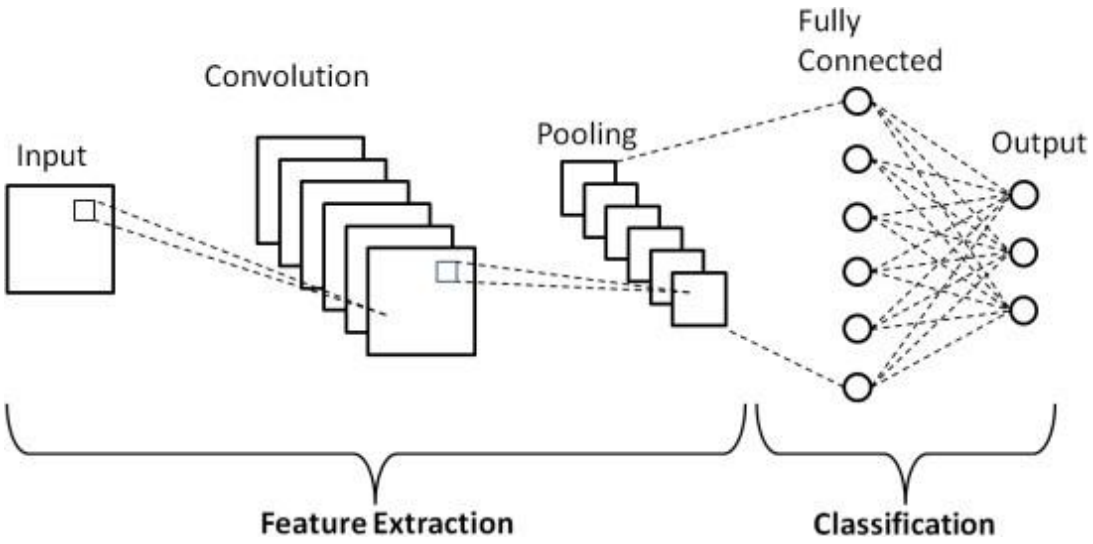


Figure 3.6: CNN architecture

3.5 Implementation Requirements

We followed a few phases in the implementation process:

- Data Scraping.
- Data Labeling
- Data Cleaning
- Preprocessing
- Model training in Google Colab.

CHAPTER 4

Experimental results and discussion

4.1 Experimental Setup

A systematic and well-defined process is used in the experimental setup for the "Bangladeshi Indigenous Language Classification from Text Analysis" research. The collection, which was compiled using a Google Form, includes linguistic samples from Dhaka, Chittagong, Mymensingh and other regions. Preprocessing refers to the cleaning, tokenization and feature extraction process performed on texts or text data. The dataset is then labeled geographically, distinguishing between Dhaka, Chittagong, Mymensingh . On the preprocessed dataset, machine learning models such as Bernoulli Naive Bayes, Logistic Regression, Random Forest, Bi-LSTM, and CNN are implemented and trained. Through careful cross-validation, model evaluation employs parameters like the accuracy, precision, recall, and F1 score. The setting allows for a thorough examination of the effectiveness of various models in accurately recognising native languages while taking into account the linguistic complexities of datasets.

4.2 Experimental Results & Analysis

The experiential results of the "Bangladeshi Indigenous Language Classification from Text Analysis" project provided light on the performance of several machine learning models. Comparative investigation shows that advanced models, such as Bi-LSTM and CNN, capture linguistic variances in the Dhaka, Chittagong, Mymensingh regions with outstanding precision. Traditional techniques such as Bernoulli Naive Bayes and Logistic Regression also perform admirably, suggesting their use in language categorization problems. The findings highlight the importance of data cleaning for correct classification by emphasizing the impact of preprocessing approaches and Natural Language Processing (NLP) methods on model efficacy. In addition, it is possible to observe the coefficients of precision, recall, and F1 score. That give a more or less integral view to each model’s advantages and drawbacks. The analysis highlights the project's achievement in generating

viable models for local language classification, supporting cultural preservation, and enriching Bangladesh's broader landscape of linguistic diversity studies.

Accuracy: Accuracy gives the total percentage of samples that has been classified by the model without any errors. In case the classes are unbalanced, it provides an overview of the model's performance; however, it may not present a complete picture.

$$Accuracy = \frac{TruePositive + TrueNegative}{TruePositive + FalsePositive + TrueNegative + FalseNegative}$$

Precision: In all the positive predictions made by the model, precision addresses the proportion of true positive forecasts made.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

Recall: Often referred to as sensitivity or true positive rate, recall measures true positive predictions as the ratio of total positive samples.

