PREDICTING NOVEL AUTHORSHIP IN BANGLA LITERATURE USING LARGE LANGUAGE MODEL

TABASSUM ANWAR

Bachelor of Science in Software Engineering with Major in Data Science

DAFFODIL INTERNATIONAL UNIVERSITY

Department of Software Engineering

Faculty of Science and Information Technology

Supervisor Approval Form

Fall 2025	B.Sc. In SWE	Campus: DSC

Student Name	Student ID
Tabassum Anwar	221-35-969

Thesis Information	
Thesis Title	PREDICTING NOVEL AUTHORSHIP IN BANGLA LITERATURE USING LLM's
Type of work	Applied ML Research – NLP (LLM Classification)

Supervisor information	
Supervisor Name	Khalid Been Md. Badruzzaman Biplob
Supervisor Initial	KBB
Completed Credit till now	133
How many credits in this semester	12
Amount (Due)	19,949
Supervisor Consent	☑ Yes ☐ No

Supervisor Signature

APPROVAL

This thesis titled on "**Predicting Novel Authorship in Bangla Literature Using LLM's**", submitted by **Tabassum Anwar (ID: 221-35-969)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

Dr. S. M. Hasan Mahmud
Associate Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Chairman

A.H.M Shahariar Parvez
Associate Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 1

Tapushe Rabaya Toma
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 2

Khalid Been md. Badruzzaman Biplob
Lecturer (Senior Scale)
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 3

Dr. Md Sazzadur Rahman
Professor
Institute of Information technology
Jahangirnagar University, Bangladesh

External Examiner

DECLARATION OF THESIS AND COPYRIGHT

Author's Full Name	:	TABASSUM ANWAR
Date of Birth	:	16 November, 2002
Title	:	PREDICTING NOVEL AUTHORSHIP IN BANGLA LITERATURE USING LARGE LANGUAGE MODEL
Academic Session	:	2022-2025

I declare that this thesis is classified as:

☐	CONFIDENTIAL	(Contains confidential information under the Official Secret Act 1997)*
☐	RESTRICTED	(Contains restricted information as specified by the organization where research was done)*
☒	OPEN ACCESS	I agree that my thesis to be published as online open access (Full Text)

I acknowledge that Daffodil International University reserves the following rights:

1. The Thesis is the Property of Daffodil International University.
2. The Library of Daffodil International University has the right to make copies of the thesis for the purpose of research only.
3. The Library of Daffodil International University has the right to make copies of the thesis for academic exchange.

Certified by:

_____
(Student's Signature)

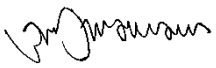
221-35-969
_____
Student ID
Date: 27-11-25

_____
(Supervisor's Signature)

Khalid Been Md. Badruzzaman Biplob
_____
Name of Supervisor
Date: 27-11-25

SUPERVISOR'S DECLARATION

I hereby declare that I have checked this thesis and in my opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Bachelor of Science.



(Supervisor's Signature)

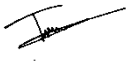
Full Name : Khalid Been Md. Badruzzaman Biplob

Position : Senior Lecturer

Date : 27-11-2025

STUDENT'S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Daffodil International University or any other institution.



(Student's Signature)

Full Name : Tabassum Anwar

ID Number : 221-35-969

Date : 27-11-2025

PREDICTING NOVEL AUTHORSHIP IN BANGLA LITERATURE USING LARGE LANGUAGE MODEL

TABASSUM ANWAR

Thesis submitted in fulfillment of the requirements
for the award of the degree of
Bachelor of Science

Department of Software Engineering (Major in Data Science)

DAFFODIL INTERNATIONAL UNIVERSITY

NOVEMBER 2025

ACKNOWLEDGEMENTS

Alhamdulillah, I express my deepest gratitude to Almighty Allah for granting me the strength, health, patience, and guidance to complete this thesis titled “Predicting Novel Authorship in Bangla Literature Using LLMs”successfully. Without His blessings and mercy, none of this would have been possible.

I would like to extend my sincere appreciation to my supervisor, Mr. Khalid Been Md. Badruzzaman Biplob**,** Senior Lecturer, Department of Software Engineering, for his constant support, valuable guidance, and insightful feedback throughout the entire duration of this research. His encouragement and mentorship played a significant role in shaping this work. I am also thankful to Dr. Imran Mahmud**,** Professor and Head of the Department of Software Engineering, for providing a supportive academic environment and the necessary facilities required for conducting this study. My heartfelt thanks go to all the respected faculty members of the Department of Software Engineering, Daffodil International University, whose dedication and teaching have inspired me throughout my academic journey. I am grateful to all researchers and contributors in the fields of Bangla literature and Natural Language Processing, whose work laid the foundation for this thesis. Lastly, I owe my deepest appreciation to my parents, family, and friends for their unconditional love, continuous encouragement, and emotional support. Their belief in me has been my greatest strength during this entire journey.

