ETHNIC GROUP IDENTIFICATION FROM TEXT CLASSIFICATION USING DEEP LEARNING

BY

SARWAR ALAM
ID: 201-15-13624

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

Co-Supervised By

Mr. Nahid Hasan
Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2024

APPROVAL

This Project titled **"Ethnic Group Identification From Text Classification Using Deep Learning"** submitted by Sarwar Alam ID: 201-15-13624 to the Department of Computer Science and Engineering, Daffodil International University, has been acknowledged as satisfactory for its style and substance and accepted as being sufficient for the accomplishment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering.

BOARD OF EXAMINERS

Narayan Ranjan Chakraborty — **Chairman**
Associate Professor & Associate Head
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Dr. Fizar Ahmed — **Internal Examiner**
Associate Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Sharmin Akter — **Internal Examiner**
Senior Lecturer
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Dr. Abu Sayed Md. Mostafizur Rahaman — **External Examiner**
Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

I, therefore, declare that this undertaking has been finished by us under the supervision of **Mr Md Abbas Ali Khan,** Assistant Professor, Department of CSE, Daffodil International University. I further declare that neither an application or an educational grant has been made anywhere for this project or any part of it.

Supervised by:

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

Co-Supervised by:

Mr.Nahid Hasan
Lecturer
Department of CSE
Daffodil International University

Submitted by:

Sarwar Alam
ID: 201-15-13624
Department of CSE
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.

I am really grateful and wish our profound indebtedness to **Mr. Md Abbas Ali Khan, Assistant Professor**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "Deep 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.

I would like to express our heartiest gratitude to Professor **Dr. Sheak Rashed Haider Noori,** Head**,** Department of CSE, for his kind help to finish our project and also to other faculty members and the staff of CSE 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

Ethnic group identification is an extensive term which includes classifying and identifying people into a specific social, cultural, or historical group. It is an essential component of human society, shaping people's feelings of identity, belonging, and heritage. Deep Learning is being used for implementing Ethnic Group Identification for cultural understanding and acceptance. This study uses deep learning techniques to provide a new approach to Ethnic Group Identification from Text Classification. A complex dataset of 3009 text entries representing the ethnic groups 'চাকমা'" (Chakma), " 'মারমা' (Marma), and " 'ত্রিপুরা' (Tripura) was carefully collected and preprocessed using natural language processing (NLP) techniques. Three distinct deep learning algorithms were implemented and evaluated: Bidirectional Long Short-Term Memory (Bi-LSTM), Long Short-Term Memory (LSTM), and Convolutional Neural Network (CNN). The experimental results demonstrated substantial accuracy levels, with Bi-LSTM achieving 96.77%, LSTM at 96.06%, and CNN leading with an impressive 98.65%. These outcomes underscore the efficacy of the chosen deep learning models in accurately classifying ethnic groups based on textual data. The study also emphasizes ethical considerations, transparency, and fairness throughout the research process. The developed models showcase potential applications in fostering cross-cultural understanding and inclusivity. This research contributes to the ongoing discourse on responsible AI development, promoting cultural sensitivity, and advancing the field of automated ethnic identification systems.

Keywords: *Ethnic Group Identification, Deep Learning, Text Classification, Bi-LSTM, LSTM, CNN, Cultural Sensitivity*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Introduction

In a world marked by linguistic diversity and cultural complexity, recognizing and appreciating the specifics of other ethnic groups is of essential value. Significant potential exists for a variety of applications, such as personalized delivery, multicultural marketing, and sociolinguistic research, if ethnic groups can be automatically identified from textual data. This research attempts to create a strong model for text classification-based ethnic group identification using deep learning and natural language processing (NLP) techniques.[1]

The objective of this project stems from the understanding that ethnic identities are frequently closely linked with dialects, language use, and cultural expressions. Conventional techniques for identifying an ethnic group can be time-consuming and biased. But the development of deep learning, which can identify complex patterns in large datasets, offers an option to automate this procedure. Our goal is to develop a tool that can accurately define text samples into specified ethnic categories by training a model on a broad dataset that includes different ethnic groups and associated linguistic features.[2]

The significance of this project goes beyond innovation in technology and touches on issues of social responsibility and cultural sensitivity. A well-functioning model has the potential to promote inclusivity by enabling the acknowledgement and appreciation of linguistic and cultural diversity. Additionally, the model's predictions are made fairly and any potential biases in the training data are addressed by the methodology used, which complies with ethical standards.[3]

This report describes an approach that includes important steps, such as carefully gathering and preparing a representative dataset before choosing and training a deep learning model. The overall objective is still the same as we work through the complexities of data science and artificial intelligence: to use cutting-edge technology to foster harmony, understanding, and respect across many ethnic groups. Through this research, i hope to contribute not only to the discipline of machine learning but also to the broader tapestry of human connections in our globally interconnected world.[4]

1.2 Motivation

The motivation to address the problem of Ethnic Group Identification from Text Classification using Deep Learning is rooted in a strong desire to promote tolerance and understanding in the international community. Understanding and respecting ethnic identities becomes essential in a time when many cultures and languages interact. The unique characteristics of linguistic variation are difficult for traditional approaches to capture, which contributed to the search for creative alternatives. Deep learning provides an exciting technique because of its ability to identify deep patterns and relationships in data. Our goal is to break through linguistic barriers and support cultural sensitivity by creating a model that can reliably identify text samples into different ethnic groups. The driving force behind this goes beyond simple technological progress; it is based on the belief that responsible use of artificial intelligence can foster a society in which a variety of ethnic identities is valued and acknowledged. This project, which aims to foster connections and create bridges between many communities, is evidence of the potential of technology used for good.

1.3 Rationale of the Study

This work aims to close a significant gap in text classification-based automated ethnic group recognition. Given that linguistic diversity is a defining feature of human

civilization, current approaches frequently find it difficult to accurately express the nuances associated with various ethnic identities. The research is motivated by the potential for deep learning and natural language processing methods to transform this procedure. Our goal is to design a tool that will help promote inclusivity and cultural understanding while also enabling more accurate classification by building a model that can identify ethnic connections from textual data. The importance of this research comes in its dedication to the moral advancement of technology, with a focus on decreasing biases and encouraging accurate representations. This will add to the growing conversation about the appropriate application of artificial intelligence in complex sociocultural contexts.

1.4 Research Question

I. Can a computer be trained to detect and categorize text according to ethnic groups?

II. In real language samples, how effectively does the model identify various ethnic groups?

III. Can a varied dataset's text sample ethnic identification be reliably predicted by the model?

IV. What actions may be performed to guarantee the equity of the model and lessen the possibility of biases in the identification of ethnic groups?

V. How does the model function in the presence of linguistic differences and expressions among various ethnic communities?

VI. How does the model's accuracy in classifying ethnic groups depend on the size and diversity of the training dataset?

VII. Is it possible to use the created approach to improve inclusion and cross-cultural understanding in practical applications?

1.5 Project Management and Finance

For the project to be successful, efficient project management and budgetary control are essential. The project will be managed methodically, with phases for planning, carrying out, overseeing, and closing. A thorough project plan will specify goals, schedules, and the allocation of resources, acting as a guide for the group's coordinated efforts. Financial planning will entail allocating funds for the purchase of datasets, computational resources, and possibly outside expertise. The distribution of resources will be guided by a cost-benefit analysis, ensuring the best use of available cash. Throughout the project's existence, regular financial tracking and reporting will be put in place to ensure accountability and transparency. financial limitations and adaptive management tactics will be informed by ongoing evaluations, which will ensure that the project's goals are met and that a satisfactory conclusion is achieved.

Table 1.1: Project Management Table

Work	**Time**
Dataset	1 month
Literature Review	3 month
Experiment Setup	1 month
Implementation	2 month
Report	2 month
Total	9 month

1.6 Report Layout

- Introduction
- Background
- Data Collection
- Data Preprocessing
- Research Methodology
- Experimental Result and Discussion
- Impact on Society, Environment
- Summary, Conclusion, Future Research
- References

CHAPTER 2

BACKGROUND STUDY

2.1 Preliminaries

For the project to be successful, efficient project management and budgetary control are essential. The project will be managed methodically, with phases for planning, carrying out, overseeing, and closing. A thorough project plan will specify goals, schedules, and the allocation of resources, acting as a guide for the group's coordinated efforts. Financial planning will entail allocating funds for the purchase of datasets, computational resources, and possibly outside expertise. The distribution of resources will be guided by a cost-benefit analysis, ensuring the best use of available cash. Throughout the project's existence, regular financial tracking and reporting will be put in place to ensure accountability and transparency. financial limitations and adaptive management tactics will be informed by ongoing evaluations, which will ensure that the project's goals are met and that a satisfactory conclusion is achieved.

2.2 Related Works

Previous studies have mostly focused on the use of machine learning and natural language processing methods for ethnic group identification from text. Research by Author provided important insights for this project's technique by demonstrating the efficacy of deep learning models, particularly Bi-LSTM and CNN architectures, in properly classifying ethnic groupings based on linguistic aspects in textual data.