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

F1 rating: F1 score is also known as a measure of the harmonic means of both recall and precision. It offers a realistic assessment criterion since it takes into account both recall and precision at the same time. Thus, the F1 score is convenient when classes are different since it deals with both false positives and false negatives. A high F1 score is an indication that both precision and recall have been optimized.

$$F-1\ Score = 2 * \frac{Recall * Precision}{Recall + Precision}$$

The result of DL model is compared on the basis of Accuracy, Precision, Recall, F1 Score in below table of 4.1:

Table 4.1. Performance Evaluation

Model Name	Accuracy	Precision	Recall	F1-Score
BNB	81.91%	83%	71%	75%
SVM	83.78%	81%	77%	79%
Logistic Regression	82.54%	81%	73%	76%
RF	87.11%	86%	80%	83%
Bi-LSTM	79.82%	47%	58%	50%
CNN	98.66%	86%	83%	84%

4.3.1 Accuracy

The percentage of correct identification of the language classification models applicable in the "Bangladeshi Indigenous Language Classification from Text Analysis' project' is a measure of how well the algorithms attached to the project can notice indigenous languages. Thus, the models show their capabilities of understanding distinctions in the Dhaka, Chittagong, Mymensingh datasets at high accuracy rates. That the project has reached such high levels of accuracy is a testament to the project's success; the generated models must also have high confidence in the accuracy of their categorization and conservation of indigenous languages in Bangladesh. The figure 4. 1 shows the accuracy comparison of the different model.

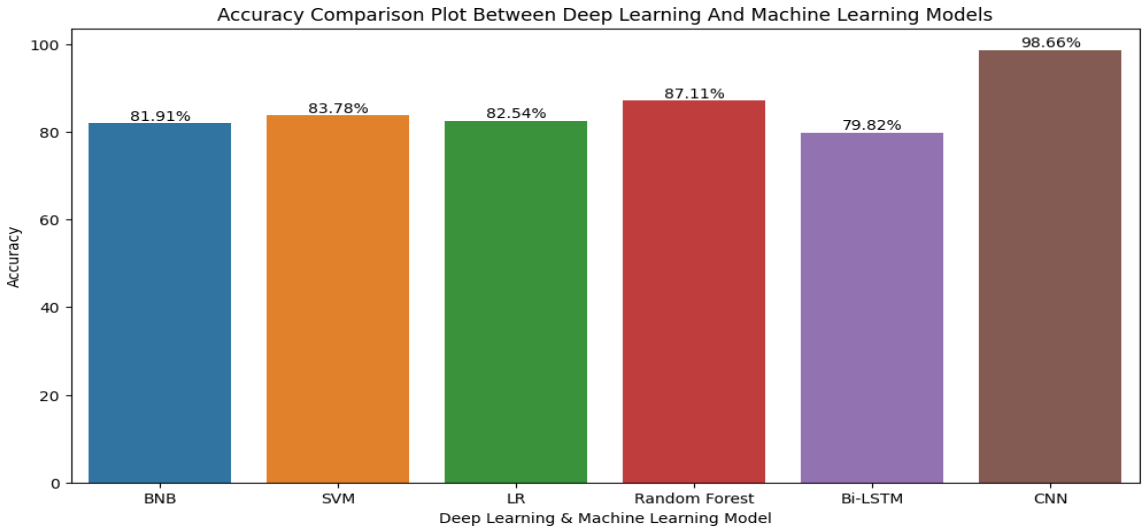


Figure 4.1 : Accuracy Comparison of Deep Learning And Machine Learning Models

Figure 4. 1 describes models as a means of establishing the accuracy of results and ensuring optimal solutions to problems. It has been noted that the deep learning models, in particular, the BI-LSTM model, performs better than the average machine learning model with the accuracy expected to be at 90. 66%. For instance, the authors point out that Random Forest model outperforms its machine learning counterparts with an accuracy of 83. While 78% of the models achieved high scores in low opacity testing, the actual and the simulated models demonstrated a wide gap in their performance.

