A METHOD FOR BENGALI AUTHOR DETECTION USING SUPERVISED CLASSIFICATION MODELS

BY

Md. Abdul Hamid
ID: 191-15-12387

Md. Tanjil Rahman
ID: 191-15-12536
AND

Md. Fahim Islam
ID: 191-15-12600

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

Supervised By

Md. Sadekur Rahman
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised By

Abu Kaisar Mohammad Masum
Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2023

APPROVAL

This Project titled **"A Method for Bengali Author Detection using Supervised Classification Models"**, submitted by **Md. Abdul Hamid, Tanjil Rahman** and **MD. FAHIM ISLAM** to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 25 January 2023.

BOARD OF EXAMINERS

______________________ **Chairman**

Dr. Touhid Bhuiyan
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

______________________ **Internal Examiner**

Dr. Md. Monzur Morshed
Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

______________________ **Internal Examiner**

Dewan Mamun Raza
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

25.1.2023
______________________ **External Examiner**

Dr. Ahmed Wasif Reza
Associate Professor
Department of Computer Science and Engineering
East West University

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Md. Sadekur Rahman, Assistant Professor, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Sadek 24.1.23

Md. Sadekur Rahman
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised by:

Kaisar

Abu Kaisar Mohammad Masum
Lecturer
Department of CSE
Daffodil International University

Submitted by:

Hamid

Md. Abdul Hamid
ID: 191-15-12387
Department of CSE
Daffodil International University

Tanjil

Md. Tanjil Rahman
ID: 191-15-12536
Department of CSE
Daffodil International University

Fahim Islam

Md. Fahim Islam
ID: 191-15-12600
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty God for His divine blessing making us possible to complete the final year project/internship successfully.

We are really grateful and wish our profound indebtedness to **Md. Sadekur Rahman**, Assistant Professor, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Natural Language Processing*" to carry out this project. His endless patience, scholarly, guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

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

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

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

ABSTRACT

Text classification is an important area of study in the field of NLP. We live in a modern world where everyone values their intellectual property. Intellectual property includes digital written ideas, blogs, poems, novels, and posts, among other things. Evil people try to steal valuable intellectual property from others and claim it as their own or pirate these properties. To avoid these problems, we created several models based on the art-of-states Supervised method for determining authorship from a given Bangla text. Because our work is a multi-class classification, we can use it to determine who created articles, news, or messages. Authorship detection can be used to identify anonymous authors as well as detect plagiarism. This article focuses on categorizing five authors in the context of Bengali text. These five authors are well-known figures in Bengali literature and poetry. Humayun Ahmed, Rabindranath Tagore, Muhammad Zafar Iqbal, Kazi Nazrul Islam, and Sarat Chandra Chattopadhyay are among those honored. Data is being gathered from over 4500 paragraphs. For the experimental evaluation, a dataset is created. We preprocess Bengali text for training purposes. Logistic regression, naive Bayes, decision trees, SVM, Random Forest, XG-Boost, and KNN are among the seven supervised classification methods used. Our deep learning Bi-Lstm model outperforms the seven supervised models in terms of accuracy. By mentioning all models, the transformers-based model, Bert uncased model learns the context very well. Bi-Lstm was used in our experiment. Bi-Lstm and Bert uncased model provides the best experimental classification report in our experiment. The Bi-Lstm model loss function yields 0.3789 with a maximum accuracy of 88% and Bert base uncased F1-Score gives 91 % accuracy.

TABLE OF CONTENTS

CONTENTS **PAGE**

CHAPTERS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

1.1 Introduction

Despite the fact that Bengali is one of the most widely spoken languages in the world, research on Bengali author detection is currently lacking [1]. Authorship identification is a computer approach that deals with the evidence of the author in a certain text [2]. The work expended in identifying the author of each article under evaluation is referred to as author identification. Stylometry is a text-mining technique that analyzes texts written by well-known authors to measure the author's style and the poetic rhythm of his mind. It does this by choosing a few traits that are present consistently throughout the author's writing, serving as the author's linguistic fingerprint. It is necessary to establish a link between the works of known writers and those by unknown authors to establish authorship identification, a subtask of stylistic detection Every day, there is a large amount of information or literary genres available in today's internet era. The practice of assigning an unknown work to a writer based on some distinctive trait or meter is known as author identification. It may be used when many people claim to have created a work, or when no one can or will identify the genuine author. As a result, while Stylometry [3] is widely used to detect theft, it is also employed in criminal investigations involving literary authorship or forensic linguistics, such as identifying the writers of anonymously published books for the police. The study of varied language styles as well as personal writing preferences is known as stylometry. Forensic [4] linguistics and the identification of genuine confessions are two legal applications of these terms. It has been successfully applied in a greater number of genres in recent years, including forums, email, blogs, chat, and other digital tools, as well as music and fine-art paintings. Binongo and Smith, Holmes et al., and Burrow [5] first used multivariate analysis (MVA) and principle components analysis (PCA) on a few function words to solve different authorship issues. Stamatatos [6] categorizes English and Arabic news as a corpus using support vector machines (SVM) on the character level of n-grams. Koppel et al [7] describe the process of assigning authorship to hundreds of prospective writers .A wide range of applications for authorship attribution exists, including author verification, plagiarism detection, author profiling or characterization, stylistic inconsistency detection, forensic linguistics, and others. Authorship attribution is generally concerned with identifying the original author of a given text from a set of given authors. There are several writing styles used by authors, and their use of particular words can be seen in their work. These are the main characteristics that can be used to categorize their work. For addressing challenges like plagiarism, email spam, online fraud, and other problems, author identification is a very significant and active area of machine learning and Deep Learning. Forensic linguistics and the identification of genuine confessions are two legal applications of these terms. There is always a danger of plagiarism and intellectual theft in modern society. People should integrate a stylometry prevention system into their systems to prevent such tasks. The first data obtained here is over 4500, and the data has been modeled so that it may be readily applied as a train and test. Data is retained as trained 80 percent of the time and tested 20 percent of the time. The model was mostly taught as a train, and its correctness was verified through testing. We estimate the author using Machine Learning (ML), Deep Learning (DL), and Bert Model technique based on the data we have. We used techniques such as Logistic Regression, Random Forest, K-Nearest Neighbor (KNN), Naive Bayes, Decision Tree, XG-Boost, Vector Machine Support (SVM), Bi-Lstm, and Bert uncased model to determine which method works best in terms of high accuracy.

1.2 Motivation

We live in a contemporary era, and robots can now speak with humans thanks to AI (Artificial Intelligence) and NLP (Natural Language Processing). Almost all information, thoughts, books, poetry, novels, and so on are available on the internet. Though it is a blessing that these are available on the internet, a significant disadvantage is that some individuals collect this information and attempt to claim their digital property. Plagiarism detection is a fantastic tool for preventing such concerns, but we know that other languages, such as English, French, and others, have a very effective system, however, Bengali has not, which is pitiful. Sometimes people copy Bengali writers' paragraphs of novels, poetry, and so on without citing copyright, but no mechanism can identify these issues when people do this consciously or unknowingly on Facebook, Twitter, blogs, forums, and so on. For these sorts of challenges, a system that can identify and forecast the author based on the author's writing style is absolutely important. These kinds of issues drive us to do this research.

1.3 Rationale of the Study

So far, when we look at the technological benefits of other languages, we realize that they are significantly superior to our own Bangla language because much work has already been done on these languages. For example, the English language has already grown in popularity. We've also made use of its technical advancements, such as Grammarly, paraphrasing tools, auto words or sentence recommendations, and so on. However, our native language lacks these features due to a lack of study in our native language. Bengali researchers should be aware of such issues and work to develop or find solutions in the present technology arena. Digital property, like other types of property, must be properly safeguarded. Ideas, poetry, books, and other author-related property should be secured in the digital age. To address this issue, we created a corpus of data from 5 Bengali writers for our research. Based on context or author writing style, we provided various methods for identifying writers from a particular Bengali paragraph.

1.4 Research Questions

While conducting our investigation, a few questions came up. These are the crucial inquiries that will allow us to systematically wrap up our investigation. The following list of crucial queries is given below:

1. Why exactly are we doing research on Bengali authors?
2. How and where from the data sources used?
3. How to perform preprocessing on Bengali data?
4. Which model is appropriate for the context?
5. How would the model evaluate?
6. What are the practical implications of this work?

7. What are the further scopes of this work?

1.5 Expected Outcome

Artificial Intelligence & Industrial Applications A2IA'2023 will be held on February 17th and 18th, 2023. Meknes, Morocco. That conference received our study paper. Thank you, lord; our work was accepted, and they requested minor revisions, which we also completed. However, in this research or thesis, we are working on a broad version. We collect additional data in order to use deep learning and transformer-based models to improve accuracy and model quality. We have discovered and are ready to pursue the following outcome:

- A given paragraph can be used to identify the proper author using our method. The algorithm will evaluate the paragraph to determine which author's writing style it is.
- A website may be developed where any author can form a genuine account and secure their ideas, poetry, novels, and so on. It will function similarly to plagiarism from the standpoint of the Bengali language.
- The more data there is, the better the model. The transformer-based model learns language more efficiently than previous models and provides context-based outputs. We anticipate creating a massive data collection for future work in the XLnet model. Which is the most recent model that produces the greatest results in this area.
- We are looking for two additional papers based on this study to be published in Q2 journal.

1.6 Research Layout

To finish our study, we organized our report into five subsections. They're listed below.

1. In Chapter 1, we discuss the overall structure of our study project and break it into many subchapters. For example, our project's introduction, motivation, rationale, and so on.

2. We reviewed earlier work and a literature review in Chapter 2, as well as the scale of the problem and the obstacles in this work.
3. In Chapter 3, we will discuss our work organization, diagram, flow, methodologies, and approaches for developing a Bengali author detector mode.
4. In Chapter 4, we will go through the Experimental Results and our Build Model Discussion.
5. In Chapter 5, we examine how our work has an impact on society, the environment, and sustainability.
6. We covered the work's summary, conclusion, and further research in Chapter 6.

CHAPTER 2

Background

2.1 Terminologies

Natural language processing has been a prominent issue for scholars in recent years. From the study, the majority of the work was completed in English but recently, there are currently more NLP model constructions based on Bangla text datasets. Because other popular languages have done adequate research on author attribution, author identification in Bengali is crucial. Because of electronic media, people are now more involved with the internet than ever before. Individuals share their thoughts on social media, blogs, forums, and other platforms; however, the problem is that occasionally individuals copy and claim as their own a certain line from another person's written poetry or novel that they find entertaining or attractive. This paper discusses a classification problem and how to simply get the highest classification accuracy by utilizing Supervised machine learning, deep learning and, transformer-based models. The data was collected in the 4500+ range and designed to be readily utilized as a train and test. Data is kept 80% of the time as taught and 20% of the time as assessed. The model was mostly taught as a train, and its accuracy was verified, so how much of the 4500+ data it can recognize correctly is clear. The seven supervised machine learning approaches are logistic regression, naive Bayes, decision trees, support vector machine (SVM), random forest, K-nearest neighbor (KNN), and XG-Boost. The best model accuracy is outperformed by the deep learning model and the Bert-based uncased model.

2.2 Related Work

This [8] work provided an autonomous author detection for Turkish literature that compares a novel classification methodology created using existing methods against established techniques. Initially, 22 style identifiers were removed and handled as equal weights, with a 67% success rate observed. The results of artificial neural networks show that MLP has a 60% success rate and the radial base function has a 72% success rate. In phase 2, 11 of the 22 style markers were assigned equal weights, increasing the success percentage to 78%. MLP success rate was 60%, while radial base function success rate was 61%. In the third phase, we used varied weights for the style markers SM3, SM13, SM17, and SM21 and calculated the success rate of 84%.

The authors of this [9] research study developed a methodology for determining the authorship of online messages. They tested their approach using English and Chinese internet newsgroup postings as the dataset. They investigated the relative strengths of the four categories of attributes as well as the three classification techniques: decision trees (C4.5), back-propagation neural networks (NN), and support vector machines (SVM). On English and Chinese datasets, those three classifiers obtained 90 to 97% and 72 to 88% accuracy, respectively.