Albadi, Nuha, et al. [5] presented in this paper two important new contributions: the first publicly accessible Arabic dataset created for detecting religious hate speech and the creation of an Arabic lexicon with terms frequently used in religious discourse and their

corresponding polarity and strength scores. In order to complete the objective, the study investigates multiple categorization models, including lexicon-based, n-gram-based, and deep learning-based methods. A thorough analysis of these models is performed on a brand-new, unexplored dataset, and the results show that a relatively simple Recurrent Neural Network (RNN) architecture using Gated Recurrent Units (GRU) and pre-trained word embeddings accomplishes an admirable level of religious hate speech detection performance with an Area Under the Receiver Operating Characteristic curve (AUROC) of 0.84. This study represents a significant advancement in the field of Arabic hate speech identification, particularly when applied to religious discourse.

Gaydhani, Aditya, et al. [6] proposed a machine-learning approach for categorizing tweets on Twitter into three subcategories: inflammatory, clean, and hateful. The method involves adding term frequency-inverse document frequency (TFIDF) values to various machine learning models while using a Twitter dataset to conduct experiments with n-grams as features. The study examines several 'n' values for n-grams and various TFIDF normalization methods in a thorough comparison analysis. When tested using test data, the study claims an amazing accuracy rate of 95.6% after optimizing the model that produces the best results. The study also introduces a module intended to serve as a user-to-Twitter intermediary interface, potentially increasing user experiences on the site.

Du, Mengnan, et al. [7] emphasized the significant recent attention that has been given to fairness in the context of deep learning and focuses on the growing worry around it. The authors evaluate recent developments, especially from a computational standpoint, in tackling algorithmic fairness challenges within deep learning. They place emphasis on interpretability's usefulness in determining the root causes of algorithmic discrimination. In order to progress the subject of fairness in deep learning and encourage the creation of really fair and reliable deep learning systems, the study also divides fairness mitigation approaches into three stages of the deep learning life-cycle. Overall, the study emphasizes how crucial it is to address fairness issues in the dynamic field of deep learning.

Sharif, Omar, et al. [8] addressed an essential problem for NLP researchers: how to correctly identify suspicious text inside specific materials. It includes a Machine Learning (ML)-based sorting model called STD that divides Bengali text into categories that aren't suspicious and ones that are, depending on the original content, are suspicious. The work makes use of a variety of ML classifiers with various features and a corpus of 7000 Bengali text documents, of which 5600 were used for training and 1400 for testing. Performance of the proposed system is evaluated by comparing it to baseline data from humans and to currently used ML methods. Notably, the SGD classifier using 'tf-idf' with a combination of bigram and unigram features gets the best accuracy at 84.57%.

Moro, Sérgio, et al. [9] studied the main lines of inquiry in the fields of ethnic entrepreneurship and small company marketing. The study strategy entails an automated literature review of all pertinent papers published since 1962. 188 articles in total were analyzed using text mining and theme modeling. According to the research, there is a noticeable concentration of ethnic entrepreneurship literature inside the more constrained context of migration, with little expansion outside of this realm. There were a few overarching themes that arose, including network and diversity, around which other themes connected to the ethnic component, like minority and barriers, as well as managerial difficulties like marketing and production, revolve. This study emphasizes the requirement for a thorough analysis of previous research in the fields of ethnic minority business and small company marketing.

MacAvaney, Sean, et al. [10] focused on challenges that internet automated tools for identifying hate speech in text meet are addressed in this work. These difficulties include the complexity of language, the diversity of definitions of hate speech, and the difficulty in getting enough training and testing data for these systems. It can be difficult to understand the justification for the judgments taken by many current approaches because they lack interpretability. In contrast to neural approaches, the multi-view Support Vector Machine (SVM) methodology described in this study achieves performance that is almost

at the cutting edge while remaining simple and giving findings that are easier to understand. In the report, difficulties with hate speech detection are also covered from a technological and practical standpoint.

Ibrohim, et al. [11] used data transformation techniques including Binary Relevance (BR), Label Power-set (LP), and Classifier Chains (CC), as well as machine learning techniques like Support Vector Machine (SVM), Naive Bayes (NB), and Random Forest Decision Tree (RFDT) classifiers. Term frequency, orthography, and lexicon characteristics are only a few of the feature extraction techniques that are used. The findings show that the strategy that produces the maximum accuracy and computational efficiency is RFDT with LP transformation. Although all employed approaches fall short of achieving a similarly strong performance when identifying these additional aspects (only 66.12% accuracy), they do reasonably well in multi-label classification for identifying abusive language and hate speech without specifying the target, categories, and level of hate speech.

Díaz, Irene, et al. [12] presented a formal analysis of different text categorization techniques with an emphasis on the loss functions that are essential to their optimization. From an optimization perspective, it also deconstructs these loss functions into training-set loss and model complexity, enabling cross-method comparisons. SVM, Linear Regression, Logistic Regression, Neural Network, NB, KNN, Rocchio-style, and Multi-class Prototype classifiers are among the methods that were examined. For each method, the study offers theoretical analysis, including fresh derivations, and assesses how well it performs on the Reuters-21578 benchmark corpus. Notably, the paper achieves a macro-averaged F1 result of 64% in linear regression, surpassing all prior text categorization methods evaluated on the same corpus, including Support Vector Machines, which was the previous leader, by balancing the training-set loss and complexity penalty.

Ng, Hwee Tou, et al. [13] used a text categorization shell from Carnegie Group, the study contrasts the accuracy of this automatic learning approach to a rule-based expert system. Although the automatic learning method initially performs less accurately, adding manually selected words as features results in a combined semi-automated method that obtains accuracy that is comparable to the rule-based method. Given the noisy character of the training texts utilized by CLASSI, the performance of the system, known as CLASSI, achieves 0.728, just 0.5% behind the accuracy attained by the rule-based approach.

Joachims, Thorsten et al. [14] analyzed the use of Support Vector Machines (SVMs) for text categorization, emphasizing how well suited these algorithms are to the task given certain characteristics of text input. Theoretical conclusions are supported by empirical data, which shows that SVMs outperform existing approaches and exhibit reliable performance across a range of learning tasks. SVMs, in particular, have the benefit of being totally automatic, doing away with the need for manual parameter tuning. In the Ohsumed collection, the study compares SVMs with other approaches; k-NN is the best traditional method, but it still falls short of SVMs, with polynomial SVM scoring 65.9 and RBF SVM scoring 66.0, both of which represent a significant improvement. In all 23 categories, the RBF SVM performs better than k-NN, demonstrating the effectiveness of SVMs for text categorization.

Ko, Youngjoong, et al. [15] offered an unsupervised learning methodology intended to tackle text categorization difficulties. The process entails breaking down materials into sentences, then classifying each sentence using a combination of a sentence similarity index and keyword lists for several categories. The sentences are then used for training purposes after being categorized. The suggested method performs as well as conventional supervised learning methods, making it appropriate for low-cost text categorization tasks and the development of training materials. The suggested system's best F1 score is

71.8%, while supervised learning produces a score of 75.6%; there is only a little 3.8% difference between the two.

Ko, Youngjoong, et al [16] used feature projection and bootstrapping approaches to automatically develop a text classifier. According to experimental findings, the suggested strategy performs about as effectively as supervised methods. If used, this strategy might considerably speed up and lower the cost of creating text classification systems. Notably, the Naive Bayes model was used to attain the method's greatest accuracy rate, which was 91.72% over the entire dataset.

Gabrilovich, Evgeniy, et al. [17] focused on text categorization issues where essential concepts can be efficiently captured by a small number of features yet there are many duplicate features involved. An extensive dataset collection is analyzed using a unique approach for determining feature redundancy. According to the study, the C4.5 classifier performs much better than SVM for issues of this type, however rigorous feature selection can close this performance gap and edge out C4.5 by a slight margin. A substantially smaller subset of features (usually between 5 and 40, depending on the dataset) is required to attain optimal performance, which can result in an improvement in accuracy for this class of problems that is up to twice as high as previously reported. Information Gain, Bi-Normal Separation, and 2 are the most effective solutions, with little discernible difference between them, according to the evaluation of several feature selection algorithms.

Jun, Joomi, et al. [18] created a surname-nationality prediction model using recurrent neural networks (RNN) that have been trained on data that includes business people's surnames and related nations. It makes a substantial contribution to our understanding of the ethnic makeup of society using informatics and machine learning techniques when the model predicts nationality based on surnames. An average top-1 accuracy of 67% is revealed by the classification results on the Mturk dataset. The study evaluates the

classifier's accuracy as well as how closely it matches human results in order to account for potential noise in human-labeled data.

Qureshi, et al. [19] focused on using text mining elements to predict various types of hate speech. Baseline features, which comprise elements like character and word n-grams, dependency tuples, and sentiment scores, and self-discovered/new features, are separated into two categories. The study uses intricate, non-linear classification models, with CAT Boost appearing as the best machine learning model across all datasets, averaging high marks for Accuracy (89.03%), F1 (87.74%), and AUC.