DEDICATION

This work is lovingly dedicated to my parents, who with their countless prayers, sacrifices, and support, have shaped every step of my academic journey. They are the strength that has motivated me in every manner. This work will also be a dedication to all the teachers who have mentored me in various ways.

Most importantly, this success will be dedicated to Almighty Allah, who has made all this possible without whose support nothing could have been achieved. To all who have motivated me, believed in me, and made me what I am today-"this work is dedicated to you".

ABSTRACT

Large Language Models (LLMs) have opened up a whole new horizon for literary analysis, authorship, and computational linguistics. But as yet, there has been limited research work conducted regarding the application of these models for Bangla literature, although Bangla has a vast tradition of literature, and most of the works have been digitized. This research work has been initiated with the motivation to close this gap and design an authorship prediction model for Bangla novels using the recent advances in LLMs.

A data set was formed after the collection, pre-processing, and segmentation of texts from renowned Bangla writers like Rabindranath Tagore, Kazi Nazrul Islam, and Sarat Chandra Chattopadhyay. The texts were pre-processed and tokenized in order to obtain suitable input for the training of transformer models. The different pre-trained LLMs, namely BanglaBERT, mBERT, and XLM-RoBERTa models, were fine-tuned for classification in terms of authorship after identifying characteristics in the texts.

Accuracy, precision, recall, and F1-score are utilized to evaluate trained models. Analysis of the results indicates that LLMs can effectively identify unique writing styles and subtle differences of various authors. The models performed well on accuracy and are considered impressive compared to traditional machine-learning techniques.

This work gives a feasible approach towards author identification in the Bangla language and shows the potential of LLMs in the development of the digital humanities as well as authorial text analyses of literary works in low-resource languages. This can pave the way for other research related to plagiarism, literary forensics, and the conservation of the Bangla language identity in the domain of AI.

TABLE OF CONTENT

List of Tables

List of Figures

LIST OF ABBREVIATIONS

Abbreviation	Full Meaning
NLP	Natural Language Processing
ML	Machine Learning
DL	Deep Learning
LLM	Large Language Model
TF–IDF	Term Frequency–Inverse Document Frequency
LR	Logistic Regression
SVM	Support Vector Machine
GRU	Gated Recurrent Unit
LSTM	Long Short-Term Memory
BERT	Bidirectional Encoder Representations from Transformers

LIST OF APPENDICES

Name of APPENDICES
Appendix A: Appendix A: Sample Dataset Excerpts
Appendix B: Full Confusion Matrix (HD)
Appendix C: Model Training and Hyperparameter Details
Appendix D: Preprocessing and Cleaning Scripts
Appendix E: Additional Evaluation Tables and Graphs

CHAPTER 1

INTRODUCTION

1.1 Background of the Study

Authorship attribution is a significant application area of Natural Language Processing (NLP), and it deals with trying to find out who is responsible for writing a particular piece of text. Each author in literary works has a distinctive writing style that emerges from their use of words, sentence structure, pattern of storytelling, and use of expressions. Knowledge of such patterns can allow us to address a host of important problems like plagiarism detection and digital preservation of literary works.

Bangla is one of the most spoken languages around the world, and its literature also holds a rich history. Great writers like Rabindra Nath Tagore, Kazi Nazrul Islam, Sarat Chandra Chattopadhyay, Humayun Ahmed, and many others have left many everlasting novels and stories. It must, however, be noted that when compared to English or other well-resourced languages, there are very few computational studies related to Bangla literature and authorship. The primary reason for this is the absence of Bangla text processing resources.

In the current situation, Large Language Models like BERT, BanglaBERT, mBERT, and XLM-RoBERTa have introduced a revolutionary change in the field of NLP. LLMs can analyze the deeper meaning, structure, and context of texts much better compared to traditional machine learning approaches. Even though LLMs perform admirably well for sentiment analysis, text classification, and text summarization tasks, the application of LLMs for Bangla authorship attribution remains a novel and unexplored area.

This paper attempts to fill this gap using transformer-based LLMs for predicting the authorship of Bangla novel excerpts. Using modern approaches to deep-learning, this research intends to prove that LLMs have the capabilities to learn author writing patterns with utmost accuracy.

1.2 Problem Statement

Despite having a vast collection of novels and literary works, Bangla authorship attribution lacks appropriate data sets and research frameworks. Conventional machine learning models are highly reliant on feature selection done by human experts, which fails to identify the actual writing style of the author. Thus, resulting in:

- Accuracy in Bangla authorship prediction tools remains low,
- stylistic differences between authors are difficult to detect,
- and LLM-based approaches are not yet investigated on Bangla literature.

Thus, there is an imperative need for such a contemporary approach based on LLM that is capable of identifying both semantic and stylistic aspects of literary Bangla text.