The authors published new distance measurements for author identification using the CNG approach in [10]. For long profile lengths, the suggested method provides a more stable solution than standard CNG. All of the assessments in this study were character-based on 3-gram. In that situation, the length of the shortest (and longest) profile has greatly grown, as have the classification results.

The authors of this [11] research study provide methods for assigning Latent Dirichlet Allocation (LDA) to the author's attribution. They investigated authorship attribution with the authors of a few candidates and proposed a novel approach that delivers sophisticated performance in the latter scenario. Three datasets are employed in this case: Judgment, IMDb62, and Blog. A striking observation is that LDAH-S gives good accuracy even with a small number of individuals, but LDAH-M takes around 50 subjects to outperform LDAH-S. In the future, they intend to investigate LDA patterns directly with model writers rather than utilizing it as a black box.

They [12] approach the PAN 2013 and find Author Identification in this study. They took numerous attributes out of the texts, such as word bags, stop words, punctuation bags, part of speech (POS) bags, and others, and used a modified weighted KNN to identify the author because it performs well with a short corpus of data.

They [13] answer the question of whether it is feasible to correctly identify the author of a very short text. They introduce the concept of an author's individual "signature" and show how such signatures are common among many writers while creating incredibly brief messages as part of their research utilizing Twitter as an experimental testbed. They aimed to outperform our baselines while also adding a new authorship attribution feature (called "flexible patterns"). Their findings show that the author of a single tweet can be recognized with high accuracy across a wide range of authorship attribution tasks, and their technique outperforms the state-of-the-art by 6.1%.

They [14] establish a system for authorship detection that is automated as part of their study. The frequency of function words was used by them as a categorization characteristic. Preprocessing the texts, obtaining classification characteristics, and executing classification are the crucial three phases. This paper contrasts the performance of an MDA classifier with an SVM classifier regarding the third step. Although both techniques are more accurate than 90% of the time, the SVM has certain limitations because it only offers binary choices.

To describe the writing style of unstructured texts by diverse authors, they [15] build an unsupervised technique to extract stylistic characteristics and identify authorship. This method can partially address the issue of the independence of different dimensions. According to comparative experimental findings on two data sets, the recommended qualities in conjunction with the classification technique in this research give a considerable boost in performance for the authorship identification task.

This study [16] investigates how difficult it is to discern who composed brief historical Arabic writings by ten different authors. They can extract several lexical and character components of each author's writing style using N-gram word levels and character levels as a text representation. The Naive Bayes classifier is then used to categorize the texts according to their authors. AAAT is a dataset that comprises three condensed texts for each book authored by a certain author. Twenty texts are used for training, while ten are used for testing. Using N-gram words at level 1, algorithms were used to get the best categorization accuracy of up to 96%. Future research might look at the resilience of various ML systems for jobs involving several authors and tiny text sets. It may also broaden the study's scope to cover additional issues.

Deep learning is used in their research to extract attributes from texts represented by variable size character n-grams. In essence, they [17] extract document characteristics using a Stacked Denoising AutoEncoder (SDAE) with varied parameters before classifying the data with a support vector machine classifier. According to the findings, the proposed solution exceeds its competitors in terms of efficacy. Using 10-fold cross-validation settings, the proposed technique outperformed their findings in terms of classification accuracy. For relatively high feature spaces, chi-square-based feature selection outperforms frequency-based feature selection, whereas the opposite is true for lower dimensional spaces. When min-max normalization is used, accuracy improves over when it is not. With their method, classification accuracy might reach up to 95.12%. For the same corpus, their implementation feature sets from earlier studies only managed an accuracy of about 80%. Their method is constrained since it only contrasts pretraining-generated code features with pre-training-generated code attributes.

In the experiment, their [18] suggested method uses a Gaussian-Bernoulli deep belief network and Gaussian units in the visible layer to characterize real-valued input. A method for integrating two qualities into one is discovered by examination of lexical, syntactic, and application-specific elements. To replicate the CA rulings, the material on the internet is divided into several brief paragraphs. Using email corpora from Twitter and Enron as well as block sizes of 140, 280, and 500 characters, the proposed method is empirically evaluated. With an error rate ranging from 8.21% to 16.73%, the results are encouraging. A counterfeit sample error rate of between 5.48% and 12.3% is attained with smaller samples. They do not work in other languages, and they have a finite number of models at their disposal.

They [19] use a strategy to simultaneously train a neural network and a classification layer to learn continuous representations for n-gram data in their study. The experimental findings show that the suggested model works equally well on all four datasets while outperforming the state-of-the-art on two of them. Four open datasets were used to train their model in this study. For authorship attribution problems, continuous n-gram representations were suggested in this work. Using four authorship attribution datasets, they showed how well our strategy worked at detecting the authors' writing styles. Their findings demonstrate that continuous representations are suitable for tasks like authorship attribution that call for stylistic (as opposed to topical) text classification.

Their study [20] presents a methodology based on the dynamics of word co-occurrence networks, which represent eight authors and their 80 texts. Following the creation of time series for 12 topological metrics, the texts were divided into sections with equal linguistic tokens. They use K-nearest neighbors and successfully predict 71 of 80 paragraphs (88.75% accuracy), which is remarkable.

The authors [21] assembled a corpus by gathering many tweets from the tweeter. Train using Bayes theorem-based Naive Bayes classifiers. Their primary goal is to determine whether a specific person sent a specific tweet. Their models performed reasonably well. Their biggest flaw was that the dataset was too small for the model to fit into, resulting in a Kappa value of 67%, an accuracy of 97%, and balanced accuracy of 76%.

Their [22] goal is to see if they can use an artificial neural network and a machine learning algorithm to identify the actual authors of some unidentified Bangla writings. To achieve this task, they made a corpus collecting articles from eight political writers. SVM classification models and multilayer feedforward neural networks are used to build an attribution system. They created two voting systems using an MLP classifier and an SVM classification. They conclude that voting yields significantly better results and is a more efficient research method than categorization methods. The voting system provides 81.7% classification accuracy compared to the Support Vector Machine's (linear) 70% accuracy.

2.3 Comparative Analysis and Summary

Previously, most authors worked in a separate language for author detection, with only a handful working in Bengali. Though several of the writers worked on author detection in Bengali, their data was restricted, and only a few classification models were employed.

Table 2.3 is some comparative analysis and contribution of some authors in the field of author identification.

Table 2.1: Published work summary based on author detection

Ref	Year	Contribution	Dataset	Model	Results and Finding
[23]	2015	A new method introduced for authorship attribution based on function WANs	Own dataset	WAN's, Naïve Bayes, 1-NN, 3-NN, DT-gdi, DTce, SVM.	Naïve bayes error rate 10.8% and Support Vector Machine error rate 11.5%
[8]	2017	Data was gathered through social media and Internet communication channels. A methodology for real-time analysis that has been explained can find information about threats being traded on IRC by hackers.	Own dataset	Stanford CoreNLP, RNTN model.	Detects threat information connected to shadow brokers leaking vulnerabilities 28 days before the WannaCry ransomware assault.

[21]	2020	Data set obtained with the Twitter API, Experiment with a Multi-label classifier.	Collected from Twitter API,	Naive Bayes classifier	Kappa value of 67%, and balanced accuracy of 76%
[24]	2021	One training set and one validation set with 11200 and 2400 problems each made up the PAN'21 SCD challenge's data set.	collected from an English-written Q&A forum	Hybrid algorithm.	task one and two with F1-scores of 86% for task one and 78% for task two on the validation set.
[15]	2021	Make the DH (Digital Humanities) community more aware of the advantages of using deep learning models.	Online Dataset (Kaggle)	DNN	Analyzing multiple use-cases of DH studies in recent literature and their possible solutions and lays out a practical decision model for DH experts for when and how to choose the appropriate deep learning approaches for their research.

CHAPTER 3

Research Methodology

3.1 Instrumentation and research subject

We develop models that can determine authorship from provided Bengali e-text. To establish this model, we must first produce a corpus data collection with enough data for the machine to understand each author's writing style. First, we choose the art of state-supervised models. Logistic Regression (LR), Support Vector Machine (SVM), Naive Bayes (NB), KNN (K-Nearest Neighbor), Decision Tree (DT), Random Forest (RF), and XB-Boost (XGB) are the seven supervised categorization methods. Our work is mainly a multi-class classification task. Because these seven models are supervised models, input and output data will be sent to the system to build the model, and the system will forecast a specific class based on unseen data. Following the completion of machine learning models, we are seeking improved models that can provide more accuracy depending on the author's context. We created another model for this classification problem, Bi-Lstm, which is a deep learning model for solving multi-class classification problems. We achieve pretty decent accuracy compared to the previous seven machine learning models. We are not satisfied with this model; we are seeking a more recent model in this respect, which is a transformer-based model. We discovered that the Bert uncased model is the ideal model for our situation; we built it for it and discovered that it provides among the greatest model accuracy and classification reports.In the next part, we will go through our proposed methodology as well as all of the models we use to achieve our goal.

3.2 Data Collection Procedure

We gather data from many sources, especially books by five authors: Humayun Ahmed (HA), Rabindranath Tagore (RT), Muhammad Zafar Iqbal (MZI), Kazi Nazrul Islam (KZI), and Sarat Chandra Chattopadhyay (SCC). These data come from a number of sources, including the Internet and books. There are almost 4500 paragraphs in the data we gathered. The XLS sheet now has two new attributes. One is the author's paragraph, and the other is the author's name. For each paragraph, we tried to gather paragraphs of the same length. Table 3.1 shows the sample data.

Table 3.1: Dataset of Bengali Author Detection

Paragraph	Author
সে হ'ল আজ তিন বছরের কথা। আমার এই খাপ-ছাড়া জীবনে তার স্মৃতিগুলো ঝড়ের মুখে পদ্মবনের মত ছিন্ন-ভিন্ন হ'য়ে গেছে। কখনও তার একটি কথা মনে পড়ে,কখনও আধখানি ছোঁওয়া আমার দাগা-পাওয়া বুকে জাগে। মানস-বনের যুঁই-কুঁড়ি আমার ফুটতে গিয়ে ফুটতে পায় না, শিউলির বোঁটা শিথিল হ'য়ে যায়। ওরই সাথে এই শাঙন-ঘন দেয়া-গরজনে আর এক দিনের অমনি মেঘের ডাক মনে পড়ে, আর আঁখি আমার আপনি জলে ভ'রে ওঠে।	কাজী নজরুল ইসলাম
এ বৎসর চারিদিকে অত্যন্ত জ্বর হইতেছিল। নারায়ণীও জ্বরে পড়িলেন। তিন-চারিটা গ্রামের মধ্যে একমাত্র খানিকটা-পাশকরা ডাক্তার নীলমণি সরকারের একটাকা ভিজিট দু'টাকায় চড়িয়া গেল এবং তাঁহার কুইনিনের পুরিয়া অ্যারারুট ও ময়দা সহযোগে সুখাদ্য হইয়া উঠিল। সাতদিন কাটিয়া গেল, নারায়ণীর জ্বর ছাড়ে না। শ্যামলাল চিন্তিত হইয়া উঠিলেন।	শরৎচন্দ্র চট্টোপাধ্যায়
তোমাদের যাদের ফুসফুসে জোর আছে তারা খুব সহজেই বেলুন ফুলাতে পার। একটা বেলুন ছবিতে যেভাবে দেখানো হয়েছে সেভাবে একটা বোতলের মাঝে লাগিয়ে ফুলানোর চেষ্টা কর দেখি! যতই চেষ্টা কর দেখবে তুমি বেলুনটা ফুলাতে পারবে না, কারণ এখন বেলুন ফুলাতে হলে একই সাথে বোতলের বাকি বাতাসকে সংকুচিত করতে হবে, তোমার ফুসফুসে সেই জোর নেই!	মুহম্মদ জাফর ইকবাল
পরীক্ষকের ডাক শুনিয়া অপরাধীর মতো আশা ভয়ে ভয়ে বইখানি লইয়া মহেন্দ্রের চৌকির পাশে আসিয়া উপস্থিত হয়। মহেন্দ্র এক হাতে কটিদেশ বেষ্টনপূর্বক তাহাকে দৃঢ়রূপে বন্দী করিয়া অপর হাতে বই ধরিয়া কহে, "আজ কতটা পড়িলে দেখি।" আশা যতগুলা লাইনে চোখ বুলাইয়াছিল, দেখাইয়া দেয়। মহেন্দ্র ক্ষুণ্‌নস্বরে বলে, "উঃ! এতটা পড়িতে পারিয়াছ? আমি কতটা পড়িয়াছি দেখিবে?" বলিয়া তাহার ডাক্তারি বইয়ের কোনো-একটা অধ্যায়ের শিরোনামটুকু মাত্র দেখাইয়া দেয়। আশা বিস্ময়ে চোখদুটা ডাগর করিয়া বলে, "তবে	রবিন্দ্রানাথ ঠাকুর
আঙুল-কাটা জগলু ভাই ডাকছে। আমি আপনেরে উনার কাছে নিয়া যাব। দৌড় দেওনের চিন্তা মাথা থাইক্যা দূর করেন। চাইরদিকে আমরার লোক আছে। আমার পিছে পিছে হাঁটেন। ডাইনে-বামে চাউখ দিবেন না। চউখ বন্ধ করে ফেলি, হাত ধরে ধরে নিয়ে যান। বাইচলামি অনেক করছেন। আর না। আঙুল-কাটা জগলু ভাইরে চিনছেন তো? নাকি চিনেন নাই। পরিচয় দিব? পরিচয়ের প্রয়োজন আছে?	হুমায়ুন আহমেদ

3.3 Statistical Analysis

In the first phase we collected 250 data for each class or author and apply our models. Later, for the experimental and final dataset we add more and more data on some authors to see the differences in each class. Table 3.2 describes the total data for each author.