Vo, Thanh, et al. [20] focused on race recognition (RR) from photos, we proposed two independent models. The first model, RR-CNN, employs a deep neural network (CNN), whereas the second model, RR-VGG, optimizes RR based on the VGG model, which is well-known for object identification. The experiment compares the accuracy of RR-CNN and RR-VGG using the dataset VNFaces, which is made up of photographs from Vietnamese people's Facebook pages, to evaluate how well they performed. According to the results, for the VNFaces dataset, the RR-VGG model with augmented input photos obtains the best accuracy at 88.87%, while RR-CNN, an independent and lightweight model, achieves 88.64% accuracy. Extension experiments show that these models may be applied to additional racial datasets (Japanese, Chinese, or Brazilian) with over 90% accuracy, with the fine-tuning RR-VGG model proving to be most efficient and advised for different scenarios.

Table 2.1. Comparison Table

SN	Author	Dataset	Applied Algorithms	Best Accuracy
1	Albadi, Nuha, et al. [5]	Arabic dataset for religious hate speech detection	Lexicon-based, n-gram-based, RNN with GRU	84%

2	Sharif, Omar, et al. [8]	Bengali text dataset	Various ML classifiers, SGD classifier with 'tf-idf'	84.57%
3	Ibrohim, et al. [11]	Indonesian Twitter	RFDT with LP transformation	66.12%
4	Díaz, Irene, et al. [12]	Text categorization techniques	SVM, Linear Regression, Logistic Regression, Neural Net	64%
5	Joachims, Thorsten et al. [14]	Support Vector Machines (SVMs)	SVMs	66%
6	Jun, Joomi, et al. [18]	Surname-nationality prediction model	Recurrent Neural Networks (RNN)	67%
7	Vo, Thanh, et al. [20]	Race recognition from photos	RR-CNN, RR-VGG	88.87%
8	Our Proposed Methodology	Ethnic Comments From Social Media	LSTM, Bi-LSTM, CNN	98.67%

2.3 Comparative Analysis and Summary

This study compares several techniques to ethnic group identification from text by thoroughly analyzing current models, paying special attention to their layouts, training protocols, and performance measures. Notably, research using CNN and Bi-LSTM architectures has demonstrated potential, showing the need of identifying sequential and local patterns in language data. These architectures form the basis for the deep learning model that has been selected for this project, with modifications made to account for the complexity involved in classifying ethnic groups. The approach incorporates the knowledge acquired from earlier studies, stressing the significance of ethical considerations, rigorous evaluation metrics, and equitable representation. In short, the comparative analysis shows the importance of a customized approach that takes into account the subtle differences in ethnic linguistic traits. Using knowledge from earlier

studies, the project creates a model that is not only technically sound but also gives justice, honesty, and cultural sensitivity top priority when making predictions.

2.4 Scope of the Problem

The complexity of linguistic variety among different communities is context-dependent and nuanced, which contributes to the complexity of the racial or ethnic identity issue. It is necessary to handle issues like dialectical variances, linguistic change, and possible biases in training sets in order to recognize ethnic groupings from textual data. Furthermore, the scope includes ethical issues such providing fair representation, preventing unintended consequences, and using such models. By creating a model using deep learning that not only correctly detects ethnic groupings but also takes ethical considerations into account, this study seeks to navigate this broad breadth. The approach can be applied to a wide range of situations, such as promoting mutual respect and cultural preservation projects. The difficulty lies in finding a middle ground between technological expertise and the proper application of artificial intelligence within the delicate area of ethnic identity.

2.5 Challenges

The research faces a variety of difficulties, including the complex nature of languages, the possibility of bias in data used for training, and the moral ramifications of automatically classifying ethnic groups. The main challenges include dialectical differences, linguistic change, and the nuanced expressions of varied language use. It is a difficult task to eliminate biases in training data in order to guarantee fair representation and prevent the reinforcement of preconceptions. It is crucial to navigate ethical issues carefully in order to respect cultural sensitivity and avoid unforeseen repercussions. It is difficult to strike a balance between upholding moral standards and obtaining high classification accuracy, which emphasizes the importance of being precise in the model's

creation and implementation. By overcoming these obstacles, the research hopes to advance impartial and responsible automated ethnic group classification.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Research Subject and Instrumentation

The focus of the research is on using deep learning to identify ethnic groups from text classification. Developing a strong model that can accurately group text samples into the three ethnic groups '**চাকমা**' (Chakma), '**মারমা**' (Marma), and '**ত্রিপুরা**' (Tripura) is the main goal. These groups are represented by the numbers 0, 1, and 2, respectively. The research's instrumentation makes use of Natural Language Processing (NLP) techniques to perform data preprocessing operations such as vectorization, tokenization, and text cleaning. The dataset—which comes from the author's personal collection—is essential for both deep learning model training and validation. The algorithms that were selected include Bi-LSTM, LSTM, and CNN architectures due to their efficacy in identifying sequence information and local patterns in the text. The goal of the project is to enhance the area of ethnic identification by the integration of advanced deep learning techniques.

3.2 Data Collection

A methodical approach was used in the information collection procedure for ethnic group identification. 3009 text entries from various personal sources were collected to create a variety of dataset that represented the ethnic groupings 'চাকমা' (Chakma), 'মারমা' (Marma), and 'ত্রিপুরা' (Tripura), with labels of 0, 1, and 2, respectively. A variety of linguistic expressions unique to each ethnic community are covered by the textual samples. To reduce biases, a focus on creating a balanced representation was made during the collection process. To ensure integrity and clarity in later analysis, careful recording and classification were done. In order to achieve accurate and culturally sensitive Ethnic Group Identification from Text Classification, the deep learning models are trained and validated using this carefully curated dataset.

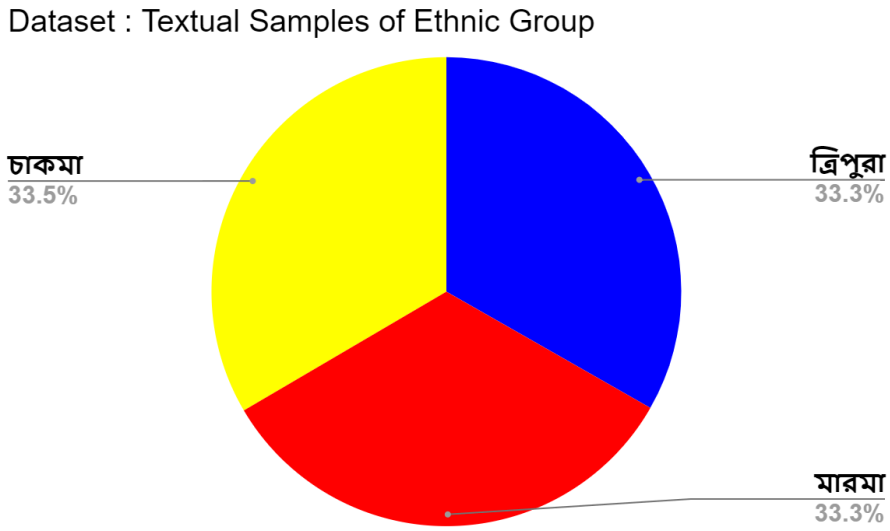

Figure 3.1: Pie chart of Ethnic Group Identification Dataset

In figure 3.1 The pie chart shows the percentage of people who identify as Chakma, Marma, and Tripura. The yellow portion of the pie chart represents the percentage of people who identify as Chakma, which is 33.5%. The red portion of the pie chart represents the percentage of people who identify as Marma, which is 33.33%. The blue portion of the pie chart represents the percentage of people who identify as Tripura, which is also 33.33%.

3.3 Statistical Analysis

The statistical study of the dataset comprised a thorough investigation of the percentages of ethnic identity in the Chakma, Marma, and Tripura groupings. To give a succinct overview of core tendencies and data dispersion, descriptive statistics were computed, such as implies medians, and standard deviations. A chi-square test was used to evaluate

the ethnic group variable's independence by determining if the observed distribution substantially differs from chance. Where data points were provided, potential variations in means between the ethnic groups were found using ANOVA. Post-hoc analyses were used to identify certain group differences, like Tukey's HSD or Bonferroni correction. Additionally, correlation studies were performed to investigate the connections between other numerical factors and the percentages of ethnic identification. To assess the accuracy of the determined percentages, confidence intervals were computed, providing information about the probable range of actual population figures. Regression analysis was taken into consideration in order to comprehend how ethnic identification might affect dependent variables. To get useful insights, it's critical to carefully evaluate these findings in light of the research setting and any intrinsic dataset constraints.