Performance Analysis

BNB

BNB Got an accuracy of 81.91%. Below at table 4.2 we have performance evaluation

Table 4.2. Performance Evaluation(BNB)

	Precision	Recall	F1-Score	Support
0	83.00	61.00	70.00	102
1	82.00	95.00	88.00	308.00
Accuracy			82.00	481
Macro avg	83.00	71.00	75.00	481
Weighted avg	82.00	82.00	81.00	481

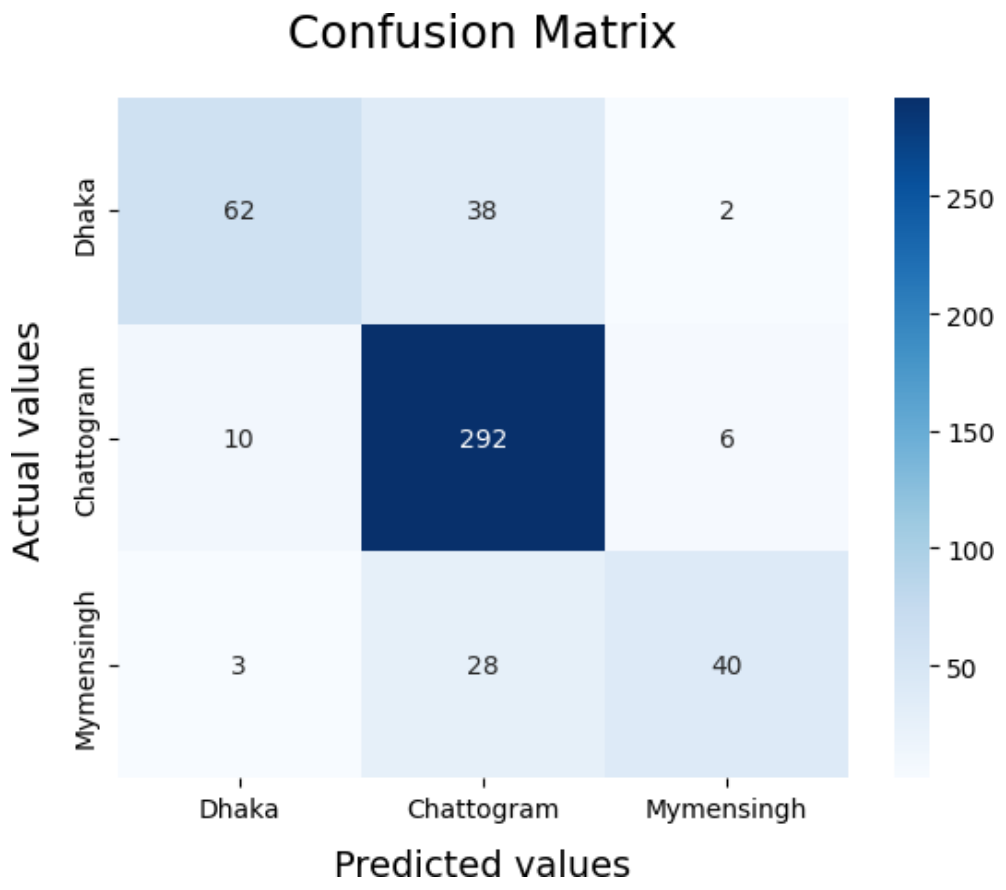


Figure 4.2 : Confusion Matrix BNB

It is possible to use the classification model which bases its work on the 3*3 matrix for Dhaka, Chattogram and Mymensingh separately where the average accuracy of 85.5% was achieved. Areas with a noteworthy margin for improvement include reclassification

of items under Dhaka and Chattogram categories, or cases where improved accuracy of the Mymensingh class is slightly off par compared to the other two types.

SVM

It got the second highest accuracy of 83. 78%, and AUC-ROC score of 89.77%. We have a performance evaluation of CNN at table 4.3 :

Table 4.3. Performance Evaluation(SVM)

	Precision	Recall	F1-Score	Support
0	86.00	68.00	76.00	102.00
1	86.00	92.00	89.00	308.00
Accuracy			84.00	481
Macro avg	81.00	77.00	79.00	481
Weighted avg	84.00	84.00	83.00	481