Table 3.2: Each Class Total Data

Class	Total data
হুমায়ুন আহমেদ	1793
কাজী নজরুল ইসলাম	1725
শরৎচন্দ্র চট্টোপাধ্যায়	505
রবিন্দ্রানাথ ঠাকুর	255
মুহম্মদ জাফর ইকবাল	251

Figure 3.1 shows the total data for each class. Different color of line shows different number and different class’s data.

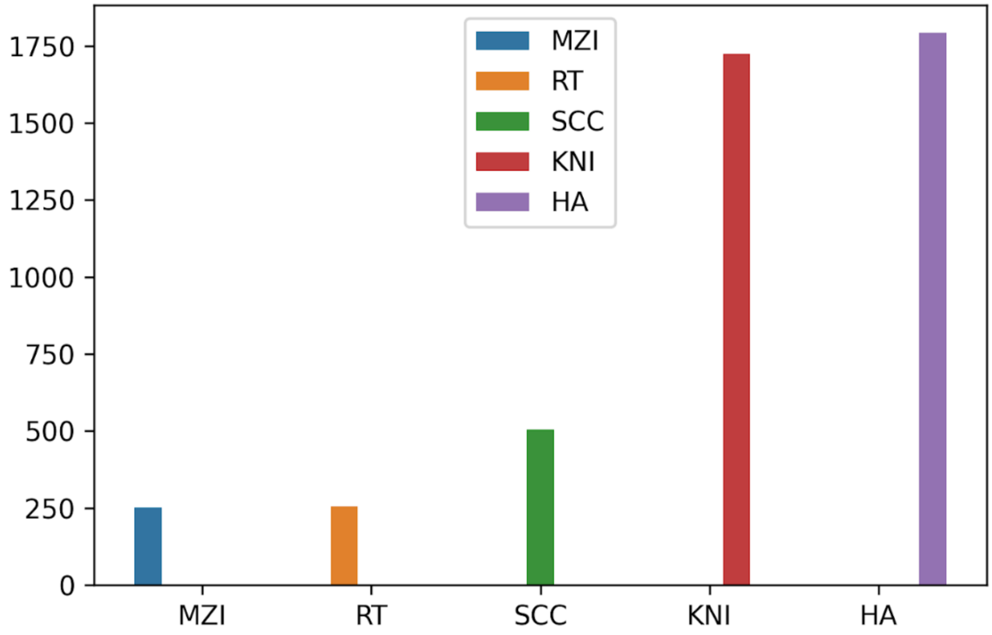


Figure 3.1: Each class Paragraph Collection

3.4 Proposed Methodology for machine learning models and data processing

Figure 3.2 describes our proposed methodology. How we accomplish our machine learning models.

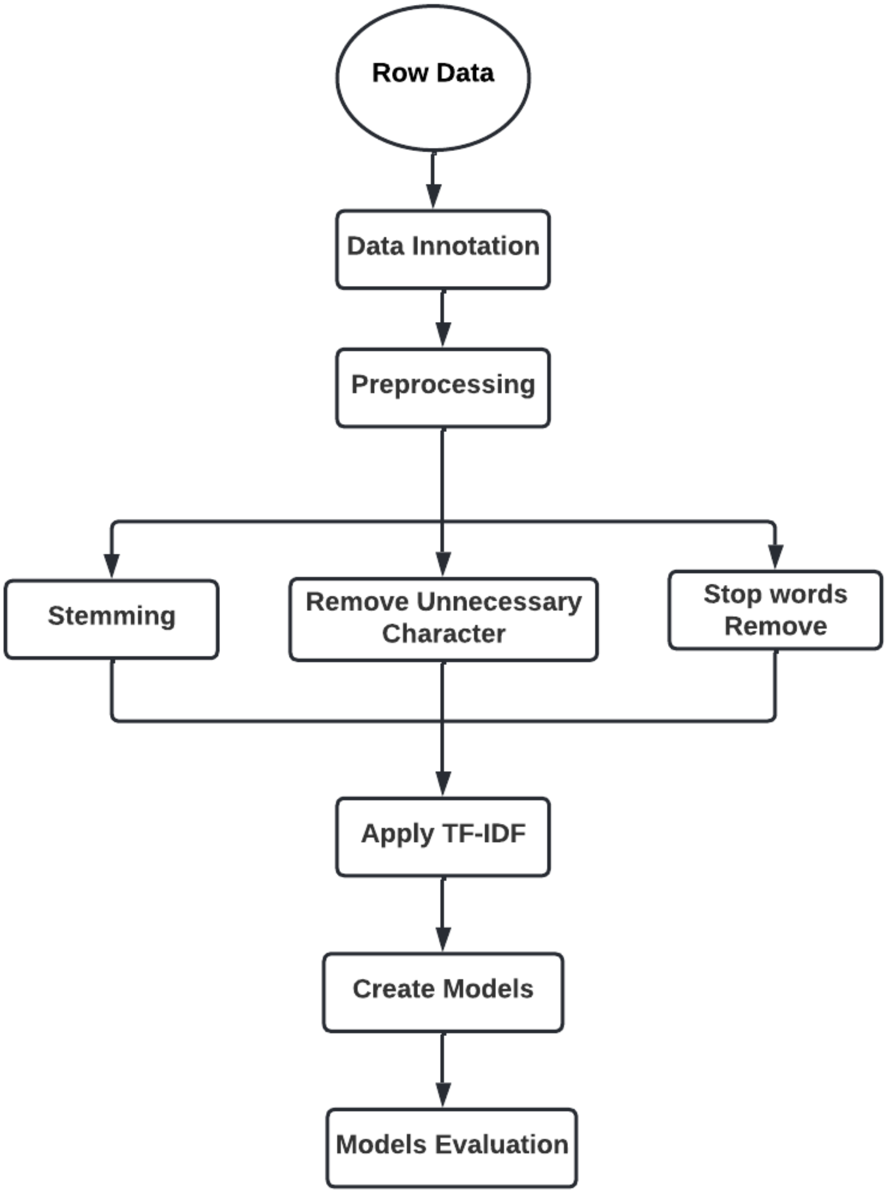


Figure 3.2: Proposed Methodology for Machine Learning Models

3.5 Data preprocessing

Preprocessing the dataset to make it appropriate for fitting into the model is a crucial step in NLP. We clean our data by using a variety of cleaning procedures as a proper part of the preparation stage. In a few stages, we finish our preparation task. The steps are listed below. By eliminating rows and deleting duplicate rows, we have taken care of null values and duplicate data in a premature state. There are several extraneous characters throughout the text. like "[", "/", "@", "|," and so forth. These extraneous characters must be eliminated from the text in order to create a clean dataset because they are not appropriate for our dataset. Table 3.3 showing removing unnecessary Bengali letters.

Table 3.3 Removing Unnecessary Letter

Text	After Removing Unnecessary Text
তুমি কে?’ অনেকক্ষণ কিছু শোনা গেল না	তুমি কে অনেকক্ষণ কিছু শোনা গেল না

Stop words are causing issues when we train our multiple models, therefore we must address them in the following approach. Stop words like " অতএব ", " অবশ্য ", " অন্তত ", " অথবা " etc. We build a corpus of stop words, and with its aid, we eliminate these stop words. Table 3.4 showing removing stop words.

Table 3.4 Removing Stop Words

Text	After Removing stops words
অতএব এখন কী করে আমার দিন কাটচে	এখন কী করে আমার দিন কাটচে

Stemming is renowned for obtaining word roots. It is significant in the field of Bengali NLP. We prepare our dataset by using a variety of relevant stemming rules to obtain root words.

We divided our dataset into two sections, with 80% of the data used for training and the remaining 20% used for investigation.

3.5.1 Feature Extraction

The TF-IDF is most commonly used in automated text analysis. It is highly useful for scoring words in NLP machine-learning algorithms. TF-IDF is an abbreviation for the term frequency-inverse document frequency. It is used to compare the frequency of occurrence of a phrase in a document to that of a dataset [26]. The TF-IDF value grows or drops proportionately to how many times a word appears in the document or how many documents in the corpus include the term.

It has two parts: Term Frequency (TF): The frequency with which a term appears in a corpus is measured by term frequency. It computes how many times a word appears in relation to the total amount of words. It is written as

$$TF = \frac{\text{Number of times the term appears in a document}}{\text{Total number of words in the document}}$$

There are more methods for determining phrase frequencies. As an example, consider employing the maximum word frequency. as well as the document's average phrase frequency

Inverse Document Frequency (IDF): The relevance of a word in a document is measured using inverse document frequency. It indicates how frequent or uncommon a term is in a corpus. The lower the score, the more frequently the term is used. IDF is written as

$$IDF = \log(\frac{\text{Number of the documents in the corpus}}{\text{Number of the documents int the corpus contain the term}})$$

Finally, it is formulated as below:

$$TF - IDF = TF * IDF$$

3.6 Machine Learning-Based Classification Model Selection

Based on the paragraph of a text, our approach efficiently produces an informed prediction about the author. To achieve our aim, we use supervised learning approaches and, as a consequence, classification algorithms. Classification algorithms will anticipate category labels from new observational data based on our dataset. Because these strategies function by teaching the model, training data must be accessible. The primary goal of classification algorithms is to find categories or labels in an existing dataset. We'll go through a few categorization approaches in the next sections.

3.6.1 Logistic Regression

One of the most extensively used machine-learning approaches is logistic regression. It is used to forecast the application of a collection of labeled structural and independent factors [27]. The output of the Logistic Regression approach is a categorized conditional variable. As a result, the result should have a distinct or categorized value. It might be binary, such as yes or no, 0 or 1. When given accurate, true binary numbers, it delivers alternative outcomes between 0 and 1, rather than between 0 and 1. The sigmoid function formula is shown below.

$$S(x) = \frac{1}{1 + \mathrm{e}^{-\mathrm{x}}}$$

3.6.2 SVM

The primary concept of the Support Vector Machine is that it is based on statistical learning theory [28, 29]. The data was displayed in a hyperplane with n-D space in the SVM method (where n is a feature number). Here, we employ a two-dimensional surface plane with a line dividing the space into two distinct pieces. Support vectors are locations that pass across the established marginal plane parallel to the found hyperplane. Figure 3.3 shows SVM models hyperplane.