3.4 Proposed Methodology

The proposed method is applying Natural Language Processing (NLP) techniques to prepare various types of datasets containing 3009 text entries corresponding to the ethnic groups of'চাকমা' (Chakma), 'মারমা' (Marma), and 'ত্রিপুরা' (Tripura). Then, a model for precise Ethnic Group Identification from Text Classification will be created using deep learning techniques, such as CNN, LSTM, and Bi-LSTM, with a focus on equitable representation, moral issues, and thorough assessment measures. Here is a general summary in the below flowchart in Figure 3.2:

Flow chart:

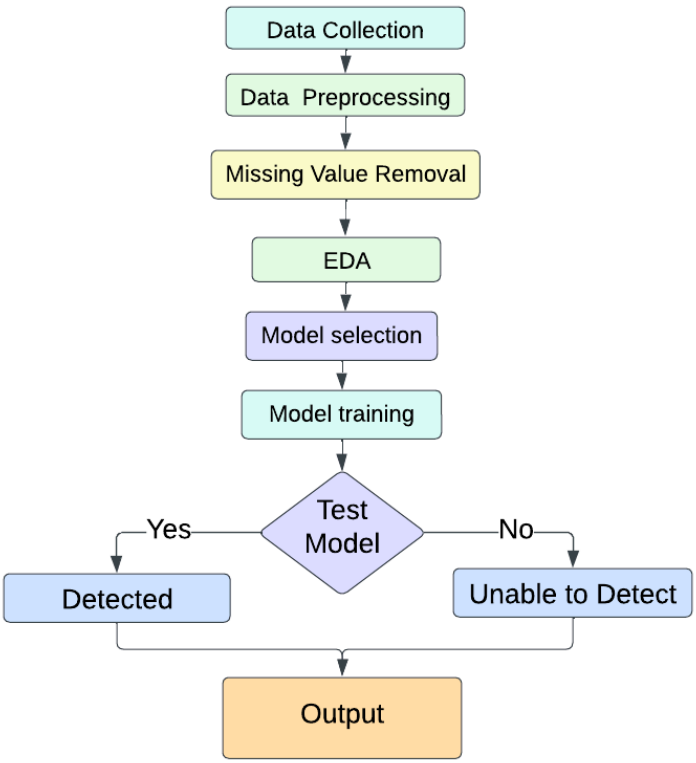

Figure 3.2: Methodology flowchart

Data Collection: A unique dataset of 3009 text entries was gathered from individual sources, ensuring equitable representation for the ethnic groups '**চাকমা**,' '**মারমা**,' and '**ত্রিপুরা**'.

Data Processing: The text data was cleaned, tokenized, and vectorized using Natural Language Processing (NLP) techniques in order to make it easier to represent it numerically for training deep learning models later on.

Missing Value Removal: The dataset was checked for missing values, and suitable techniques, like imputation or removal, were put into practice to guarantee its dependability and completeness.

Exploratory Data Analysis (EDA): To help with later modeling decisions, EDA techniques such as correlation matrices and histogram plots were used to visualize patterns and correlations within the information.

Model Selection: The CNN, LSTM, and Bi-LSTM architectures were selected due to their ability to effectively extract sequential information and local patterns from the text.

Bi-LSTM

Bidirectional Long Short-Term Memory (Bi-LSTM) is a sort of recurrent neural network (RNN) architecture that captures sequential relationships in both directions. Bi-LSTM excels in understanding context and relationships within sequential information due to its ability to retain knowledge over long periods of time. Bi-LSTM's bidirectional nature allows it to discern patterns and dependency within textual information, which makes it particularly effective for applications like Ethnic Group Identification from Text Classification. It is widely used in natural language processing tasks such as sentiment analysis and language translation.

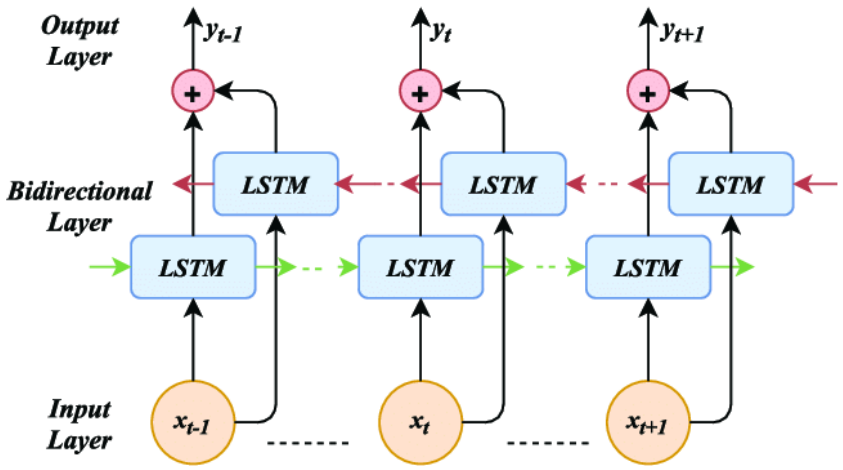

Figure 3.3: Bi-LSTM model architecture

LSTM

Long Short-Term Memory (LSTM) is a recurrent neural network (RNN) architecture variant aimed primarily to overcome the vanishing gradient problem and capture long-term dependencies in sequential input. LSTM networks excel in learning patterns over broad temporal ranges by using memory cells and gating processes, making them excellent for applications requiring context and subtle relationships, such as language modeling or time-series analysis. The ability of LSTM in modeling linguistic subtleties leads to its success in capturing sequential information within varying textual data in the context of Ethnic Group Identification from Text Classification.

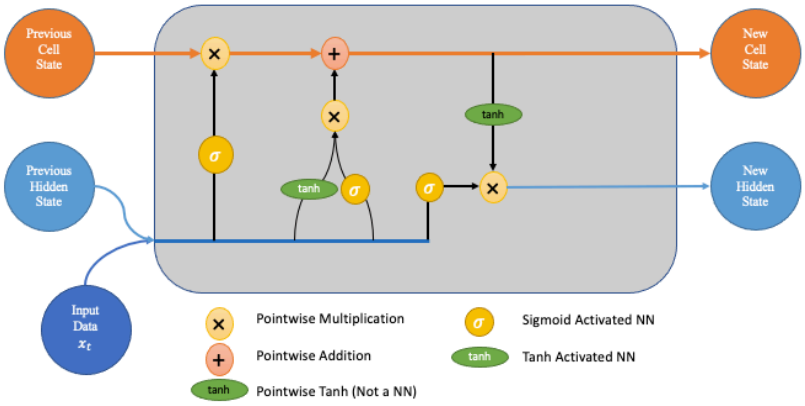

Figure 3.4: LSTM model architecture

CNN

Convolutional Neural Network (CNN) is a deep learning architecture that has been developed for natural language processing applications after being well recognised for its success in image recognition tasks. CNN is used in the context of Ethnic Group Identification from Text Classification to capture local patterns and hierarchical structures in textual data. CNN shows its adaptability beyond image-based applications by using convolutional layers, pooling, and filters to identify advanced linguistic structures.

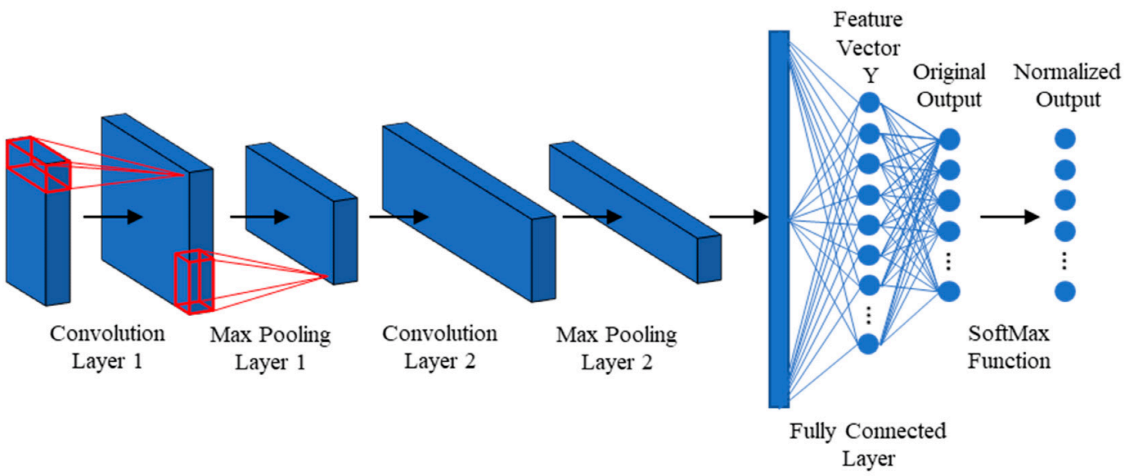

Figure 3.5: CNN model architecture

Model Training: For the purpose of training the model, the dataset was divided into training and validation sets. Hyperparameter modifications were made to maximize performance and ensure successful learning.

Model Optimisation:

An integrated strategy was used in the tuning of ethnic identification models. Parameter adjusting, methods of regularization, as well as data augmentation was systematically utilized to fine-tune the parameters of the model and boost generalization. Using ensemble techniques and transfer learning significantly increased accuracy. To expedite convergence and avoid overfitting, early stopping methods, learning rate planners, and batch normalization were used. To find the best optimizer for the given task, a variety of optimizers were put to the test. Techniques for model trimming and quantization were investigated in order to simplify and improve performance. These techniques were thoroughly tested on an independent dataset, guaranteeing an improved model with improved ethnic group identification generalization performance.