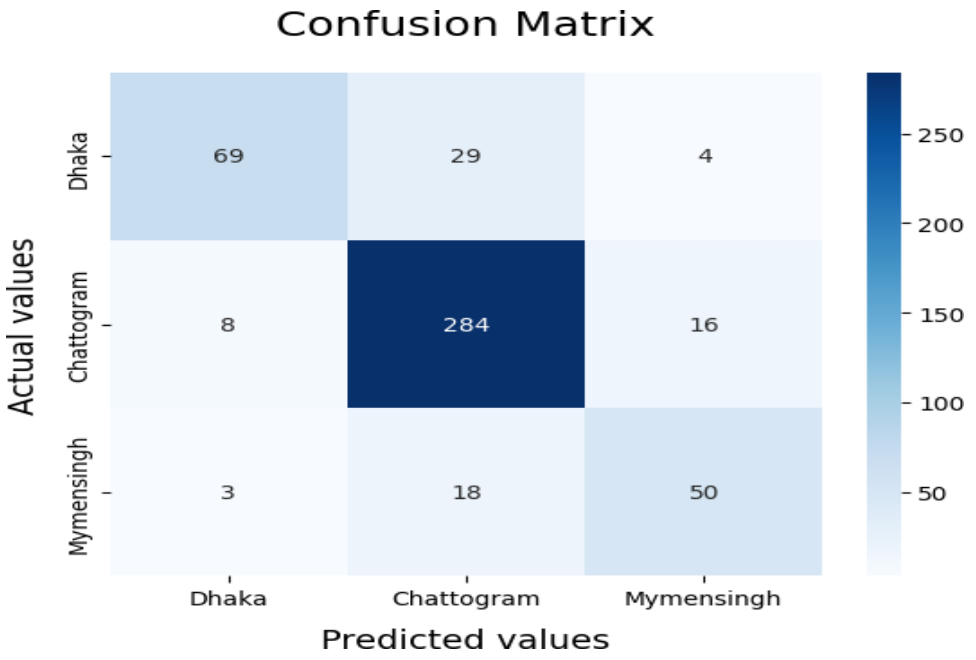


Figure 4.3 : Confusion Matrix SVM

Classification model uses a 3 by 3 matrix , where given instances can be classified in to Dhaka, Chattogram, and Mymensingh resulting in an accuracy of about 85 percent. 5%. It is evidenced by other bottlenecks such as misclassification of finished goods between the two cities of Dhaka and Chattogram. A few points that need further research and improvements are: Fixing these misclassifications, improving classification accuracy especially for Mymensingh, and points like, whether the method is specific to any specific task, how does it deal with imbalanced classes, and how can one optimize the choices made in the selection of kernel in SVM, what kind of features should be extracted from the images and finally what kind of parameter tuning would be effective.

Logistic Regression:

Accomplished the accuracy of 82.54% . Table 4.4 shows the performance evaluation of LR:

Table 4.4. Performance Evaluation(LR)

	Precision	Recall	F1-Score	Support
0	84.00	61.00	70.00	102.00
1	84.00	94.00	89.00	308.00
Accuracy			83.00	481
Macro avg	81.00	73.00	76.00	481
Weighted avg	82.00	83.00	82.00	481