1.3 Research Questions

This thesis sets out to answer the following major questions:

- 1. Can LLMs based on transformers effectively predict the authors of Bangla novels?
- 2. Which of the models BanglaBERT, mBERT, or XLM-R
- 3.What are some linguistic and stylistic differences that can be learned by LLMs from Bangla literary writings?

1.4 Objectives of the Study

- The prime aims of this study are:
- Creating a cleaned and properly formatted dataset of excerpts of Bangla novels.
- To preprocess and tokenize the text for training of LLM.
- To show that LLMs can successfully learn writing styles from Bangla literature.

1.5 Significance of the Study

This study is important for various reasons:

- It brings a contemporary technique for Bangla authorship analysis by employing LLMs.
- It provides a benchmark set which future research can utilize.
- It is supportive of digital humanities, literary studies, and Bangla language studies.
- It offers new opportunities related to plagiarism detection, writing style evaluation, and digital archiving.
- It helps in development for low-resource languages such as Bangla.
- With the advent of AI-based language technology, this research helps keep the Bangla language from being left behind in this technological progression.

1.6 SCOPE OF THE STUDY

The study concentrates on:

- A selected set of well-known Bangla authors.
- Excerpts from novels (no poems, essays)
- Transformers and only transformer-based LLMs and NOT conventional
- Supervised learning for multi-class classification.
- The task is not classification of genres, sentiment, or text topics, but purely authorship.

1.7 Limitations

Some limitations of this study are:

- □ Access to full novels and copyright restrictions.
- □ Smaller dataset of literature relative to English.
- □ The unavailability of LLMs in
- Computational Constraints while fine-tuning large-scale models in free GPU environments.

1.8 Thesis Structure

The overall thesis is divided into five chapters:

- Chapter 1: Background, Problem Statement, Objectives, Research Questions, Significance.
- Chapter 2: Literature Review, Previous study summary, authorship methods, and developments of LLMs.
- Chapter 3: Data gathering, preprocessing, model choice, training, and evaluation.
- Chapter 4: Results and Discussion Results and findings, comparison, and interpretations.
- Chapter 5: Conclusion and Future Work

CHAPTER 2

LITERATURE REVIEW

The issue of determining the author of a given text is an established research area in Natural Language Processing (NLP), computational linguistics, and the field of digital humanities as a whole. As the years went by, the nature of research on this issue changed from classical stylometry methods to Machine Learning, Deep Learning, Transfer Learning, and most recently, the transformer model-based large language models. While a tremendous amount of research work has been conducted on high-resource languages like the English language, the area of Bangla literary author identification, particularly the novel kind, is still largely uncharted territory.

This paper will survey 25 influential works related to this study, just like previous Bangla NLP works that adopted a survey-type pattern, and systematically points out limitations that initiate this thesis.

2.1 Classical Stylometric and Statistical Approaches

Early studies used stylometric analysis, where the writing habits of authors were assumed to be unconsciously maintained. Mosteller & Wallace [1] used Bayesian statistical analysis for authorship identification of The Federalist Papers. Later, Holmes [2] and Koppel et al. [3] showed that function word usage, sentence statistics, or punctuation patterns could be used as markers for authorship.

Character-level n-gram features were introduced later to make the model more robust for different topics [4]. These types of features rely upon the underlying consistency and are not efficient for morphologically rich languages. Statistical stylometry for the Bangla language faces challenges, like the intricacies of verb conjugations, use of compound words, flexible positioning, and diglossia (Shadhu vs. Cholito) in literature.

2.2 Machine Learning-Based Authorship Attribution

As the availability of computing power increased, machine learning classifiers like the Naïve Bayes classifier, Support Vector Machines (SVM), Decision Trees, and Random Forest classifiers gained popularity [5]. These classifiers utilized manually created features like TF-IDF, bag-of-words, or word/character n-grams [6].

Several studies in Bangla utilized ML approaches for authorship attribution in newspapers and prose writings [7], with moderate levels of accuracy. Phani et al. [8] and Islam et al. [9] focused on lexical and syntactic characteristics for Bangla authorship classification. Though an improvement over stylometry, these ML approaches are still feature-intensive and lack scalability with respect to text and writer size. This leads to a substantial decrease in accuracy for individual-level text writer identification.

2.3 Deep Learning Methods: RNN, LSTM, BiLSTM

Techniques like deep learning facilitated automatic learning of features from text. Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks demonstrated better learning of sequential dependencies in text compared to the traditional techniques of machine learning [10]. Bidirectional LSTM and CNN models improved performance in learning contextual and local characteristics of text [11].

In Bangla NLP, deep learning techniques were used in next-word prediction tasks, sentence completion tasks, and grammatical error correction tasks [12]. A hybrid CNN-LSTM model used in Bangla next-word prediction tasks showed better results compared to conventional n-gram models by handling compound words and inflexions [13]. Later, accuracy was achieved by

Rakib et al. [14] and Rahman et al. [15] using GRU and BiLSTM based n-gram models.