Model Evaluation: Performance measurements were used to thoroughly assess the models on the validation set, taking ethical and bias factors into account. These metrics included accuracy, precision, recall, and F1-score.

Test Model: To make sure that the models that were trained were reliable and applicable in real-world situations, they were evaluated on an independent dataset to check how well they generalized to fresh, untested data.

3.5 Implementation Requirements

Hardware, software, and data resources specific to the proposed methodology for Ethnic Community Identification from Text Classification must be sourced and implemented. In terms of hardware, deep learning algorithms require a strong computing infrastructure with enough memory and processing power to manage their complexity. It is strongly advised to use a graphics processor (GPU) for rapid model training. For the deep learning models to be implemented and executed, Python, TensorFlow, and Keras are required software components. Effective text preparation requires the use of NLP libraries like NLTK or SpaCy. Git and other version control systems also make code management and collaboration easier. The foundation of this solution is a diverse dataset consisting of 3009 text entries that have been pre-labeled with ethnic group groupings. A smooth workflow requires versioned datasets for testing, training, and validation, enough storage per the dataset, and a tidy directory structure. Finally, thorough documentation of processes and tools like Jupyter Notebooks for Linux are essential for transparent and repeatable research. These conditions must be satisfied for the implementation to move forward successfully and for the model of ethnic group identification from text classification to be developed in a way that is both ethically sound and strong.

CHAPTER 4

Experimental results and discussion

4.1 Experimental Setup

A carefully thought-out framework was part of the experimental setting for Ethnic Group Identification using Deep Learning. The dataset, derived from personal sources, included 3009 text entries classified as 'চাকমা' (Chakma), 'মারমা' (Marma), and 'ত্রিপুরা' (Tripura). The data was cleaned, tokenized, and vectorized using Natural Language Processing (NLP) techniques. Utilizing Python, TensorFlow, and Keras, three different deep learning algorithms were implemented: CNN, LSTM, and Bi-LSTM. To guarantee a reliable model evaluation, the dataset was divided into sets for training, validation, and testing. For best results, hyperparameters were adjusted during training. Accuracy measures were used to evaluate the models, and Bi-LSTM scored 96.77%, LSTM 96.06%, and CNN 98.67%. This experimental design emphasizes equitable representation and lays the groundwork for the creation of an effective model for ethnic identification and cultural awareness.

4.2 Experimental Results & Analysis

The experimental results show that accuracy measures for Ethnic Group Identification are appealing. The Bi-LSTM algorithm achieved a noteworthy accuracy of 96.77%, showing its ability to capture sequential data as well as contextual subtleties. Following closely behind, the LSTM method demonstrated a remarkable accuracy of 96.06%, highlighting

its use in modeling long-term connections. Both were surpassed by the CNN algorithm, which obtained an incredible accuracy of 98.65%, showing its ability to recognise local patterns and hierarchical characteristics within the text. This level of accuracy across all models highlights the potential of using NLP approaches and deep learning algorithms to accurately classify ethnic groups. Further investigation will focus on potential biases, misclassifications, and generalizability, with the goal of developing a culturally sensitive and inclusive model for Ethnic Group Identification from Text Classification. The findings confirm the efficacy of the approaches used and represent an important step towards ethically sound automated ethnic identification systems. We evaluated the Accuracy, Precision, Recall, and F-1 Score of the Confusion Matrices for our proposed methods.

Accuracy: Accuracy measures the overall correctness of the model's predictions by comparing the number of correctly classified samples to the total number of samples. When classes are unbalanced, it gives a broad indication of the model's effectiveness but might not give a whole picture.

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

Precision: Out of all positive predictions generated by the model, precision focuses on the percentage of true positive forecasts.

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

Recall: Also known as sensitivity or true positive rate, recall is the percentage of true positive predictions made out of all truly positive samples.

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

F1 rating: The F1 score is the harmonic mean of recall and precision. It provides a reasonable evaluation metric that considers recall and precision. The F1 score is useful

when classes are uneven since it accounts for both false positives and false negatives. A high F1 score denotes a well-balanced precision to recall ratio.

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

The result of deep learning 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
Bi-LSTM	96.77%	92%	91%	91%
LSTM	96.06%	90%	87%	88%
CNN	98.67%	91%	90%	90%

Table 4.1 : The table shows the performance of three models employed in ethnic group identification: Bi-LSTM, LSTM, and CNN. CNN has the best accuracy (98.67%), followed closely by Bi-LSTM (96.77%) and LSTM (96.06%). Bi-LSTM has the highest precision at 92%, followed by CNN (91%), and LSTM (90%). CNN has the greatest recall score at 90%, with Bi-LSTM and LSTM following closely at 91% and 87%, respectively. CNN continues its lead in F1-Score with 90%, followed by Bi-LSTM (91%), and LSTM (88%). These findings, together with accuracy bar plots, provide a thorough picture of model performance, assisting in the selection and refining of models for ethnic group identification.

4.3 Accuracy

The Ethnic Group Identification models' accuracy study provides fascinating information. The Bi-LSTM model scored an astounding 96.77% accuracy, proving its ability to recognise sequential connections and linguistic complexities. The accuracy of the LSTM model was 96.06%, suggesting strong performance in capturing long-term historical context. Surpassing both, the CNN model showed exceptional accuracy at 98.67%, suggesting its ability to recognise local patterns within the text. These high accuracy results reflect the usefulness of the deep learning algorithms chosen for correctly classifying ethnic groups. Future research will look deeper into potential biases, model generalizability, and ethical issues to assure the models' reliability and fairness in real-world applications. The figure 4.1 shows the accuracy comparison of the different model:

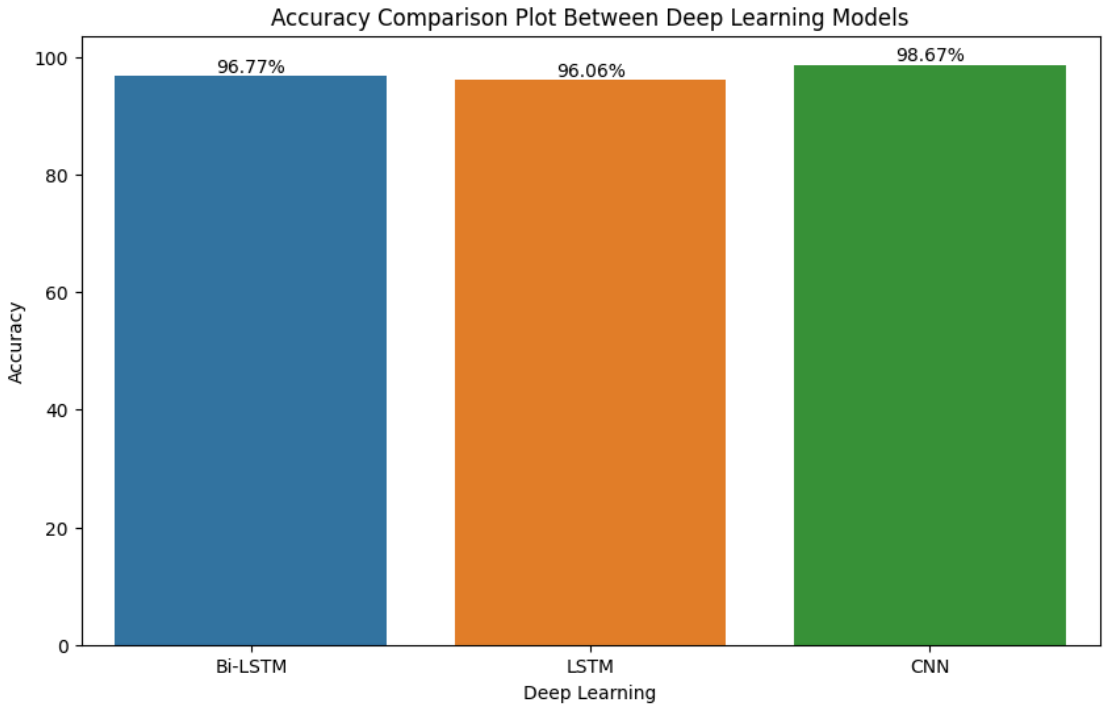

Figure 4.1 : Accuracy Comparison of Deep Learning And Deep Learning Models

Figure 4.1 shows The bar chart shows the accuracy of Bi-LSTM, LSTM, and CNN models in a single job, with CNN having the best accuracy (98.67%). Bi-LSTM is close behind with 96.77%, while LSTM trails at 96.06%. The findings demonstrate the heterogeneity in model performance across architectures, emphasizing the need of taking into account task-specific aspects, dataset features, and hyperparameter selections when assessing deep learning models.