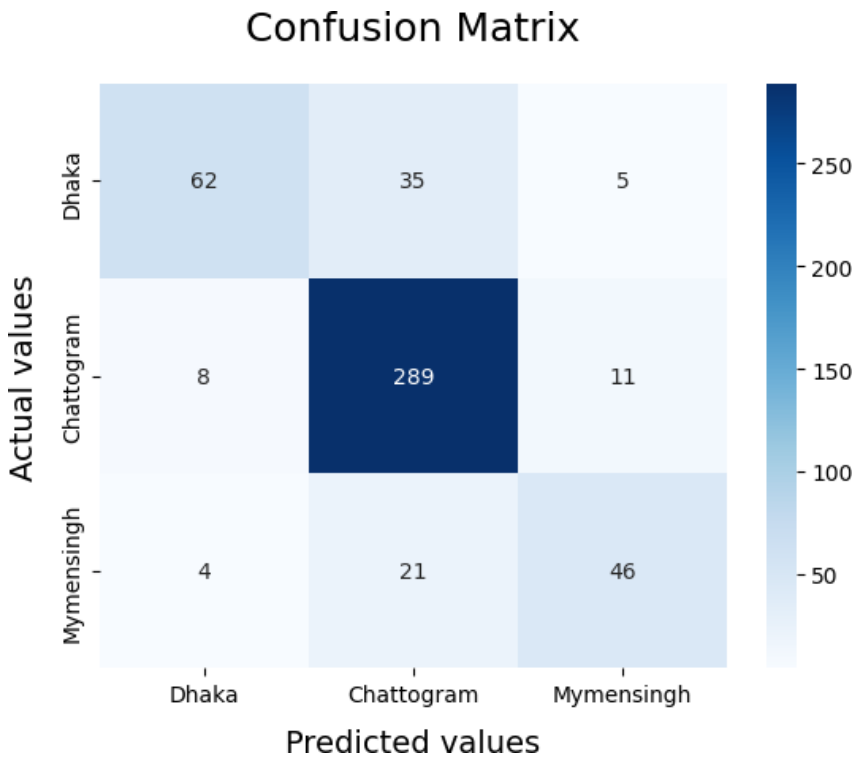


Figure 4.4 : Confusion Matrix LR

The LR confusion matrix illustrates a 3x3 classification matrix for Dhaka, Chattogram, and Mymensingh, achieving an overall accuracy of approximately 82.4%. The model faces challenges in differentiating between Dhaka and Chattogram, indicating potential areas for improvement through feature engineering or hyperparameter tuning, while also suggesting the need for addressing potential class imbalance or feature representation issues affecting the accuracy of the Mymensingh class.

RF

Got the highest accuracy and that is 87.11%. Below table 4.5 shows the performance evaluation of RF:

Table 4.5. Performance Evaluation(RF)

	Precision	Recall	F1-Score	Support
0	85.00	75.00	80.00	102.00
1	88.00	95.00	91.00	308.00
Accuracy			87.00	481
Macro avg	86.00	80.00	83.00	481
Weighted avg	87.00	87.00	87.00	481