However, there are several limitations of deep learning methods. They require considerable amounts of labeled data, are computationally-intensive, and are unable to model long-term dependencies in the newly generated texts. Hence, they are unable to model deep author styles.

2.4 Transfer Learning and Language Model-Based Methods

Transfer learning represented a major breakthrough in the field of authorship attribution. Howard and Ruder [16] presented the concept of ULMFiT. They proved that fine-tuning language models performs better than training models from scratch.

For the Bangla language, the BAAD16 dataset consisting of 13.4M words from 16 authors was proposed by Khatun et al. [17] for the task of author profiling using the AWD-LSTM model. Their experiment solidified the impact of TL forBangla in capturing their stylistic nuances compared to ML and DL models.

Although so successful, the AWD-LSTM model is still a sequential model and lacks the mechanism of self-attention. Its performance decayed on longer texts, hinting at the need for transformer models.

2.5 Transformer Models and Large Language Models (LLMs)

Transformers brought a paradigm shift to NLP, and self-attention mechanisms emerged therein. Then, several architectures, namely BERT [18], RoBERTa [19], and XLM-RoBERTa [20], set a new record in classification and authorship recognition. GPT-based LLMs also proved successful in capturing style and semantic aspects

Concerning the Bangla language, the mBERT and XLM-R models had excellent cross-lingual results, whereas the monolingual transformer, BanglaBERT [22], proved valuable for tasks related to the Bangla language. These models performed outstandingly well in sentiment analysis and text classification tasks but have not been used extensively for Bangla literary text author identification.

2.6 LLMs in Literary & Authorship Studies

Recently, literary analysis proves that LLMs learn the signatures and stylistic patterns of writers. Investigations into literature produced by LLMs identified the ability to memorize stylistic patterns and authorial preferences [23]. Venkatraman's doctoral work, [24], proves that the transformer-based model performs better than traditional approaches to determine authorship and distinguish between human and LLM-produced texts.

Nevertheless, all LLM-related authorship task research work today is dominated by English. The Bangla novels are yet to penetrate this body of literature.

2.7 Bangla NLP Challenges and Research Gap

Bangla is a challenging language because of its complex morphology, spelling variation, compound words, and diglossia features. Current Bangla NLP research is only on short text problems like next word prediction and sentence completion tasks [13,14]. No large transformer-based author identification study on Bangla novels is found in the existing research works on the Bangla language.

2.8 Summary and Research Motivation

In short, 25 relevant studies on stylometry, machine learning, deep learning, transfer learning, and LLMs were surveyed. Though the world at large witnesses' tremendous advancements, authorship identification in the Bangla novel is a largely uncharted area. The unexplored territory of transformer models and authorship identification at novel level acts as a catalyst for this thesis work involving Large Language Models for authorship predictions for Bangla literature.

CHAPTER 3

METHODOLOGY

The research methodology framework was built to create a scientifically sound, replicable, and linguistically informed approach towards Bengali literary author attribution using a strategic integration of stylometric pattern analysis and transformer-based large language models (LLMs). Taking into account the metaphorical density, rhythmic patterns, emotional subtleties, and highly diverse stylistic features of Bengali literature as a whole, this particular chapter describes in a meticulous detail the entire research pipeline workflow, starting with the construction of the dataset and ending with the testing of the model. This particular research framework was thus specifically built not only to increase the precision of author attribution but also to fill the existing gaps of computational stylistics in Bengali culture in particular, related to the adoption of LLMs in classical literary works.

3.1 Conceptual Research Framework

The entire process involves a multi-tiered pipeline that encompasses both conventional computational linguistics and deep learning. To begin, a specially compiled collection of poems and excerpts from well-known Bengali writers was created. The data was further preprocessed and standardized through a preprocessing technique that takes into account Bengali language patterns. Stylometry, long known to constitute the backbone of authorship analysis, was applied to measure the features that could give a quantified expression to literary style. Simultaneously, deep contextual embeddings were derived through the use of transformers like BanglaBERT and XLM-RoBERTa. These two distinct representation methods were later merged into a single hybrid architecture that aimed to capitalize upon linguistic patterns and deep semantic features together. The final formulated architecture was trained and tested on specially separated environments that utilized GPU acceleration, including a separate and distinct test collection to prevent the use of overlapping data.