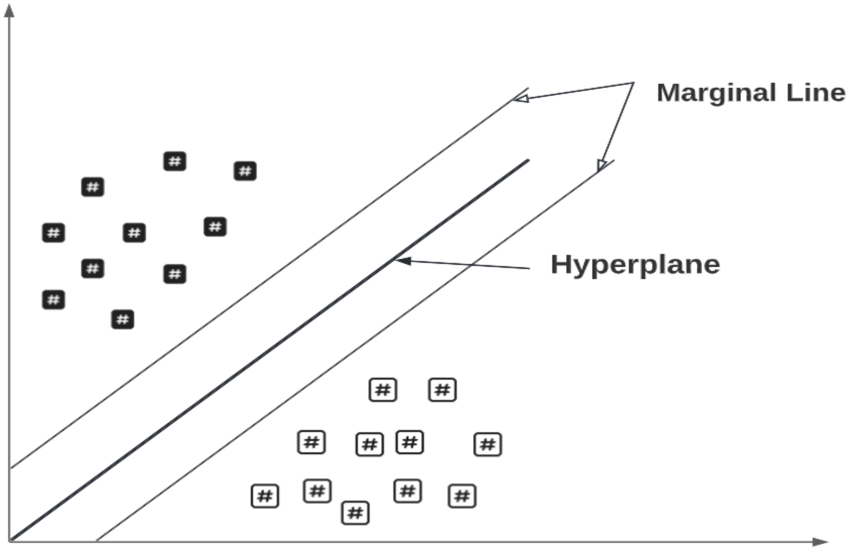


Figure 3.3 SVM Model Hyperplane

3.6.3 KNN

K-Nearest Neighbors (KNN) is commonly referred to as a lazy type algorithm since it does not foresee but instead memorizes the process. It classifies new points based on their similarities using Euclidean distances. To train the k-nearest-neighbor classifier, the Euclidean distance between a test sample and the requisite training samples is commonly utilized [30]. KNN is widely used because of its simple explanation and speedy calculation time. The Euclidean Distance formula is shown below.

$$Euclidean\ Distance = \sqrt{(x_2 - x_1) + (y_2 - y_1)}$$

3.6.4 Decision Tree

For classification and regression, a supervised learning model known as a decision tree is used.

problem. It splits the dataset by selecting a feature. Features can be expressed as a nominal or continuous number. It is a tree-structured classifier. The internal nodes of this classifier represent a property of the dataset, The result is represented by the leaf nodes, while the decision rules are represented by the branches [31]. The decision tree algorithm starts at the root node of the tree. It compares the values of the root. This comparison is used to determine whether to go to the next node or to follow the brunch. This process is repeated until the leaf node is reached [32]. Figure 3.5.4 demonstrates the decision tree working principle.

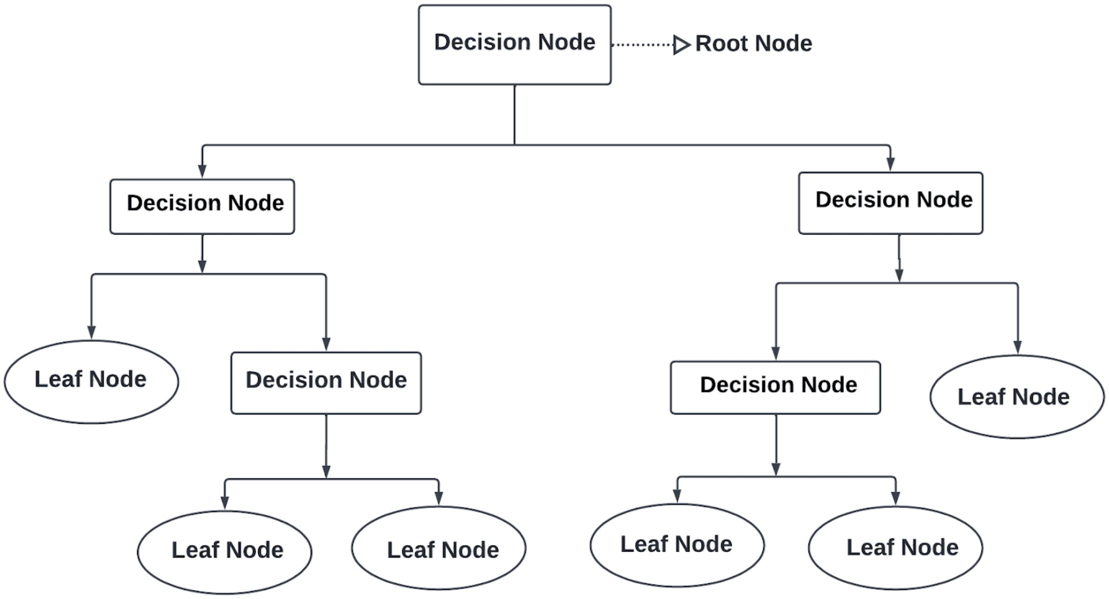


Figure 3.4: Decision Tree Architecture.

3.6.5 Random Forest

Random Forest is based on the ensemble learning idea [33]. When numerous classifiers tackle a difficult issue and improve the performance of a model, the process is referred to as an ensemble. To increase the projected accuracy of other datasets, the Random Forest Classifier makes various judgments on subsets of other datasets and averages them.

3.6.6 Naïve Bayes

In the field of classification, Naive Bayes classifiers outperform more powerful alternatives, particularly for small data sets [34]. Its data are mutually independent and are based on the Bayes theorem and the term Naive. According to the dataset, Naive Bayes is a linear classifier that is often robust, easy to use, and operates quickly with improved accuracy [35]. When the dataset is predefined and labeled, the naive Bayes model works well. The naive Bayes formula is illustrated below.

$$P\left(\frac{x}{y}\right) = \frac{\mathrm{P}\left(\frac{y}{x}\right)\mathrm{P(x)}}{\mathrm{P(y)}}$$

P(x/y) is the hypothesis's probability, while P(y/x) is proof that the hypothesis is correct. P(x) is the Prior probability, whereas P(y) is the Marginal probability.

3.6.7 XG-Boost

The XG-Boost is a Gradient Boosted decision tree execution. In this case, decision trees are generated in the following sequence. The fed decision tree predicts the results by assigning weights to each independent variable. It selects the best result from among them.

3.7 Deep Learning-Based Bi-LSTM Classification Model

The recurrent model is the foundation of the Bidirectional Long Short-Term (Bi-Lstm), which analyzes text as sequential information [36]. The Bi-Lstm models make use of data from the neurons' previous status [37]. The output of each LSTM is combined using their total as information processing moves forward and backward simultaneously [38].

Bi-LSTM models have been demonstrated to be useful in voice recognition problems [39], being a cutting-edge model for classifying sequential input into several classes. Because documents are word sequences and our problem is multi-class, Bi-LSTM is an appropriate model. Our model architecture is made up of three primary layers and 250 input tokens. Each token in a distributed array of 100 dimensions is transformed by the layer. The recurrent layer has two hidden LSTM models, one forward and one backward, each with 200 memory blocks and one cell. The outputs of the two hidden LSTMs are added together. The last layer has 5 output neurons and a SoftMax activation function. Figure 3.5 shows our Bi-Lstm Architecture Diagram.

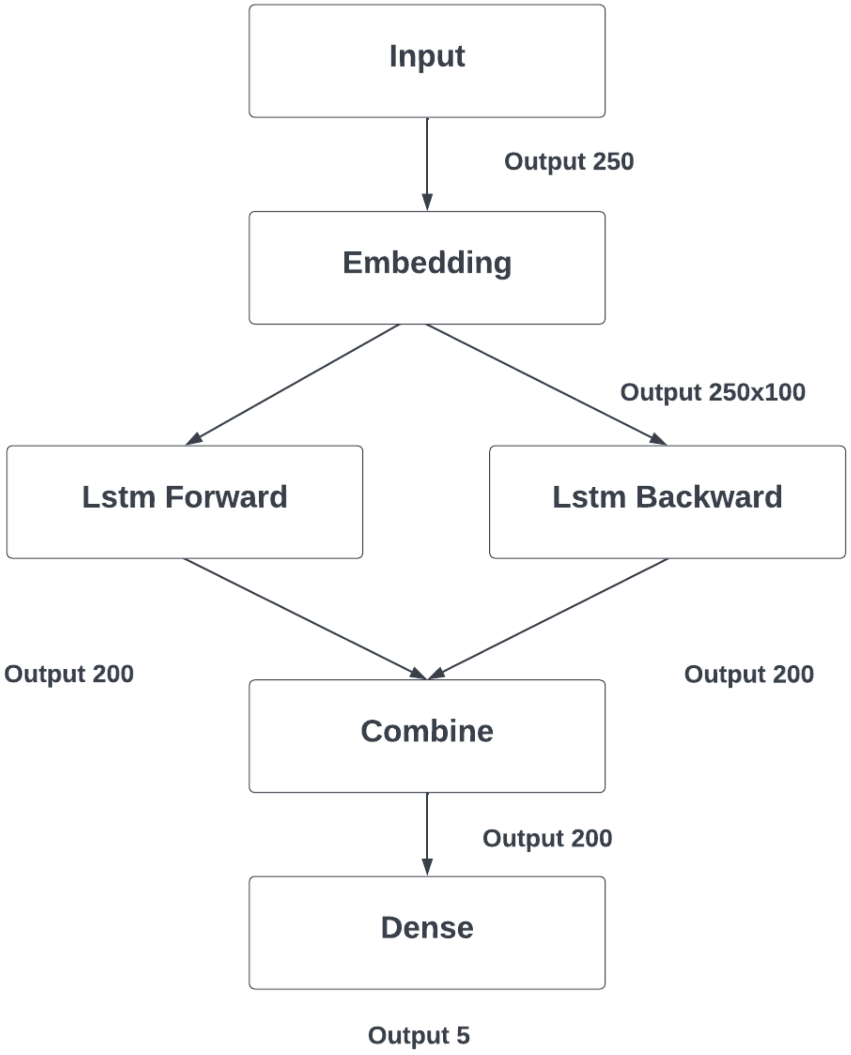


Figure 3.5: Bi-Lstm Architecture Diagram

3.8 Transformer Bert-Based uncased Classification Model

BERT(Bidirectional Encoder Representation from Transformer) is unquestionably a milestone in the application of Machine Learning to Natural Language Processing. Many other NLP domains have been attracted to BERT's strong performance, and efforts have been made to build BERT variants trained on multilingual data [40]. Transformer is a well-known attention mechanism that discovers contextual linkages between words in a text. A simple Transformer consists of an encoder that reads the text input and a decoder that produces a task prediction. Bert makes use of a transformer. Mask Language Modeling, which predicts the job, is relevant for our investigation. Predict the author in our situation. BERT just requires the encoder component because its objective is to construct a language representation model. The total of the token embeddings, segmentation embeddings, and position embeddings is the input embeddings. Figure 3.6 shows the combination of three embeddings.

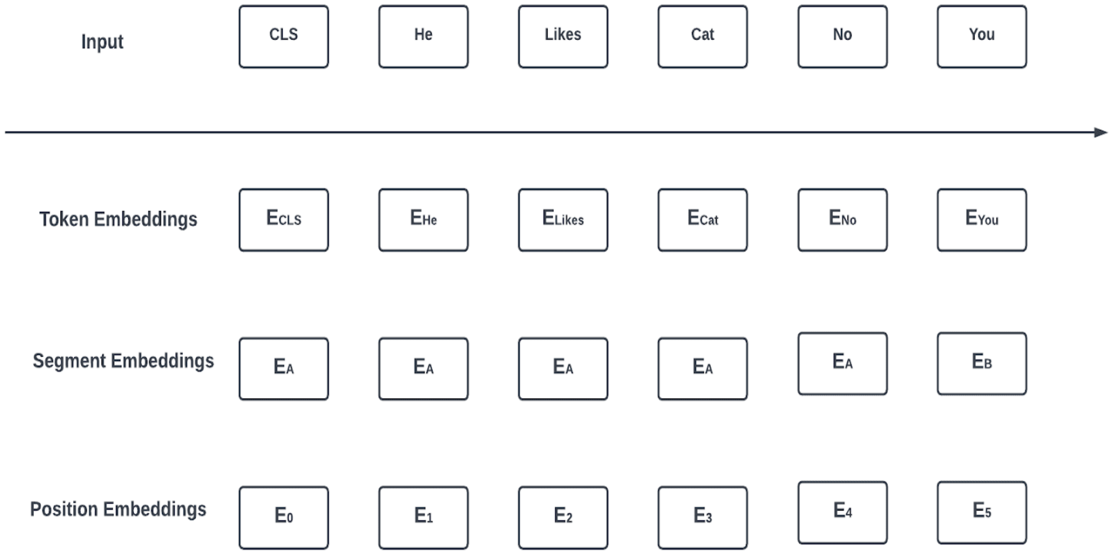


Figure 3.6: The input representation for BERT

We choose the BERT-Base(un-cased) model among the four types of pre-trained versions of BERT that are suited for our model. It has 12 layers, 768 hidden nodes, 12 attention heads, and 110M parameters. Figure 3.7 depicts a basic un-cased bert-based architecture.