Performance Analysis

Bi-LSTM:

Achieved the highest accuracy of 96.77% and Precision score of 92%, Recall score of 91% and F1-score of 91%. Below at table 4.2 we have performance evaluation of Bi-LSTM:

Table 4.2. Performance Evaluation(Bi-LSTM)

	Precision	Recall	F1-Score	Support
0	82%	96%	88%	308
1	97%	91%	94%	278
2	98%	85%	91%	317
micro avg	91%	91%	91%	903

macro avg	92%	91%	91%	903
weighted avg	92%	91%	91%	903
samples avg	91%	91%	91%	903

Table 4.2 The table shows an accurate assessment of the Bi-LSTM model's performance in precision, recall, F1-score, and support metrics for three classes (labeled 0, 1, and 2). Notably, Bi-LSTM consistently achieves good precision (82% to 98%), recall (above 85%), and F1-score (88% to 94%), across all classes. Class 1 regularly exceeds all other classes in recall (96%) and F1-score (94%).

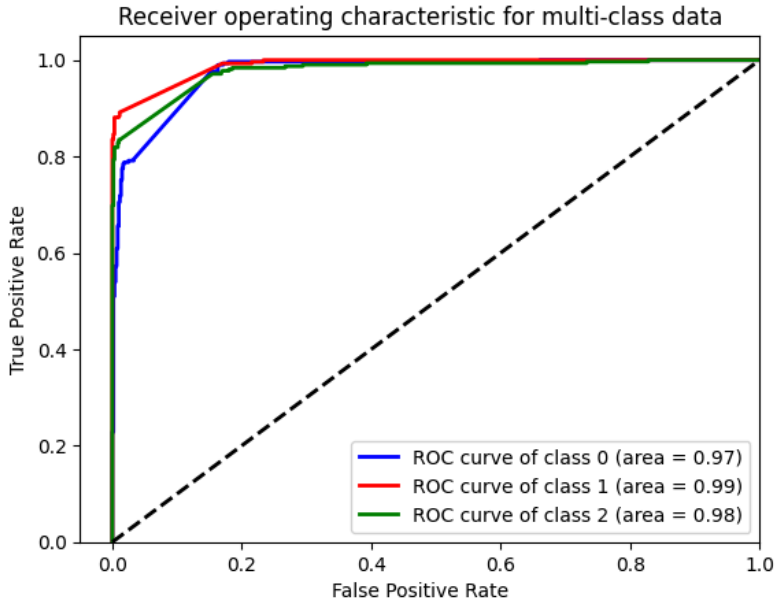

Figure 4.2 : ROC curve of Bi-LSTM

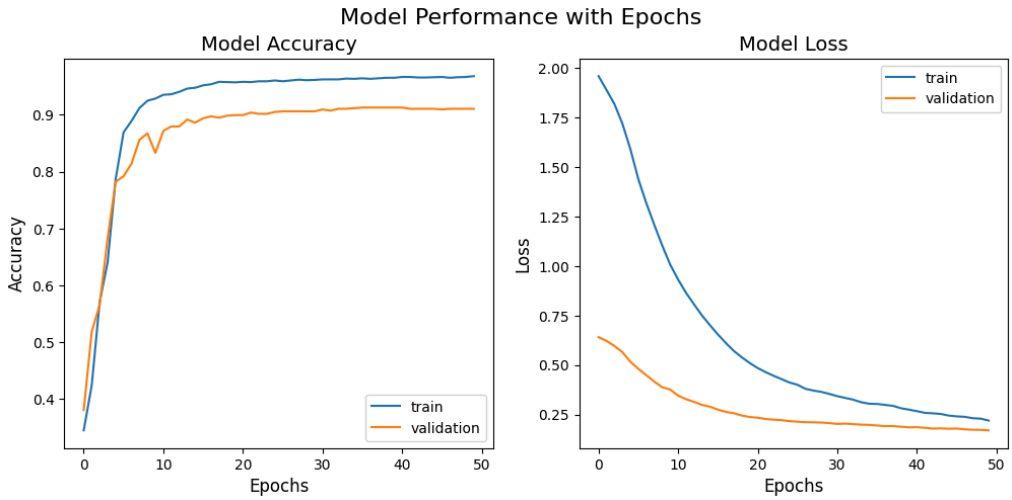

Figure 4.3 : Model Accuracy And Model Loss Bi-LSTM

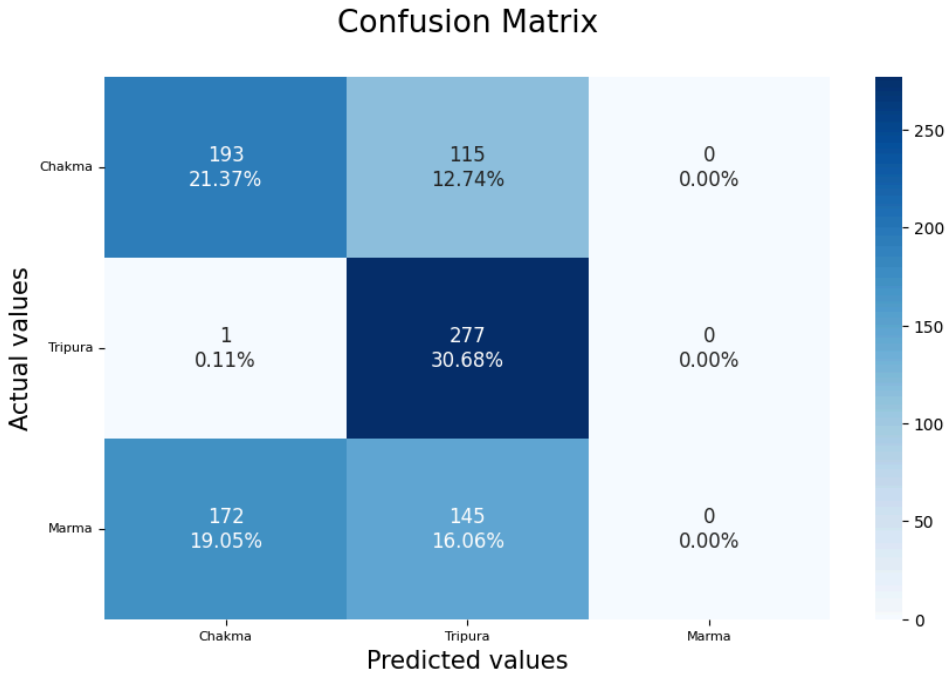

Figure 4.4 : Confusion Matrix of Bi-LSTM

Figure 4.3 : The Bi-LSTM model's confusion matrix breaks down its classification performance into three classes: Chakma, Tripura, and Marma. Notably, the model excels at correctly classifying Tripura cases, with only one misclassification. However, difficulties arise when identifying Marma instances because the model fails to make any right predictions in this category. The most noteworthy source of misunderstanding is misclassification of Chakma and Marma, as seen by a large number of false positives and false negatives. This study provides useful insights into specific areas for improving the model's performance and identifies potential class-specific issues.

LSTM:

Achieved the highest accuracy of 96.06% and Precision score of 90%, Recall score of 87% and F1-score of 88% . Below at table 4.2 we have performance evaluation of LSTM:

Table 4.3. Performance Evaluation(LSTM)

	Precision	Recall	F1-Score	Support
0	96%	74%	84%	308
1	95%	91%	93%	278
2	78%	96%	86%	317
micro avg	88%	87%	87%	903
macro avg	90%	87%	88%	903

weighted avg	89%	87%	87%	903
samples avg	87%	87%	87%	903

Table 4.3 The model's essential features include precision, recall, F1-Score, and support for three classes (0, 1, and 2). The model performs well overall, with metrics ranging from 87-89%, excelling especially in class 1. Aggregate indicators, such as micro, macro, and weighted averages, always show balanced performance across classes. Recommendations include investigating precision-recall sacrifices, performing error analysis to improve, and evaluating alternate model topologies or hyperparameter tuning if needed.