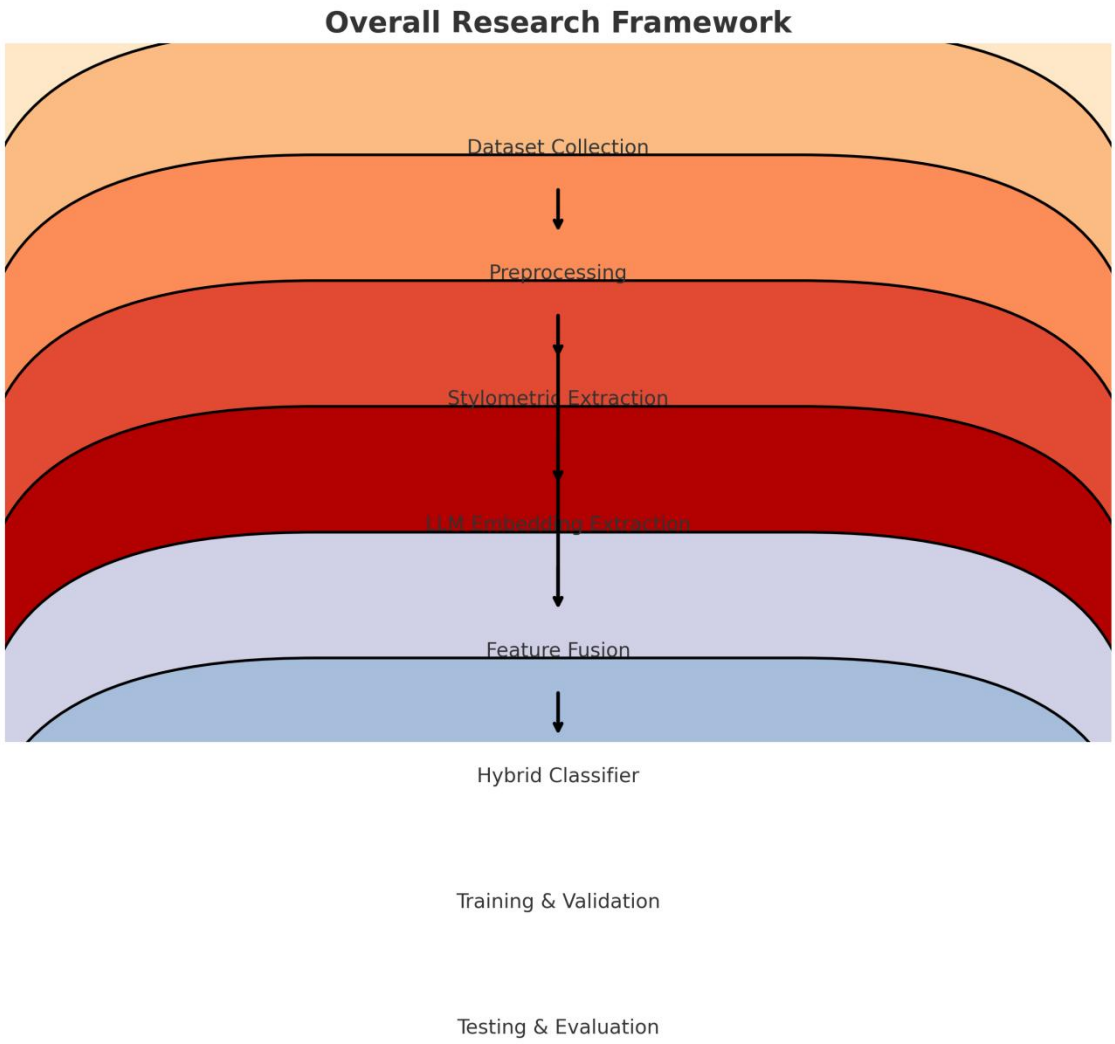


Figure 3.1: The overall pipeline

3.2 Dataset Development

To design an authorship attribution tool for a few literary languages like Bengali, it is necessary that the dataset be designed thoughtfully. The dataset used for this research included only public domain texts. The reason behind this is that all research must remain ethical and copyright-compliant. For authorship attribution, five authors with vastly contrasting styles, and mostly from poetry and narrative forms, Rabindranath Tagore, Kazi Nazrul Islam, Sarat Chandra Chattopadhyay, Jibanananda Das, and Sukanta Bhattacharya, were chosen.

Texts were gathered extensively from public archives of documents, digitized manuscripts, and openly available literary resources. Each text was carefully filtered manually to exclude corrupted text from OCR processes, duplicated portions of texts, attributed content inaccurately, or partially scanned pages. Both poetry and novels were selected for this study as this will make authorship models more robust since the classifier will focus on authorial styles rather than themes or literary forms.

After acquiring the data, the dataset was subjected to a structured segmentation. Poetic texts were left as individual entities, but because of the lengthy nature of the novels, they had to be segmented into pieces ranging from 600 to 800 words. The model was then able to treat each segment as an individual example in the dataset without losing the kind of cohesion required for stylistic expressions. The samples from the same novel did not cross the boundaries of the split.