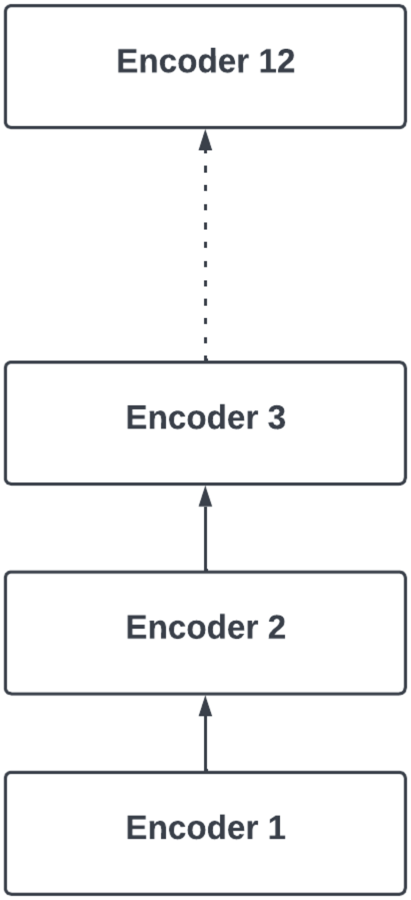


Figure 3.7: Bert-Base un-cased architecture.

3.9 Model Evaluation

A model cannot be evaluated, only on the basis of training and testing accuracy. We require some reports in order to evaluate the model. First and foremost, in order to obtain reliable results from our model, we must use cross validation. Following that, we must create a classification report in order to analyze our model.

3.9.1 K-Fold Cross Validation

Cross-validation is a validation approach that allows us to assess the correctness of a model. It randomly splits data. As a result, test data may contain information that is not present in train data. As a result, our model's accuracy might be variable. k-Fold Cross-validation assists us in resolving this issue. There is a parameter(k) in this approach that represents the number of folds that a dataset is divided into. Cross-validation divides the dataset k times at random and evaluates how well the model performs when confronted with any randomly selected unseen test data.

3.9.2 Confusion Matrix, ROC Curve and Classification Report

The confusion matrix is generally used in the evaluation of multi-class or single-label classification models to evaluate performance [41]. Sensitivity, specificity, accuracy, and the area under the ROC curve are all common discrimination measurements (or, equivalently, the c-index). There are statistical tests for all these criteria to assess if one model outperforms another in discrimination ability [42]. ROC curves are a measure of a classifier's prediction quality that compares and visualizes the tradeoff between the model's sensitivity and specificity. We evaluate our models through different evaluation metrics, such as precision, recall, F1 score, and model accuracy. These formulas are given below,

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \times 100\%$$

$$Recall = \frac{TP}{TP + FN} \times 100$$

$$Precision = \frac{TP}{TP + FP} \times 100\%$$

$$F1\,Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \times 100\%$$

3.10 Implementation Requirements

Our work is classified as Bengali NLP. We gathered Bengali data from many sources in order to build a system that can determine authorship from Bengali paragraph data. To analyze and assess the complete job, we require a high-end computer system with a GPU and other essential instruments, since we use deep learning and the BERT model. The hardware, software, and advanced tools required to do this job are listed below.

- Hardware and Software:
    1. Intel Core i7 11th gen integrated with 16GB ram
    2. 1 TB Hard Disk
    3. Google Collaboratory with 12GB GPU and 350GB ram
    4. High-Speed Broadband Internet Connection

- Advance Libraries and Tools:
    1. Python 3.8 or upper.
    2. Pandas
    3. NumPy
    4. NLTK
    5. Matplotlib
    6. Scikit-Learn
    7. Transformers
    8. Torch
    9. Karas

CHAPTER 4

Experimental Results and Discussion

4.1 Experimental Setup

We choose Google Colab as our working environment. We employ a variety of built-in libraries to achieve our goal. For example, the Panda preprocessing library, Matplotlib, Seaborn, Nltk, and others. Scikit-learn (Sklearn) is Python's most user-friendly and capable machine-learning library. It provides a wide range of powerful machine learning and statistical modeling approaches, particularly classification models. Sklearn is used to train our classification models.

4.2 Experimental Result and Analysis

We will go through Three different types of works, such as Machine Learning Models, Deep Learning Models, and BERT Models. These are given below.

4.2.1 Experimental Result and Analysis for Machine Learning Models

This is primarily a multiclass problem. The five members of the class are Humayun Ahmed (H.A), Rabindranath Tagore (R.T), Muhammad Zafar Iqbal (M.J.I), Kazi Nazrul Islam (K.N.I), and Sarat Chandra Chattopadhyay (S.C.C). Different performance criteria, such as model accuracy, precision, recall, and F1-score, are used to evaluate the seven classifier models. Table 4.1 displays the performance metrics measurement.

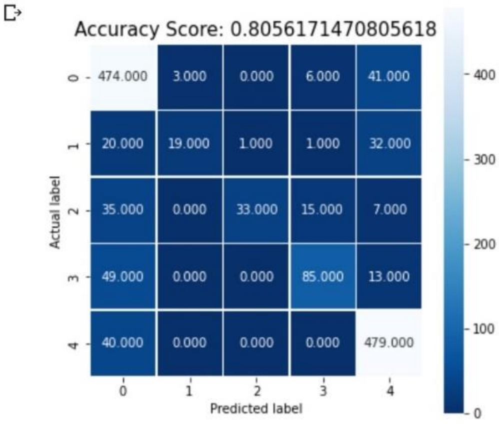
Table 4.1: Classification Report and Accuracy.

	Class	Precision	Recall	F1-Score	Accuracy
Naive Bayes	H.A	0.67	0.87	0.75	71%
	R.T	0.00	0.00	0.00	
	M.J.I	0.00	0.00	0.00	
	K.N.I	0.91	0.07	0.13	
	S.C	0.73	0.97	0.84	

	Class	**Precision**	**Recall**	**F1-Score**	**Accuracy**
Logistic Regression	H.A	0.77	0.90	0.83	81%
	R.T	0.86	0.26	0.40	
	M.J.I	0.97	0.37	0.53	
	K.N.I	0.79	0.58	0.67	
	S.C	0.84	0.92	0.88	
KNN	H.A	0.58	0.89	0.70	67%
	R.T	0.62	0.25	0.35	
	M.J.I	0.79	0.34	0.48	
	K.N.I	0.74	0.19	0.30	
	S.C	0.81	0.70	0.75	
Random Forest	H.A	0.73	0.87	0.80	77%
	R.T	0.85	0.15	0.26	
	M.J.I	0.95	0.21	0.35	
	K.N.I	0.84	0.42	0.56	
	S.C	0.79	0.95	0.86	
Decision Tree	H.A	0.64	0.66	0.65	64%
	R.T	0.33	0.29	0.31	
	M.J.I	0.52	0.32	0.40	
	K.N.I	0.59	0.56	0.57	
	S.C	0.71	0.76	0.73	
SVM	H.A	0.78	0.91	0.84	82%
	R.T	0.80	0.38	0.52	
	M.J.I	0.96	0.48	0.64	
	K.N.I	0.86	0.61	0.71	
	S.C	0.85	0.91	0.88	
XG-Boost	H.A	0.72	0.86	0.79	77%
	R.T	0.68	0.32	0.43	
	M.J.I	0.95	0.40	0.56	
	K.N.I	0.78	0.50	0.61	
	S.C	0.81	0.88	0.84	

Working with five writers, it generated a confusion matrix of 5x5. Seven classification models' confusion matrix is given below. Figure 4.1 shows the machine learning models confusion matrix.

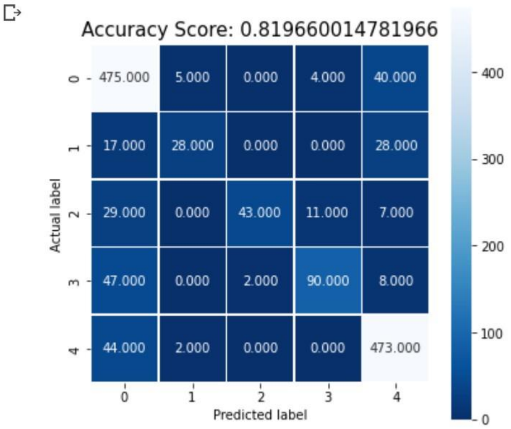
Logistic Regression



Accuracy Score: 0.8056171470805618

Actual label \ Predicted label	0	1	2	3	4
0	474.000	3.000	0.000	6.000	41.000
1	20.000	19.000	1.000	1.000	32.000
2	35.000	0.000	33.000	15.000	7.000
3	49.000	0.000	0.000	85.000	13.000
4	40.000	0.000	0.000	0.000	479.000

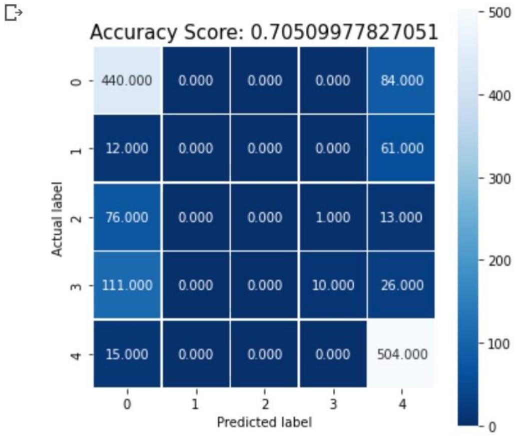
SVM



Accuracy Score: 0.819660014781966

Actual label \ Predicted label	0	1	2	3	4
0	475.000	5.000	0.000	4.000	40.000
1	17.000	28.000	0.000	0.000	28.000
2	29.000	0.000	43.000	11.000	7.000
3	47.000	0.000	2.000	90.000	8.000
4	44.000	2.000	0.000	0.000	473.000

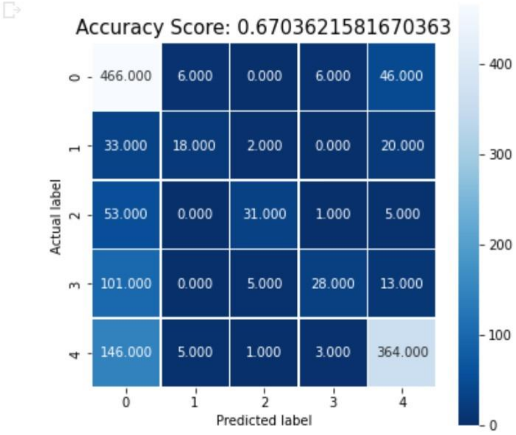
Naive Bayes



Accuracy Score: 0.70509977827051

Actual label \ Predicted label	0	1	2	3	4
0	440.000	0.000	0.000	0.000	84.000
1	12.000	0.000	0.000	0.000	61.000
2	76.000	0.000	0.000	1.000	13.000
3	111.000	0.000	0.000	10.000	26.000
4	15.000	0.000	0.000	0.000	504.000