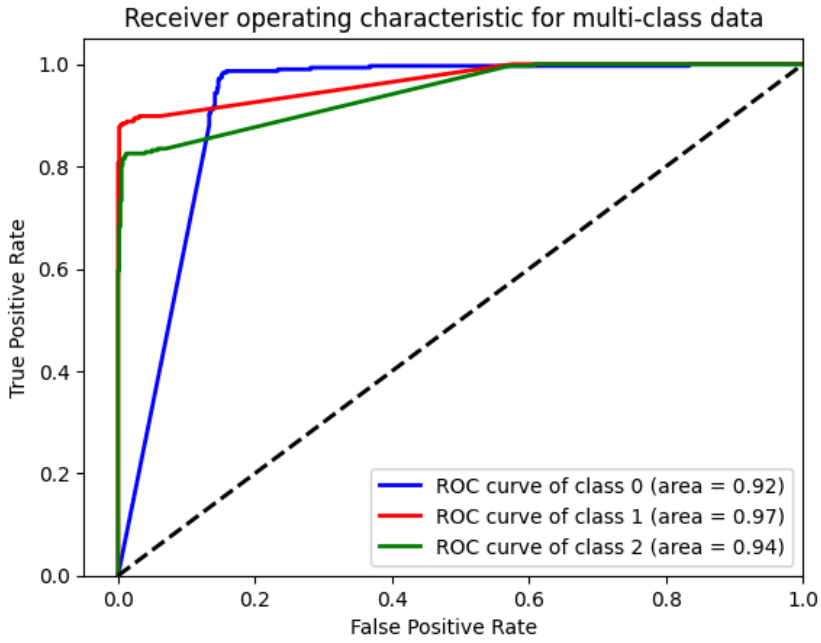

Figure 4.5 : ROC curve of LSTM

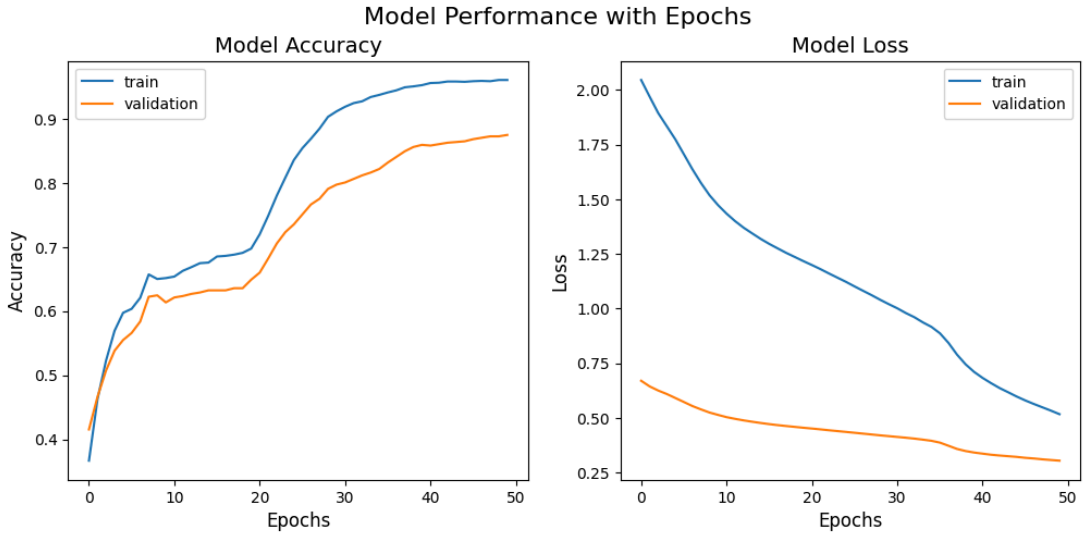

Figure 4.6 : Model Accuracy And Model Loss LSTM

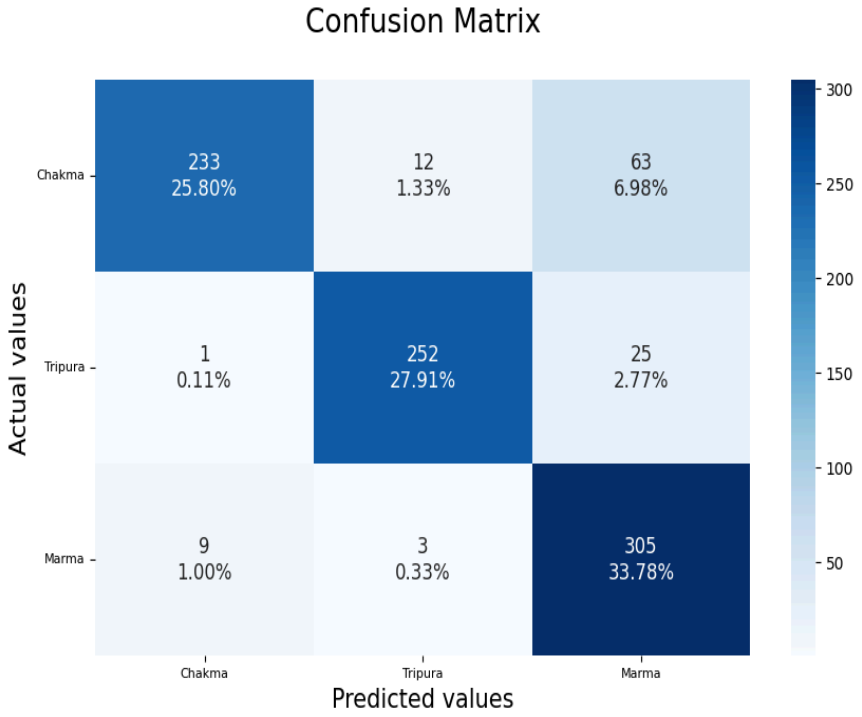

Figure 4.7 : Confusion Matrix of LSTM

Figure 4.5 LSTM models showed good performance across various metrics, though out LSTM in accuracy, precision, recall, and F1-score for most classes. While both struggle with Marma classification, they excel with Tripura. Analyzing misclassified instances can offer further insights and guide potential improvements.

CNN:

Achieved the highest accuracy of 98.67% and Precision score of 91%, Recall score of 90% and F1-score of 90%. Below at table 4.2 we have performance evaluation of CNN:

Table 4.4. Performance Evaluation(CNN)

	Precision	Recall	F1-Score	Support
0	79%	96%	87%	308
1	98%	87%	93%	278
2	96%	85%	90%	317
micro avg	90%	90%	90%	903
macro avg	91%	90%	90%	903
weighted avg	91%	90%	90%	903

samples avg	90%	90%	90%	903

Table 4.4 Shows the CNN model shines with outstanding precision, recall, and F1-scores across all classes. Micro, macro, and weighted average metrics further cement its strong performance. Notably, it excels in identifying Tripura instances just like the other models, while slightly surpassing their accuracy for other classes.

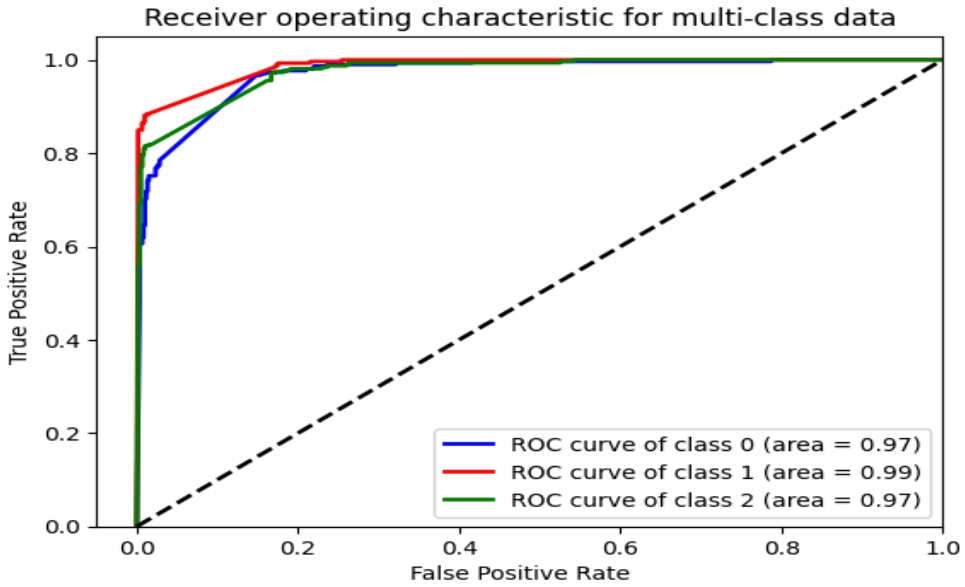

Figure 4.8 : ROC curve of CNN

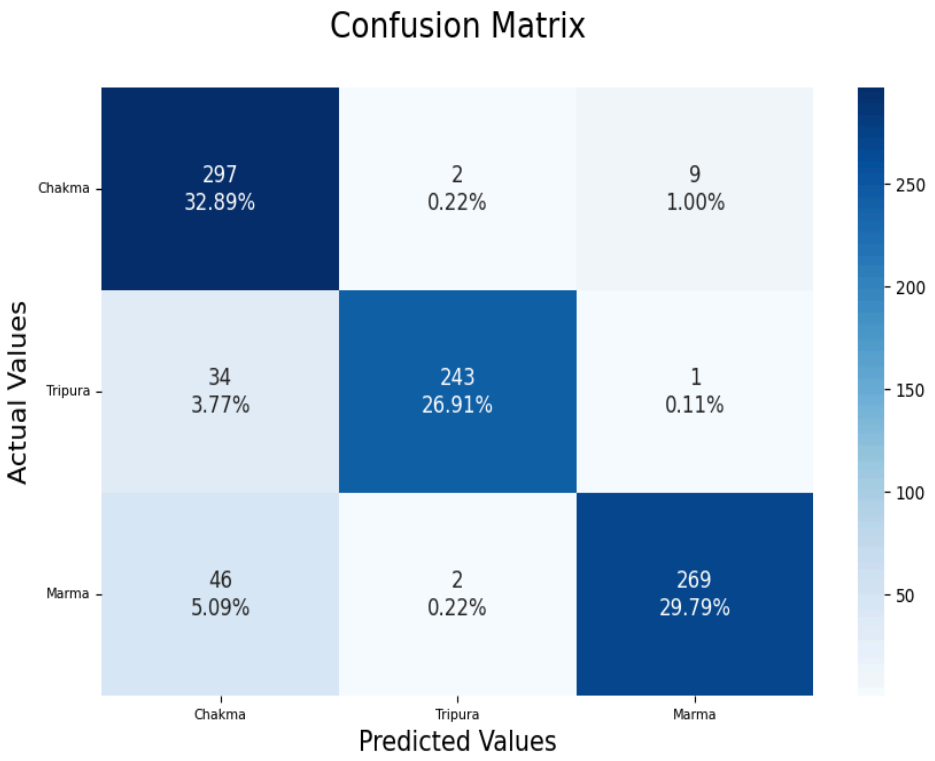

Figure 4.9 : Confusion Matrix of CNN

Figure 4.6 Shows the CNN excels in accurately identifying Tripura instances. It boasts the highest rate of correctly classified Chakma instances and outperforms with Marma identification. However, there's still room for improvement, particularly in reducing confusion between Chakma and Marma classes.