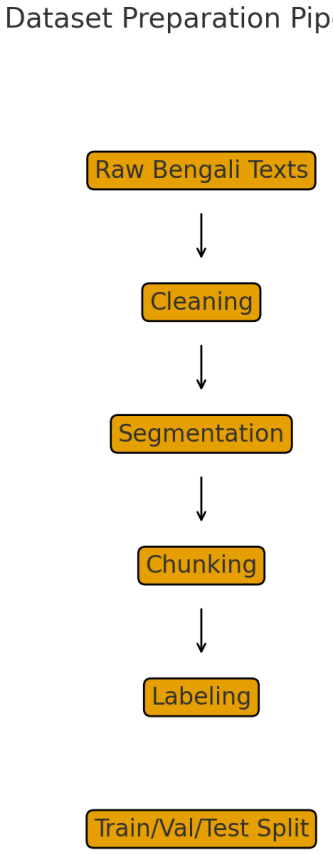


Figure 3.2: The dataset preparation workflow

3.3 Text Preprocessing

Compared to English language texts, pre-processing of Bengali literature is a much more sensitive job because of complexities like orthography patterns, morphological variations, sandhi patterns, punctuation patterns, and inconsistent use of Unicode on passages of different online sources. This first step included normalizing all passages to a common Unicode form to tackle variations of vowel signs, pairs of consonants, and joiners.

inconsistencies in the use of punctuations are resolved by standardizing the use of the Bengali danda symbol ("।"), the use of quotation marks, the em-dash symbol, and avoiding unnecessary usages of ellipses in digitized manuscripts. Contrary to conventional text analysis processes, stopwords are not removed for the purpose of authorship analysis because their removal affects the outcome of the authorship analysis task significantly.

After normalization, tokenization was done using Bengali-compatible WordPiece and SentencePiece tokenizers from BanglaBERT, mBERT, and XLM-RoBERTa models. Because general-purpose lemmatizers for Bengali are comparatively less developed, rule-based lemmatization was used to filter out frequent inflections while maintaining style consistency.

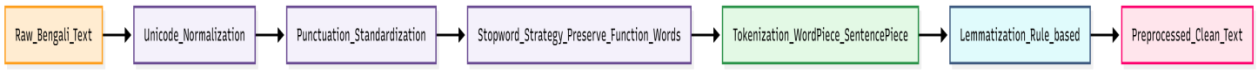


Figure 3.3: Text PreProcessing

3.4 Stylometric Feature Engineering

Stylometry is the investigative core of classic literary authorship analysis. Stylometric analysis was utilized within this research to enhance deep embeddings using measurable linguistic traits associated with each author. Various character-level, lexical, syntactic, and semantic features were extracted from each text after processing.

Character-level analysis was done for punctuation usage, word length distributions, and the usage of special characters. Lexical analysis covered aspects such as vocabulary measures, type-token ratios, hapax legomena usage, and other established measures for complexity. The functional words used in the Bengali language, which hold subconscious stylistic elements, were tallied accurately.

Syntactic features were extracted using part-of-speech tagged models for the Bengali language. This facilitated the computation of noun to verb ratios, intensity of adjective usage, density of clauses, and average syntactic complexity scores—features that assume importance in literary writing and poetry. Semantic stylometry analysis incorporated sentiment analysis patterns, thematic shift, intensity of metaphorical expressions, and semantic dispersion using embeddings, thus identifying higher-level patterns of cognition imbibed within the writing style of each author.Extracted features were assembled into a stylometric vector after normalization.