KNN

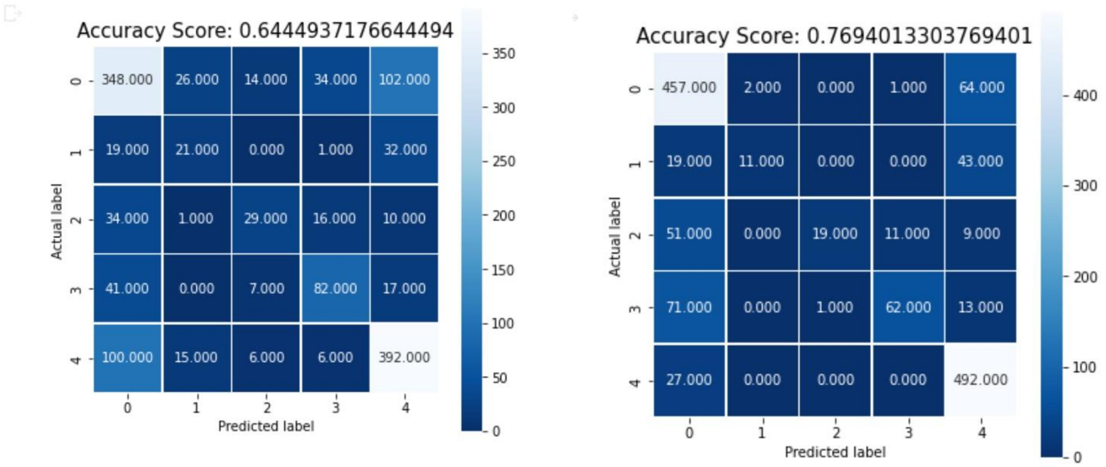


Accuracy Score: 0.6703621581670363

Actual label \ Predicted label	0	1	2	3	4
0	466.000	6.000	0.000	6.000	46.000
1	33.000	18.000	2.000	0.000	20.000
2	53.000	0.000	31.000	1.000	5.000
3	101.000	0.000	5.000	28.000	13.000
4	146.000	5.000	1.000	3.000	364.000

Decision Tree

Random Forest



XB-Boost

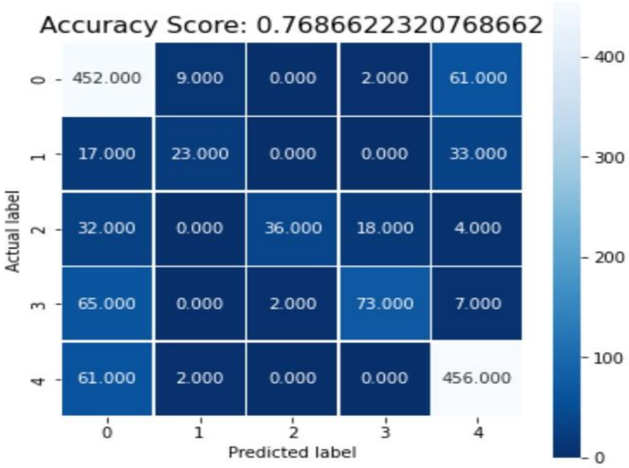


Figure 4.1: Seven Classification Models Confusion Matrix

The ROC curve depicts the true positive rate on the Y axis and the false positive rate on the X axis. We all know that the greater the AUC, the better the model in general. As shown in Figure 4.2, all of our classes' AUC scores are excellent, and this result is based only on the Logistic Regression model which one's accuracy is 81%.

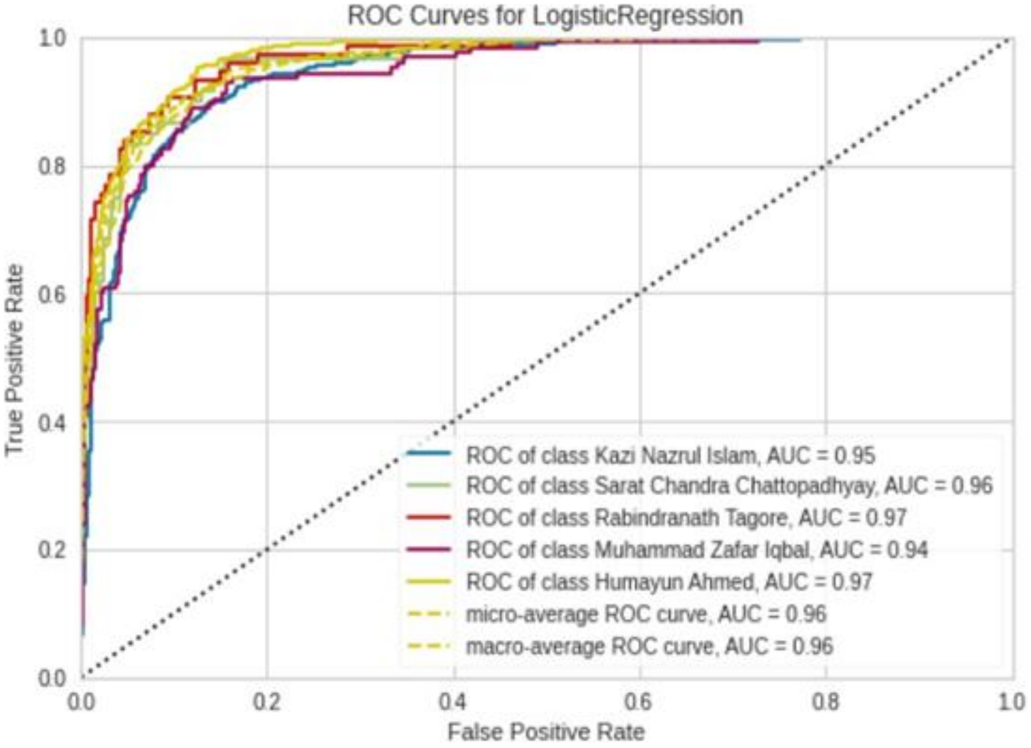


Figure 4.2: Roc Curve of Logistic Regression.

4.2.2 Experimental Result and Analysis for Bi-Lstm Model.

An important factor is how the loss function and model accuracy perform on training and test data. Figure 4.3 depicts the Loss and Accuracy of the model based on Training and Test Data.

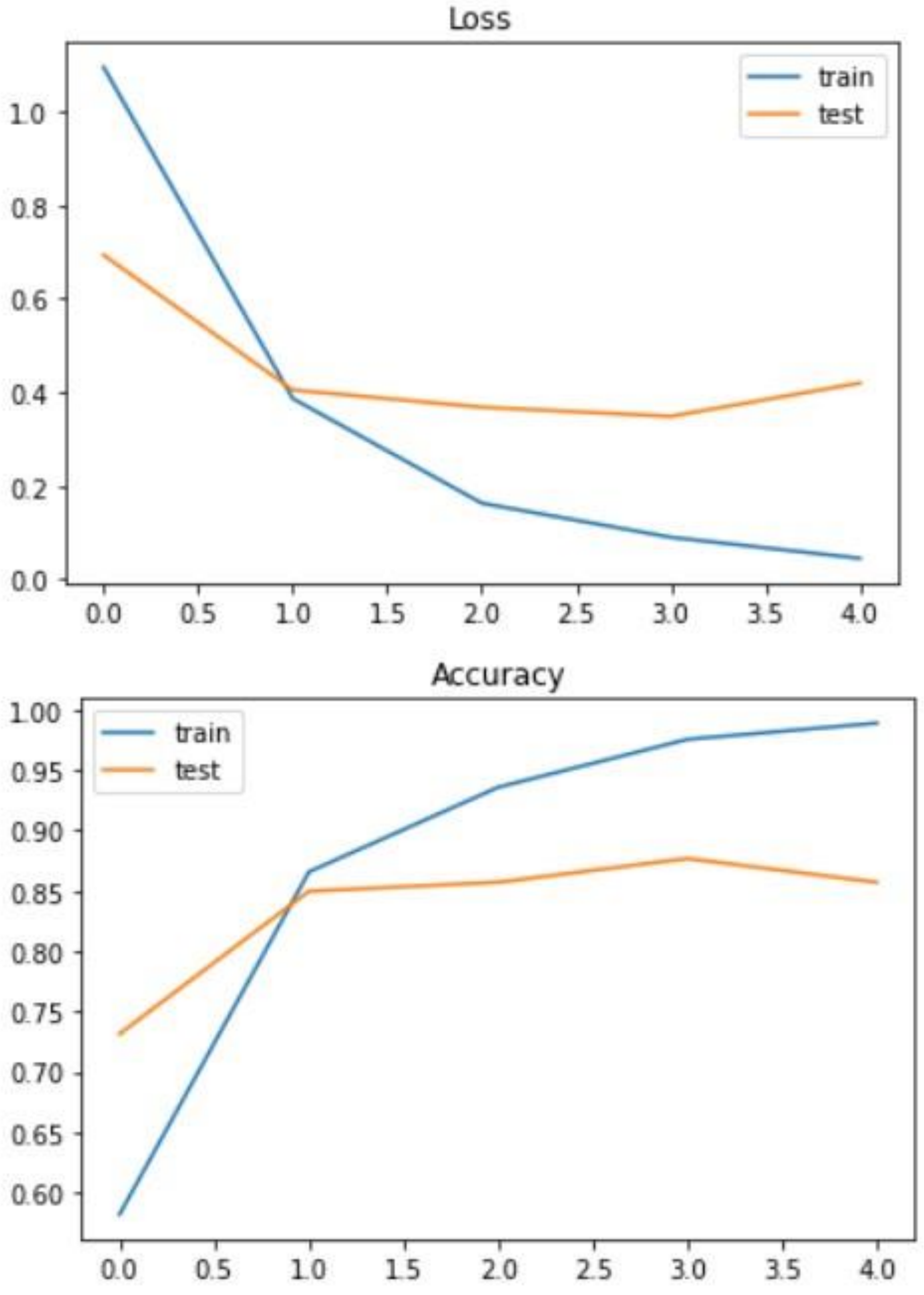


Figure 4.3: Loss and Accuracy based on Training and Test Data.

We established 5 epochs and everything works well. Our categorization outcome is shown in the 4.2 table.

Table 4.2: Classification Report of Bi-Lstm model.

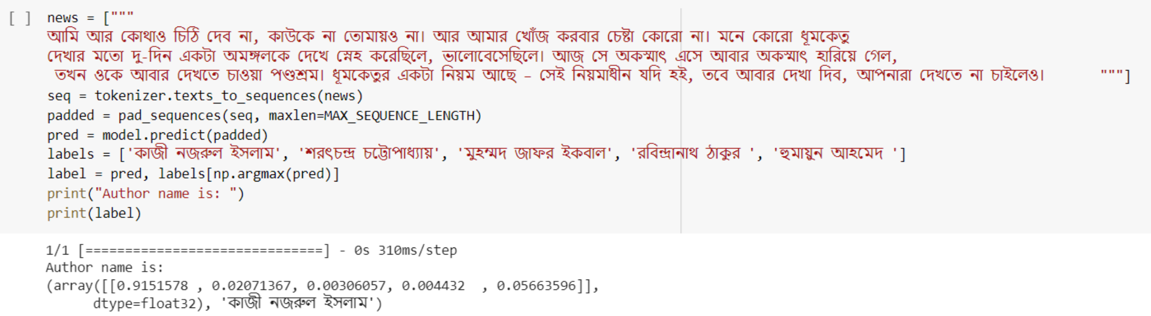
Epoch	Model loss function	Model Accuracy	Validation Loss function	Validation Accuracy
1	1.0954	0.5820	0.6937	0.7315
2	0.3870	0.8658	0.4052	0.8498
3	0.1630	0.9359	0.3683	0.8571
4	0.0900	0.9756	0.3488	0.8768
5	0.0451	0.9890	0.4200	0.8571

We shall now identify the true author from the given text. This passage was taken from কাজী নজরুল ইসলাম novel, which was not included in our dataset. Table 4.3 demonstrates the কাজী নজরুল ইসলাম paragraph that is not included in our dataset.

Table 4.3: Author paragraph and not knowing who is author.