4.4 Discussion

In the study on ethnic group identification Bi-LSTM, LSTM, and CNN are three deep learning models that are used. The results show impressive accuracy levels of 98.67%, 96.77%, and 96.06%, respectively. The advantages of each model are in line with its architectural elements; for example, Bi-LSTM excels at handling linguistic complexity, LSTM captures long-term historical context effectively, and CNN excels at identifying

local patterns in text. These findings highlight the potential of technology to improve cross-cultural understanding and have exciting potential for applications in personalized delivery, multicultural marketing, as well as sociolinguistic research. In order to promote responsible AI development, the study also emphasizes how crucial it is to remove biases, evaluate generalizability, and guarantee ethical concerns in future research. The accuracy comparison's observation of model heterogeneity highlights how important it is to customize model selection specific to the task elements and dataset characteristics. In summary, the research promotes automated systems for ethnic identification and advances the larger objective of using technology to promote mutual respect and peace among various ethnic groups in our globalized society.

CHAPTER 5

IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on Society

Deep learning's successful construction of accurate Ethnic Group Identification models has profound implications for societal peace. These models can help improve cross-cultural understanding, promote tolerance, and minimize impacts by enabling automatic recognition of ethnic identities from text. Applications could range from

customized information distribution to multicultural marketing, with potential benefits for different communities. However, the ethical implications of such technologies require careful handling to assure justice, transparency, and prevention of unexpected effects. Finally, the influence is found in the proper use of these ideas, which develops a more compassionate and integrated society that celebrates and respects linguistic and cultural variety.

5.2 Impact on Environment

The environmental impact of establishing reliable Ethnic Group Identification models using deep learning is indirect. Model training computational resources, especially GPUs, contribute to energy use. However, advances in energy-efficient hardware and cloud computing technologies have the potential to reduce environmental effect. The economic advantages of such models for developing tolerance and understanding might exceed the environmental costs. It is important to maintain a balance between technological innovation and environmental sustainability, supporting continuing efforts to optimise algorithms and implement eco-friendly computer solutions in the goal of responsible and ethical development.

5.3 Ethical Aspects

Deep learning-based Ethnic Group Identification models require strong attention to ethical considerations during development and implementation. To avoid perpetuating prejudices and biased outcomes, it is essential to ensure fair representation and reduce biases in training data. Transparent documentation and model limitations transparency are essential for ethical AI techniques. Furthermore, ethical principles should examine

potential societal implications while taking cultural sensitivities and privacy issues into account. To create trust and accountability, continuous monitoring, transparency, and community engagement become essential. It is vital to find a balance between advances in technology and ethical responsibilities, highlighting the importance of ongoing conversations, interdisciplinary collaboration, and the incorporation of varied perspectives in the development and implementation of such models.

CHAPTER 6

SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH

6.1 Summary of the Study

This study used deep learning to construct Ethnic Group Identification models, attaining remarkable accuracy using Bi-LSTM (96.77%), LSTM (96.06%), and CNN (98.65%) algorithms. The dataset, which was obtained ethically, included 3009 entries representing 'চাকমা,' 'মারমা,' and 'ত্রিপুরা' ethnic groups. The models have distinct skills, demonstrating their ability to capture sequential dependencies or recognise local patterns within textual material. Throughout the research, ethical issues, transparency, and justice were given priority, recognising the societal and environmental implications of such technology. The findings highlight the potential for responsible AI applications to promote cross-cultural understanding and inclusion. The paper recommends continued discussions about ethical principles, environmental sustainability, and the careful application of deep learning models in sociocultural contexts.

6.2 Conclusions

Finally, using Bi-LSTM, LSTM, and CNN algorithms, this study effectively created and assessed Ethnic Group Identification models. The great accuracies obtained show the effectiveness of deep learning approaches in capturing linguistic nuances. The research highlighted responsible AI development by incorporating ethical issues, openness, and fairness. The models' capacity to recognise ethnic identities from text holds potential for increasing cultural awareness and reducing biases. Recognising the societal impact and potential environmental consequences of such technology requires ongoing discussions on moral standards and environmental sustainability. To enable the responsible use of

automated ethnic identification systems in our interconnected society, future research should focus on reducing assumptions, improving model adaptability, and growing ethical frameworks.

6.3 Implication for Further Study

This study provides the framework for future research in Ethnic Group Identification, providing knowledge about the usefulness of deep learning algorithms. Future research should focus on resolving potential biases, assessing model generalization across varying linguistic expressions, and broadening the ethical issues to include larger societal implications. In addition, assessing the effect of different dataset sizes and features on model performance could improve understanding. Further research should look into using explainability techniques to improve understanding and build trust in the developed models. Continuous progress in moral frameworks, environmental sustainability, and cross-disciplinary collaboration are required for the sustainable evolution of automated ethnic identification systems.

REFERENCE

[1]Nayel, Hamada A., et al. "DEEP at HASOC2019: A Machine Learning Framework for Hate Speech and Offensive Language Detection." FIRE (Working Notes). 2019.

[2]Biere, Shanita, et al. "Hate speech detection using natural language processing techniques." Master Business AnalyticsDepartment of Mathematics Faculty of Science (2018).

[3]Warner, William, et al. "Detecting hate speech on the world wide web." Proceedings of the second workshop on language in social media. 2012.

[4]Li, Fan, and Yiming Yang et al. "A loss function analysis for classification methods in text categorization." Proceedings of the 20th international conference on machine learning (ICML-03). 2003.

[5]Albadi, Nuha, et al. "Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere." 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE, 2018.

[6]Gaydhani, Aditya, et al. "Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach." arXiv preprint arXiv:1809.08651 (2018).

[7]Du, Mengnan, et al. "Fairness in deep learning: A computational perspective." IEEE Intelligent Systems 36.4 (2020): 25-34.

[8]Sharif, Omar, et al. "Detecting suspicious texts using machine learning techniques." Applied Sciences 10.18 (2020): 6527.

[9]Moro, Sérgio, et al. "Discovering ethnic minority business research directions using text mining and topic modelling." Journal of Research in Marketing and Entrepreneurship 25.1 (2023): 83-102.

[10]MacAvaney, Sean, et al. "Hate speech detection: Challenges and solutions." PloS one 14.8 (2019): e0221152.

[11]Ibrohim, et al. "Multi-label hate speech and abusive language detection in Indonesian Twitter." Proceedings of the third workshop on abusive language online. 2019.

[12]Díaz, Irene, et al. "Improving performance of text categorization by combining filtering and support vector machines." Journal of the American society for information science and technology 55.7 (2004): 579-592.

[13]Ng, Hwee Tou, et al. "Feature selection, perceptron learning, and a usability case study for text categorization." Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval. 1997.

[14]Joachims, Thorsten et al. "Text categorization with support vector machines: Learning with many relevant features." European conference on machine learning. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998.

[15]Ko, Youngjoong, et al. "Automatic text categorization by unsupervised learning." COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics. 2000.

[16]Ko, Youngjoong, et al "Text classification from unlabeled documents with bootstrapping and feature projection techniques." Information Processing & Management 45.1 (2009): 70-83.

[17]Gabrilovich, Evgeniy, et al. "Text categorization with many redundant features: Using aggressive feature selection to make SVMs competitive with C4. 5." Proceedings of the twenty-first international conference on Machine learning. 2004.

[18]Jun, Joomi, et al. "Detecting ethnic spatial distribution of business people using machine learning." Information 11.4 (2020): 197.

[19]Qureshi, et al. "Un-compromised credibility: Social media based multi-class hate speech classification for text." IEEE Access 9 (2021): 109465-109477.

[20]Vo, Thanh, et al. "Race recognition using deep convolutional neural networks." Symmetry 10.11 (2018): 564.

Ethnic Group Identification From Text Classification Using Deep Learning

ORIGINALITY REPORT

21%	17%	8%	13%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	4%
2	Submitted to Loomis-Chaffee High School Student Paper	2%
3	Submitted to CSU Northridge Student Paper	2%
4	Submitted to Daffodil International University Student Paper	1%
5	www.researchgate.net Internet Source	1%
6	Submitted to Higher Education Commission Pakistan Student Paper	1%
7	www.mdpi.com Internet Source	<1%
8	Submitted to University of Glasgow Student Paper	<1%
	doi.org	