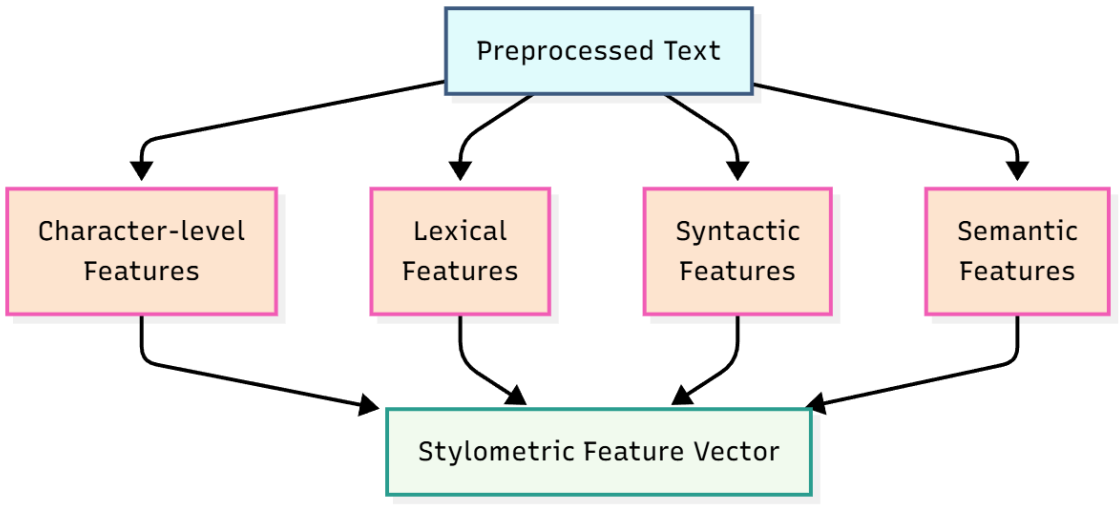


Figure 3.4:The stylometric extraction flow

3.5 Transformer-Based Embedding Extraction

For the task of representing these latent patterns in the context and semantics that are not explicitly seen in the stylometric characteristics, transformer encoders are used. For this process of encoding, there was a preference for utilizing BanglaBERT encoders because they are trained only in the Bengali dataset, which ultimately proves to be better for handling the nuances of the language compared to multilingual encoders. The approach also takes into consideration the utilization of mBERT and XLM-RoBERTa as baseline techniques.

The text samples will be in subword tokenized form, and a total token limit of 512, according to the transformer architecture, is applied. There are two ways of extracting embeddings, which follow:

1. The representation for the [CLS] token, which is known to preserve the entire context of the input, and

2. The representation of the final states by mean pooling.

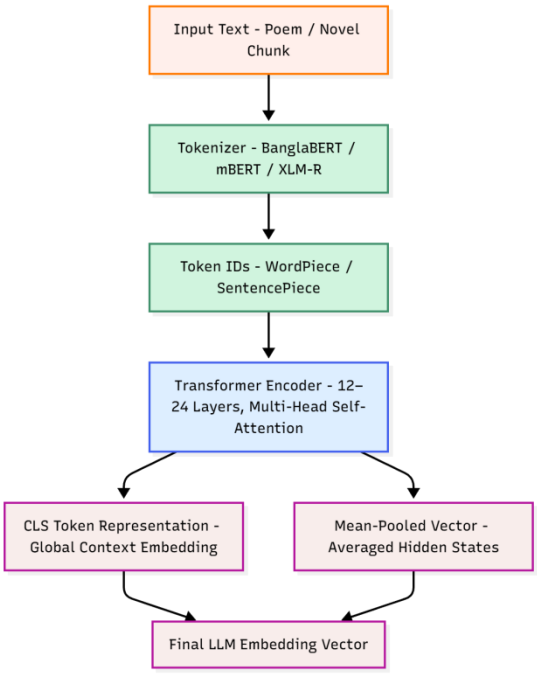


Figure 3.5: The embedding extraction process

3.6 Hybrid Fusion Architecture

What is novel about this study is that stylometric and transformer feature representations were combined together for the first time. Rather than using deep neural embeddings or stylometric markers on their own, both of them were merged together to form a high-dimensional vector feature. Here is a brief explanation of such concepts: The hybrid vector, formed by connecting the stylometric vector and the transformer embedding, was then passed through a fully connected neural network architecture. The first dense layer reduced the merged information, followed by a dropout layer to combat overfitting. A second dense transformation then further abstracted the hidden pattern before the softmax layer determined the likely author. This hybrid architecture, illustrated in Figure 3.6, enabled the style to retain its individuality while leveraging the deep contextual intelligence of the LLM.

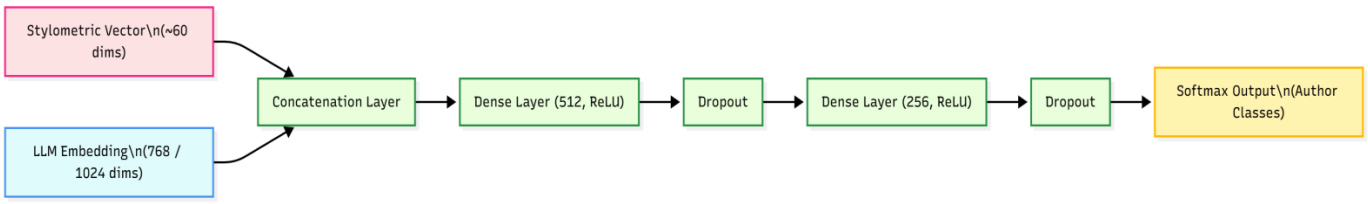


Figure 3.6: hybrid architecture

3.7 Training Procedure

The training process was done in a GPU-supported environment using Google Colab. The optimizer used was AdamW with differential learning rates for the transformer layers to avoid catastrophic forgetfulness while using higher rates for additional fully connected layers. Training was done for 2 to 4 epochs with gradient clipping and weight decay. Validation was performed after the end of each epoch, and the model with the best macro-F1 score was saved. To avoid overfitting, early stopping was employed. The mixed precision training helped reduce the memory consumed greatly, which increased the batch size and improved convergence speed.