Author’s paragraph	Author's name
আমি আর কোথাও চিঠি দেব না, কাউকে না তোমায়ও না। আর আমার খোঁজ করবার চেষ্টা কোরো না। মনে কোরো ধূমকেতু দেখার মতো দু-দিন একটা অমঙ্গলকে দেখে স্নেহ করেছিলে, ভালোবেসেছিলে। আজ সে অকস্মাৎ এসে আবার অকস্মাৎ হারিয়ে গেল, তখন ওকে আবার দেখতে চাওয়া পণ্ডশ্রম। ধূমকেতুর একটা নিয়ম আছে – সেই নিয়মাধীন যদি হই, তবে আবার দেখা দিব, আপনারা দেখতে না চাইলেও।	??

Figure 4.4 shows that our model correctly identifies the correct author.



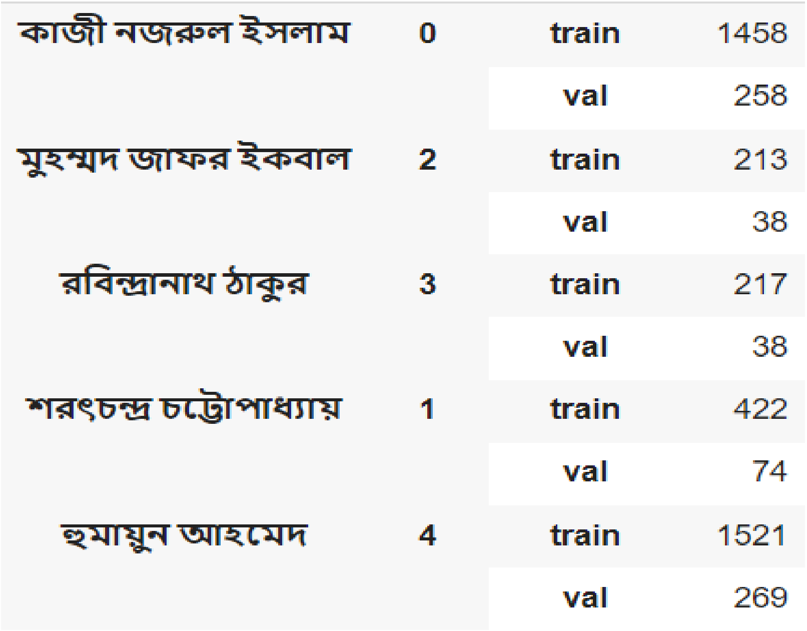
```
[ ]  news = ["""
     আমি আর কোথাও চিঠি দেব না, কাউকে না তোমায়ও না। আর আমার খোঁজ করবার চেষ্টা কোরো না। মনে কোরো ধূমকেতু
     দেখার মতো দু-দিন একটা অমঙ্গলকে দেখে স্নেহ করেছিলে, ভালোবেসেছিলে। আজ সে অকস্মাৎ এসে আবার অকস্মাৎ হারিয়ে গেল,
      তখন ওকে আবার দেখতে চাওয়া পণ্ডশ্রম। ধূমকেতুর একটা নিয়ম আছে - সেই নিয়মাধীন যদি হই, তবে আবার দেখা দিব, আপনারা দেখতে না চাইলেও।       """]
     seq = tokenizer.texts_to_sequences(news)
     padded = pad_sequences(seq, maxlen=MAX_SEQUENCE_LENGTH)
     pred = model.predict(padded)
     labels = ['কাজী নজরুল ইসলাম', 'শরৎচন্দ্র চট্টোপাধ্যায়', 'মুহম্মদ জাফর ইকবাল', 'রবিন্দ্রানাথ ঠাকুর ', 'হুমায়ুন আহমেদ ']
     label = pred, labels[np.argmax(pred)]
     print("Author name is: ")
     print(label)

     1/1 [==============================] - 0s 310ms/step
     Author name is:
     (array([[0.9151578 , 0.02071367, 0.00306057, 0.004432  , 0.05663596]],
           dtype=float32), 'কাজী নজরুল ইসলাম')
```

Figure 4.4: Correctly predict the Author.

4.2.3 Experimental Result and Analysis for BERT Based un-cased Model.

We divide the dataset into two halves, one for training and one for testing. Figure 4.5 depicts the training and test/validation data in numeric format.



কাজী নজরুল ইসলাম	0	train	1458
		val	258
মুহম্মদ জাফর ইকবাল	2	train	213
		val	38
রবিন্দ্রানাথ ঠাকুর	3	train	217
		val	38
শরৎচন্দ্র চট্টোপাধ্যায়	1	train	422
		val	74
হুমায়ুন আহমেদ	4	train	1521
		val	269

Figure 4.5: Train test split of our dataset

We set up 5 epochs to build our model. The 4.4 table displays the results of our classification.

Table 4.4: Classification Report of BERT Based un-cased model.

Epoch	Training Loss	Validation Loss	F1-Score
1	0.797	0.749	0.80
4	0.448	0.707	0.862
8	0.203	0.664	0.889
14	0.03179	0.76149	0.908
15	0.03265	0.72129	0.912

The crucial moment has arrived. If we look at the 4.5 table, we can see that as data increases, so does the accuracy ratio. As a result, the greater the amount of data for the author class, the higher the accuracy ratio. Because Transformer-based models learn the linguistic context, a large amount of data is required. Our 4.5 table demonstrates this.

Table 4.5: Correctly identity paragraph each author's accuracy ratio.

Author's name	Validation Data	Correctly Identify Data	Accuracy Ratio
মুহম্মদ জাফর ইকবাল	38	19	50%
রবিন্দ্রানাথ ঠাকুর	38	24	63%
শরৎচন্দ্র চট্টোপাধ্যায়	74	62	84%
কাজী নজরুল ইসলাম	258	233	90%
হুমায়ুন আহমেদ	269	263	98%

4.3 Discussion

SVM has the best overall model accuracy of the seven classifier models. SVM models from the R.T., M.J.I., and K.N.I classes had the highest model accuracy, recall, and f1 score. However, in Logistic Regression models, the H.A. and S.C classes yield the highest accuracy, recall, and f1 score. Bi-Lstm also performs fairly enough and it gives 98% model accuracy but validation accuracy is 88%. But the Transformer based classification model, BERT Based un-cased model gives overall best performance. More the data, the better the class is. Here, হুমায়ুন আহমেদ class accuracy ratio 98%, which is almost 100%.

CHAPTER 5

Impact On Society, Environmental, and Sustainability

5.1 Impact on Society

We currently live in a modern world where individuals place a high value on digital property. We are all connected thanks to the internet and social media, especially in the society in which we live. We are highly mindful of other people's digital property, including authors of novels, poetry, and books, among others. However, there is a dilemma in the digital world where some dishonest people steal other people's digital property and attempt to claim it as their own. We developed a methodology that can identify the true author of electronically generated books, novels, poetry, etc. to avoid this type of issue.

5.2 Impact on Environment

We will aim to construct a fantastic project so that certain digital property can remain secure utilizing the copyright problem of Bengali writers so that it will have a significant influence on society and the environment. Because an environment is created by a society, and societies are created by people. People should have enough knowledge of digital property in this digital arena, as well as their native language. We hope that our effort will contribute to the development of an environment in which any Bengali writer can create any sort of Bengali poetry, book, or other work in digital form without fear of theft, and where people can identify the correct author.

5.3 Ethical Aspects

As there is no powerful system that can recognize Bengali writers' novels, poetry, books, and other digital property, we try to create a model that can do so in this technological day. Our concept also has certain ethical implications. They're listed below.

1. Stop wicked individuals from stealing the digital property of a Bengali author.
2. The Bengali writers can be predicted from the provided text.
3. Contribute a system in our language and enhance our language value.

4. The project and website will be open source, which means that anybody may contribute to our project if they think it would be useful, and users will be able to easily identify the correct creators of a particular intellectual property.

5.4 Sustainability Plan

People nowadays are unwilling to use other people's comments, phrases, and sentences unless they are aware of a copyright issue. Using another person's digital property without permission is a crime in the English language. As a result, we developed a model for protecting intellectual property in the Bengali language, such as poetry, novels, author books, and so on. We hope that the next generation understands digital property, how to utilize it, and what not to do with it. Our deep learning models are the best for our system, but they require a large amount of data. This massive amount of data will be added gradually, and the website will be made open to everyone. We will also need to consider how to support this model indefinitely.

CHAPTER 6

Summary, Conclusion, Recommendation, and Implication For Future Research

6.1 Summary

Our work is connected to the Bengali NLP. Working with NLP is a difficult task for researchers. The dataset is the most significant part of any research project. In this project, we are developing a machine learning, deep learning, and BERT model to recognize the Bangla author from a given text. We encountered certain difficulties when developing this model, as well as other issues. All of the stages and work summary are provided here, step by step.

1. Planning about this work
2. formulation of a problem
3. Data collection from various books and websites.
4. Data Labeling
5. Data Annotation.
6. Data cleaning
7. Data Feature selection.
8. Data Vectorization
9. Split data into train and test.
10. Model Selection
11. Build environment in google Collaboratory.
12. Model Evaluation.
13. Measure performance of Model.

6.2 Conclusion

Author detection is a feature that makes determining which article belongs to which author intuitive. The support vector machine produced the best results of the seven classification reports, with 82% accuracy and 85% precision. The Bi-Lstm provides 88% validation accuracy, whereas the BERT Model provides 91% validation accuracy. Our work might be used on the backend of social media, forums, blogs, or websites to identify the original author or detect plagiarism, among other things. Author detection is a method for identifying dishonest people who try to pass off someone else's content as their own, which is prohibited. Everyone should be more conscious of this situation.

6.3 Future Scope

We've previously shown that massive amounts of data are essential in the transformer-based approach. We hope to do so and implement the newest transformer-based approach known as XL-net. In practice, we shall include this line in our Bengali website or preserve the author's novels.

Reference

[1] N. Islam, M. M. Hoque, and M. R. Hossain, “Automatic authorship detection from Bengali text using stylometric approach,” *IEEE Xplore*, Dec. 01, 2017. https://ieeexplore.ieee.org/abstract/document/8281793 (accessed Dec. 28, 2022).

[2] T. Chakraborty, “Authorship Identification in Bengali Literature: a Comparative Analysis,” *arXiv:1208.6268 [cs]*, Feb. 2013, Accessed: Dec. 28, 2022. [Online]. Available: https://arxiv.org/abs/1208.6268

[3] S. Das and P. Mitra, “Author Identification in Bengali Literary Works,” *Lecture Notes in Computer Science*, pp. 220–226, 2011, doi: 10.1007/978-3-642-21786-9_37.

[4] A. Rocha *et al.*, “Authorship Attribution for Social Media Forensics,” *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 1, pp. 5–33, Jan. 2017, doi: 10.1109/TIFS.2016.2603960.

[5] F. Mosteller and D. L. Wallace, *Applied Bayesian and Classical Inference: The Case of The Federalist Papers*. Springer Science & Business Media, 2012. Accessed: Dec. 29, 2022. [Online]. Available: https://books.google.com.bd/books?hl=en&lr=&id=LJXaBwAAQBAJ&oi=fnd&pg=PA1&dq=Mosteller+F

[6] E. Stamatatos, “A survey of modern authorship attribution methods,” *Journal of the American Society for Information Science and Technology*, vol. 60, no. 3, pp. 538–556, Mar. 2009, doi: 10.1002/asi.21001.

[7] M. Koppel, J. Schler, S. Argamon, and E. Messeri, “Authorship attribution with thousands of candidate authors,” *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR ’06*, 2006, doi: 10.1145/1148170.1148304.

[8] B. Diri and M. Fatih Amasyali, “Automatic Author Detection for Turkish Texts.” Accessed: Dec. 29, 2022. [Online]. Available: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=f1422461024fcec79c94fe2671923ce79be0e4ef

[9] R. Zheng, J. Li, H. Chen, and Z. Huang, “A framework for authorship identification of online messages: Writing-style features and classification techniques,” *Journal of the American Society for Information Science and Technology*, vol. 57, no. 3, pp. 378–393, 2006, doi: 10.1002/asi.20316.

[10] E. Stamatatos, “Author Identification Using Imbalanced and Limited Training Texts,” *IEEE Xplore*, Sep. 01, 2007. https://ieeexplore.ieee.org/abstract/document/4312893 (accessed Feb. 12, 2022).