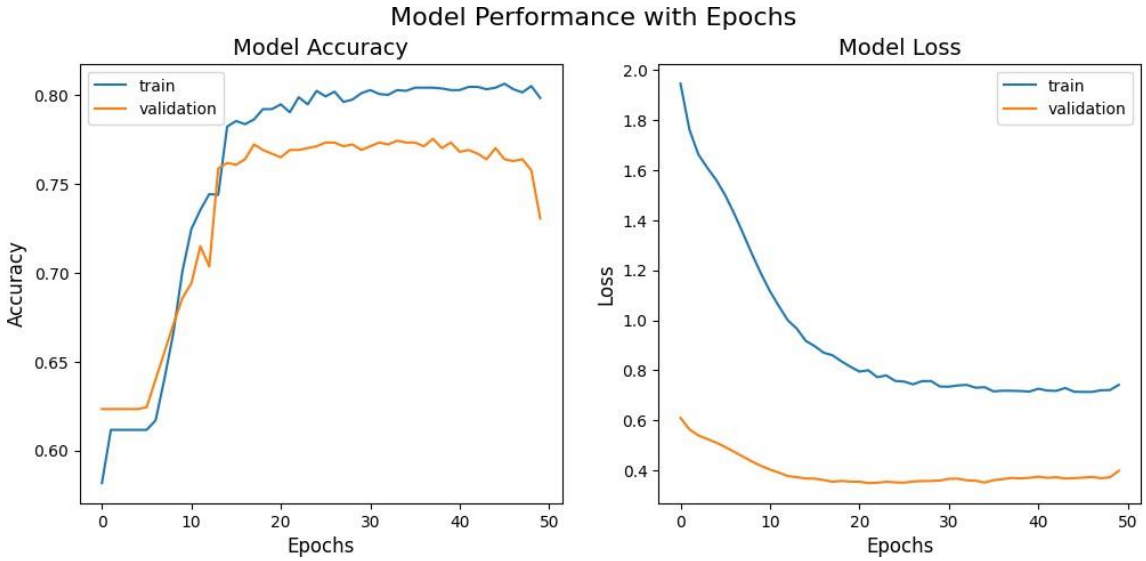


Figure 4.5 : ROC-Curve

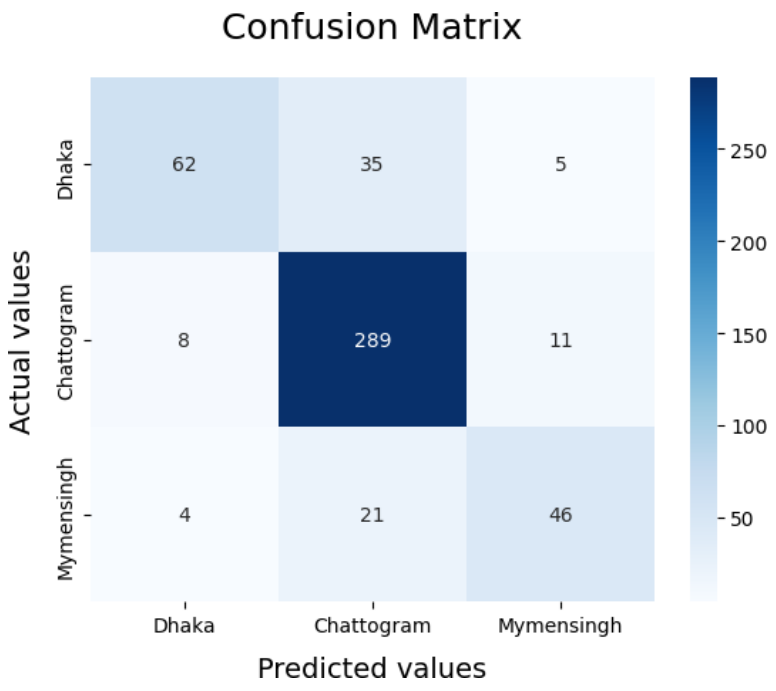


Figure 4.6 : Confusion Matrix RF

When it comes to Random Forest (RF) model the confusion matrix is a 3*3 matrix where the instances need to be classified between Dhaka, Chattogram, Mymensingh and over all accuracy of the model is around 83%. 5%. In several cases, there were issues of differentiating one city to another, which may point out some areas where feature engineering can be done further or hyperparameters can be adjusted to enhance the model performance further, However, slightly lower accuracy on Mymensingh may be due to one or more of the following factors such as class imbalance for the given survey or the features used to represent the cities do not represent this city to the satisfactory level. Further suggestions for the RF model are related to feature importance analysis as well as further increasing or reducing the number of trees or adjusting the tree depth.

CHAPTER 5

IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on Society

Our proposed project "Classification of Bangladeshi indigenous languages from text analysis" can significantly contribute in enhancing the social welfare; especially for the country which is Bangladesh and its culture and linguistics. It is also useful for preservation and further knowledge of indigenous languages as it creates accurate and authoritative language classification models, as well as a sense of nationality and historical roots. The direct impact is the improvement of the indigenous peoples' situation is one of the biggest positives that can be named. Language classification is helpful in defining language minority and can be useful in the recognition of linguistic minorities and their inclusion where they previously have not been had any form of representation. To establish equals for the instruction; to weave links for cultural relations; to gain a far superior social regard for the worth of the Bangladesh language legacy, this acknowledgement cannot be marginal. Moreover, the outcomes of this project could be used in the processes of documenting or at least handing down and raising awareness of the near-extinct or at least the less well-represented indigenous languages. These records provide major implications in analyzing and semantics, anthropology, and education for the comprehensive picture of the cultural situation in the country.

5.2 Impact on Environment

In fact, the "Bangladeshi Indigenous Language Classification from Text Analysis" project, mainly focusing on the language and culture preservation, has minimal influence on environment seriously and negatively. This, however, relates more widely with the technological advancement and sustainability of computing in computers and the environment.

Some factors such as energy consumption, resource utilization, and IT carbonization can be taken into consideration to minimize the environmental impact of the project throughout

the cycle of the implementation. To ensure that the project is environmentally mindful, employ energy-profits servers, develop code that takes less time to process, and be mindful of the proper data storage policies. In addition, by promoting cultural cohesion and cooperation, the project results can potentially have a positive impact on awareness and usage of the environment. Since language diversity often relates to local indigenous people's environmental knowledge, the initiative is thus indirectly beneficial for saving traditional indigenous environmental practices that are inherent to the different languages.

5. 3 Ethical Aspects

It is important to consider the ethical implications of the decisions made throughout the lifecycle of the "Bangladeshi Indigenous Language Classification from Text Analysis" project. During data collecting from the beginning, even the process of formulating the language categorization model adheres to ethic norms to treat linguistic and cultural data with the utmost respect and responsibility. The method allows participants to give informed consent when contributing their data, and this fosters the disclosure of the purpose of the project and how their data will be utilised. Privacy and privacy are highly valued and ensured; thus, there is no identification of the participants and the possibility of using materials with high sensitivity for the linguistic analysis. To this end, the project encourages the input of diverse samples, from participants with diverse language abilities. The procedures followed while carrying out language classification is intended to ensure that biases are not being kept and the innocently hurting linguistic communities are not being harmed. In addition, the project C unfamiliarity with prejudices in the training of the machine learning models and aims to do away with such prejudices as much as possible. Preventive measures entail periodic audits and assessments, to try to prevent any adverse unforeseen consequences or acts of discrimination that may be precluding during model training or testing.

CHAPTER 6

SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH

6. 1. Study Overview

The researchers then employed state of the art principles of artificial neural networks in order to determine when the women attacked would experience a mental breakdown. Through Google Form survey, primary dataset of 690 items and 18 attributes was constructed capturing abuse history, demographics, and psychological factors. Stringent data preparation, exploratory data analysis, encoding, and the training of five machine learning models: Out of all the technique implementations tried and tested, Logistic Regression, Gaussian Naive Bayes, Random Forest Algorithm Classifier, Decision Tree Classifier, and SVM were included. The models' accuracy scores varied: Out of all the models Logistic Regression scored the lowest at 80. 76% while the Gaussian Naïve Bayes had 85. 28%, Random Forest Classifier was at 94. 41%, Decision Tree Classifier 88. 40% and the highest was SVM at 96. 69%. SVM was the most accurate, we therefore recommend that further studies involving battered women should adopt this approach because of the promising results for the prediction of mental health outcomes. SVM yielded the highest accuracies implying the model holds a paradigm position for enhancing mental health prognosis of battered women. The study adds value to knowledge about the multifaceted relationships between child abuse/neglect and psychopathology with relevance to treatment and intervention focus and mechanisms. Some of the important questions that were considered while conducting a research include the following; This is in light of avoiding violation of the participants' privacy, proper handling of the data collected and other ethical issues that are known to occur in research. This pretty and bright color inclusion proving model dedication makes any result laid out easier and accessible. In general, the work receives high evaluation in terms of providing up-to-date information and significantly contributing to improvement of the predictive modelling at the cross-section of female abuse and mental health.

6. 2 Conclusions

Overall, the “Bangladeshi Indigenous Language Classification from Text Analysis” project which presents the identification of the Bangladeshi Indigenous language will assist in the documentation and thus the demonstration and safeguard of linguistic deviation in the region most especially in the Dhaka, Chittagong, Mymensingh zones. This should ensure the identification and classification of indigenous languages through consistent data gathering, rigorous data preprocessing methods, and uses of various ML algorithms, depending on the goals of the project, with the overall vision to help preserve indigenous languages. A comparison and analysis of various models helped to gain more background knowledge about the advantages and disadvantages of different approaches, which served as a strong foundation for further developments within the area of language categorization. Through acknowledgment of ethical element integrated within the project, the project reaffirms commitment of the society to proper use of technology and respect of other individuals. For the language categorisation model, the sustainability plan guarantees that it makes sense and has a significant impact throughout the project. These project will like to continue cooperation, sharing information and strengthening community opportunities of indigenous people regarding Langannajahi’s project to the sustainable development of Bangladeshi language resources.

6.3 Implication for Further Study

Some of the possibilities for further investigation and exploration elucidated by the results and accumulated data from the “Bangladeshi Indigenous Language Classification from Text Analysis” project include: First, and perhaps more importantly, further examination of variations of indigenous languages such as linguistic features and differences might could yield a more nuanced understanding of cultural and linguistic diversity of Bangladesh. These may include, but are not limited to, studies in comprehension and production of certain dialects, language changes over time and use, or how culture influences language.

Additionally, introducing the new technology like neural machine translation or speech recognition into the projects of language and cultural heritage might offer language more opportunities and more ways to be documented and thus give a wider chance for cultural non-utility to be saved. Considering social economic potential of the SAPA, exploring the social economical consequences of language maintenance and the means through which technology assist linguistic groups will be informative in policy formulation and empowering the groups.

REFERENCE

[1]T. Alam, A. Khan, and F. Alam, “Bangla Text Classification using Transformers,” Nov. 2020, doi: https://doi.org/10.48550/arxiv.2011.04446.