[11] Y. Seroussi, I. Zukerman, and F. Bohnert, “Authorship Attribution with Topic Models,” *Computational Linguistics*, vol. 40, no. 2, pp. 269–310, Jun. 2014, doi: 10.1162/coli_a_00173.

[12] M. Ghaeini, “Intrinsic Author Identification Using Modified Weighted KNN Notebook for PAN at CLEF 2013.” Accessed: Dec. 29, 2022. [Online]. Available: https://ceur-ws.org/Vol-1179/CLEF2013wn-PAN-Ghaeini2013.pdf

[13] R. Schwartz, O. Tsur, A. Rappoport, and M. Koppel, “Authorship Attribution of Micro-Messages,” Association for Computational Linguistics, 2013. Accessed: Dec. 29, 2022. [Online]. Available: https://aclanthology.org/D13-1193.pdf

[14] M. Ebrahimpour, T. J. Putniņš, M. J. Berryman, A. Allison, B. W.-H. . Ng, and D. Abbott, “Automated Authorship Attribution Using Advanced Signal Classification Techniques,” *PLoS ONE*, vol. 8, no. 2, p. e54998, Feb. 2013, doi: 10.1371/journal.pone.0054998.

[15] C. Zhang, X. Wu, Z. Niu, and W. Ding, “Authorship identification from unstructured texts,” *Knowledge-Based Systems*, vol. 66, pp. 99–111, Aug. 2014, doi: 10.1016/j.knosys.2014.04.025.

[16] F. Howedi and M. Mohd, “Text Classification for Authorship Attribution Using Naive Bayes Classifier with Limited Training Data Authorship Attribution of Short Historical Arabic Texts using Stylometric Features and a KNN Classifier with Limited Training Data View project Text Classification for Authorship Attribution Using Naive Bayes Classifier with Limited Training Data,” *Online)*, vol. 5, no. 4, 2014.

[17] O. Binnani, “Author Identification using Traditional Machine Learning Models,” *Computer Science and Information Technology Trends*, Aug. 2022, doi: 10.5121/csit.2022.121402.

[18] M. L. Brocardo, I. Traore, I. Woungang, and M. S. Obaidat, “Authorship verification using deep belief network systems,” *International Journal of Communication Systems*, vol. 30, no. 12, p. e3259, Jan. 2017, doi: 10.1002/dac.3259.

[19] C. Akimushkin, D. R. Amancio, and O. N. Oliveira, “Text Authorship Identified Using the Dynamics of Word Co-Occurrence Networks,” *PLOS ONE*, vol. 12, no. 1, p. e0170527, Jan. 2017, doi: 10.1371/journal.pone.0170527.

[20] C. Akimushkin, D. R. Amancio, and O. N. Oliveira, “Text Authorship Identified Using the Dynamics of Word Co-Occurrence Networks,” *PLOS ONE*, vol. 12, no. 1, p. e0170527, Jan. 2017, doi: 10.1371/journal.pone.0170527.

[21] R. Abascal-Mena and E. López-Ornelas, “Author detection: Analyzing tweets by using a Naïve Bayes classifier,” *Journal of Intelligent & Fuzzy Systems*, pp. 1–9, Jun. 2020, doi: 10.3233/jifs-179894.

[22] A. S. Hossain, N. Akter, and Md. S. Islam, “A Stylometric Approach for Author Attribution System Using Neural Network and Machine Learning Classifiers,” *Proceedings of the International Conference on Computing Advancements*, Jan. 2020, doi: 10.1145/3377049.3377079.

[23] A. S. Hossain, N. Akter, and Md. S. Islam, “A Stylometric Approach for Author Attribution System Using Neural Network and Machine Learning Classifiers,” *Proceedings of the International Conference on Computing Advancements*, Jan. 2020, doi: 10.1145/3377049.3377079.

[24] R. Deibel and D. Löfflad, “Style Change Detection on Real-World Data using an LSTM-powered Attribution Algorithm Notebook for PAN at CLEF 2021.” Accessed: Dec. 29, 2022. [Online]. Available: https://ceur-ws.org/Vol-2936/paper-163.pdf

[25] O. Suissa, A. Elmalech, and M. Zhitomirsky-Geffet, “Text analysis using deep neural networks in digital humanities and information science,” *Journal of the Association for Information Science and Technology*, vol. 73, no. 2, pp. 268–287, Jun. 2021, doi: 10.1002/asi.24544.

[26] S. Phani, S. Lahiri, and A. Biswas, “A machine learning approach for authorship attribution for Bengali blogs,” *IEEE Xplore*, Nov. 01, 2016. https://ieeexplore.ieee.org/abstract/document/7875984 (accessed Dec. 29, 2022).

[27] S. Dreiseitl and L. Ohno-Machado, “Logistic regression and artificial neural network classification models: a methodology review,” *Journal of Biomedical Informatics*, vol. 35, no. 5–6, pp. 352–359, Oct. 2002, doi: 10.1016/s1532-0464(03)00034-0.

[28] D. Madigan, A. Genkin, D. Lewis, S. Argamon, D. Fradkin, and L. Ye, “Author Identification on the Large Scale.”

[29] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/bf00994018.

[30] L. Peterson, “K-nearest neighbor,” *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009, doi: 10.4249/scholarpedia.1883.

[31] S. Segarra, M. Eisen, and A. Ribeiro, “Authorship Attribution Through Function Word Adjacency Networks,” *IEEE Transactions on Signal Processing*, vol. 63, no. 20, pp. 5464–5478, Oct. 2015, doi: 10.1109/tsp.2015.2451111.

[32] R. Abascal-Mena and E. López-Ornelas, “Author detection: Analyzing tweets by using a Naïve Bayes classifier,” *Journal of Intelligent & Fuzzy Systems*, pp. 1–9, Jun. 2020, doi: 10.3233/jifs-179894.

[33] O. Sagi and L. Rokach, “Ensemble learning: A survey,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, Feb. 2018, doi: 10.1002/widm.1249.

[34] S. Raschka, “Naive Bayes and Text Classification I - Introduction and Theory,” *arXiv.org*, 2014. https://arxiv.org/abs/1410.5329 (accessed Jul. 09, 2019).

[35] P. Domingos and M. Pazzani, “On the Optimality of the Simple Bayesian Classifier under Zero-One Loss,” *Machine Learning*, vol. 29, no. 2/3, pp. 103–130, 1997, doi: 10.1023/a:1007413511361.

[36] F. A. Braz *et al.*, “Document classification using a Bi-LSTM to unclog Brazil’s supreme court,” *arXiv:1811.11569 [cs, stat]*, Nov. 2018, Accessed: Dec. 29, 2022. [Online]. Available: https://arxiv.org/abs/1811.11569

[37] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

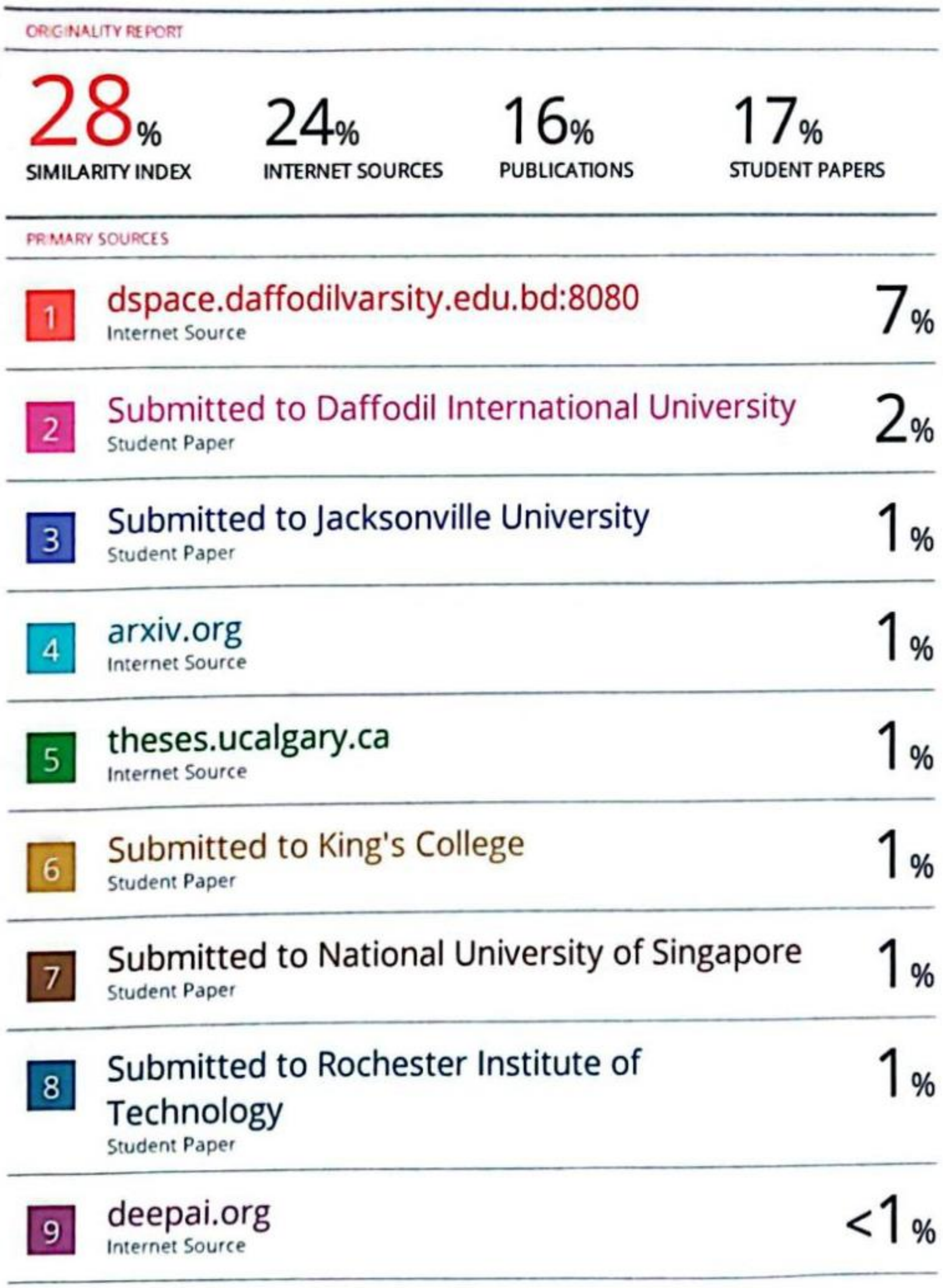
[38] A. Graves and J. Schmidhuber, “Framewise phoneme classification with bidirectional LSTM and other neural network architectures,” *Neural Networks*, vol. 18, no. 5–6, pp. 602–610, Jul. 2005, doi: 10.1016/j.neunet.2005.06.042.

[39] R. Ghaeini *et al.*, “DR-BiLSTM: Dependent Reading Bidirectional LSTM for Natural Language Inference,” *arXiv:1802.05577 [cs]*, Apr. 2018, Accessed: Dec. 29, 2022. [Online]. Available: https://arxiv.org/abs/1802.05577

[40] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez, “SPANISH PRE-TRAINED BERT MODEL AND EVALUATION DATA.” Accessed: Dec. 29, 2022. [Online]. Available: https://users.dcc.uchile.cl/~jperez/papers/pml4dc2020.pdf

[41] D. Krstinić, M. Braović, L. Šerić, and D. Božić-Štulić, “Multi-label Classifier Performance Evaluation with Confusion Matrix,” *Computer Science & Information Technology*, Jun. 2020, doi: 10.5121/csit.2020.100801.

[42] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, “Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach,” *Biometrics*, vol. 44, no. 3, p. 837, Sep. 1988, doi: 10.2307/2531595.



ORIGINALITY REPORT

28%	24%	16%	17%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	7%
2	Submitted to Daffodil International University Student Paper	2%
3	Submitted to Jacksonville University Student Paper	1%
4	arxiv.org Internet Source	1%
5	theses.ucalgary.ca Internet Source	1%
6	Submitted to King's College Student Paper	1%
7	Submitted to National University of Singapore Student Paper	1%
8	Submitted to Rochester Institute of Technology Student Paper	1%
9	deepai.org Internet Source	<1%