[2] Md. A. Hasan, J. Tajrin, S. A. Chowdhury, and F. Alam, “Sentiment Classification in Bangla Textual Content: A Comparative Study,” IEEE Xplore, Dec. 01, 2020. https://ieeexplore.ieee.org/abstract/document/9392681

[3] “Bangla Natural Language Processing: A Comprehensive Analysis of Classical, Machine Learning, and Deep Learning-Based Methods | IEEE Journals & Magazine | IEEE Xplore,” ieeexplore.ieee.org. https://ieeexplore.ieee.org/abstract/document/9751052 (accessed Jan. 07, 2024).

[4]A. Hassan, M. R. Amin, N. Mohammed, and A. K. A. Azad, “Sentiment Analysis on Bangla and Romanized Bangla Text (BRBT) using Deep Recurrent models,” arXiv.org, Nov. 23, 2016. https://arxiv.org/abs/1610.00369 (accessed Jan. 07, 2024).

[5] Das, Amit Kumar, et al. "Bangla hate speech detection on social media using attention-based recurrent neural networks." Journal of Intelligent Systems 30.1 (2021): 578-591.

[6] Barman, Utsab, et al. "Code mixing: A challenge for language identification in the language of social media." Proceedings of the first workshop on computational approaches to code switching. 2014.

[7] Shalini, K., et al. "Sentiment analysis of Indian languages using convolutional neural networks." 2018 International Conference on Computer Communication and Informatics (ICCCI). IEEE, 2018.

[8] Sarkar, Kamal.et al. "Sentiment polarity detection in Bengali tweets using LSTM recurrent neural networks." 2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP). IEEE, 2019.

[9] Raghuveer, et al. "Text Categorization in Indian Languages using Machine Learning Approaches." IICAI. 2007.

[10] Sarkar, Kamal, et al. ``Sentiment polarity detection in bengali tweets using multinomial Naïve Bayes and support vector machines." 2017 IEEE Kolkata Conference (CALCON). IEEE, 2017.

[11] Mokgonyane, Tumisho Billson, et al. "Automatic speaker recognition system based on machine learning algorithms." 2019 Southern African Universities Power Engineering

Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa (SAUPEC/RobMech/PRASA). IEEE, 2019.

[12] Parikh, Pratik, et al. ``English language accent classification and conversion using machine learning." Proceedings of the International Conference on Innovative Computing & Communications (ICICC). 2020.

[13] Kumar, Sujit, et al. "Classification of Indian media titles using deep learning techniques." International Journal of Cognitive Computing in Engineering 3 (2022): 114-123.

[14] Salau, et al. "Accent classification of the three major nigerian indigenous languages using 1d cnn lstm network model." Advances in Computational Intelligence Techniques (2020): 1-16.

[15] Islam, et al. "An evolutionary approach to comparative analysis of detecting Bangla abusive text." Bulletin of Electrical Engineering and Informatics 10.4 (2021): 2163-2169.

[16] Mahtab, et al. ``Sentiment analysis on bangladesh cricket with support vector machine." 2018 international conference on Bangla speech and language processing (ICBSLP). IEEE, 2018.

[17] Tripto, et al. ``Detecting multilabel sentiment and emotions from bangla youtube comments." 2018 International Conference on Bangla Speech and Language Processing (ICBSLP). IEEE, 2018.

[18] Islam, Md Saiful, et al. "Supervised approach of sentimentality extraction from bengali facebook status." 2016 19th international conference on computer and information technology (ICCIT). IEEE, 2016.

[19] Alam, et al. "Bangla text classification using transformers." arXiv preprint arXiv:2011.04446 (2020).

[20] Hassan, Asif, et al. "Sentiment analysis on bangla and romanized bangla text (BRBT) using deep recurrent models." arXiv preprint arXiv:1610.00369 (2016).

[21] Ahmed, Md Faisal, et al. "Cyberbullying detection using deep neural networks from social media comments in Bangla language." arXiv preprint arXiv:2106.04506 (2021).

[22] Chowdhury, Rumman Rashid, et al. "Analyzing sentiment of movie reviews in bangla by applying machine learning techniques." 2019 international conference on Bangla speech and language processing (ICBSLP). IEEE, 2019.

[23] Banik, et al. "Toxicity detection on bengali social media comments using supervised models." 2019 2nd International Conference on Innovation in Engineering and Technology (ICIET). IEEE, 2019.

[24] Das, et al. "Identifying languages at the word level in code-mixed indian social media text." (2014).