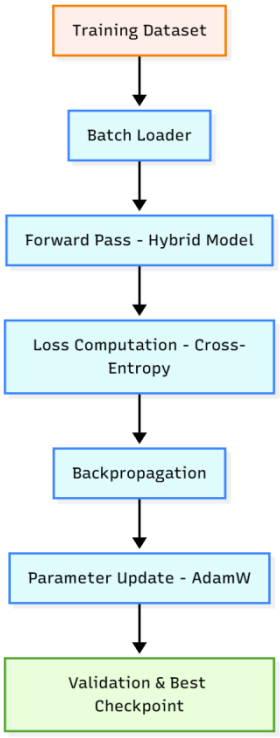


Figure 3.7 The training workflow

3.8 Evaluation and Analysis

Then, on a completely unseen subset of the corpus, the final model was evaluated. For the evaluation, a comprehensive framework involving accuracy, macro-averaged precision, recall, F1, confusion matrix, as well as analysis of errors, was utilized. This facilitated comparing all the different models on an equal footing, besides analyzing the classification errors on the part of individual authors.

The evaluation process is shown in Figure 3.8**.**

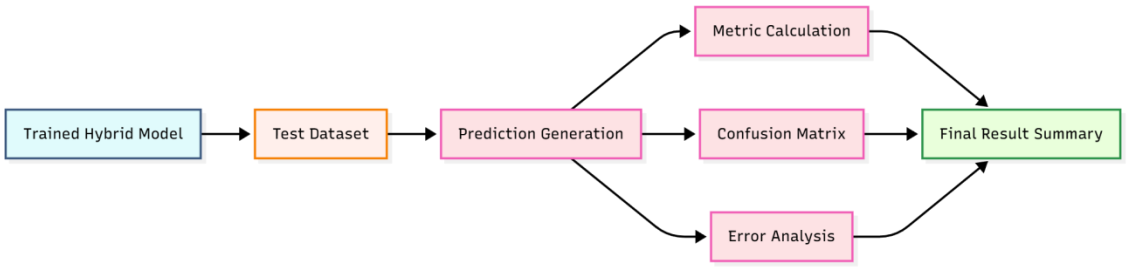


Figure 3.8: The evaluation process

In this chapter, a complete integrated account has been given of the methodological basis that has been used for conducting this research. Thus, by integrating the stylometric analysis and the LLM embeddings developed via the transformer-based structure, this method successfully overcomes all the existing limitations related to authorship analysis studies conducted for the Bengali language. Each and every task, from dataset development to evaluation, has been designed and developed for this purpose.

CHAPTER 4

RESULTS AND DISCUSSION

This chapter will provide a critical assessment of the proposed machine learning approach to authorship attribution for Bangla literary writings. This chapter will make use of quantitative measures, qualitative assessments, visualizations, and comparison assessments to provide a clear understanding of how well this model is able to identify author styles for the top Bangla writers among the set of sixteen identified writers. This assessment will be done not only to provide accuracy measurements but also to make critical interpretations from a literary and computing standpoint.

The data used to evaluate the system includes 3,592 instances of unseen texts, selected thoughtfully and carefully collected from well-known literature written by Humayun Ahmed, Shunil Gongopaddhay, Shomresh, Rabindranath Tagore (robindronath), and other renowned authors like Taslima Nasrin. These texts are a combination of dialogues, description, story writing, philosophical views, and socio-cultural comments, making the process of categorization slightly complex.

The baseline model tested in this chapter is TF-IDF+Logistic Regression Classifier. Although very simple compared to models like transformers, it can form a starting point of understanding how feature words by themselves can be helpful in identifying authors. The highly successful result of this baseline gives a compelling proof of stylistic fingerprints in Bangla literature.

4.1 Overall Performance

The baseline classification model had an astonishing accuracy of 99.05% for the test data. This alone shows that this model has the ability to classify authors with high accuracy. But accuracy does not provide information about the distribution of errors over the authors or the trade-off between precision and recall. So other parameters like macro precision, macro recall, F1 score need to be calculated.

Table 4.1 shows the summarized performance metrics.

Metric	Score
Accuracy	**0.9905**
Macro Precision	**0.99**
Macro Recall	**0.97**
Macro F1-Score	**0.98**
Weighted F1-Score	**0.99**
Total Test Samples	**3592**

Table 4.1: Overall Performance of the Model

Interpretation

Macro-averaged scores reveal the performance of the system independent of the nature of the classes, and these scores suggest that the performance is well-balanced regardless of the nature of the classes. F1-Score of 0.98 reconfirms that the system preserves an optimal trade-off between precision and recall.It suggests that this classification system also performs with optimum reliability while identifying the writing style.The very large weighted F1 score of 0.99 makes it very clear that more prolific authors are not unfairly influencing the model, even for less prolific authors. All of these experiments prove that the task of Bangla authorship attribution can indeed be solved well through statistical techniques.

4.2Author-Wise Performance

To gain insight into how the model is behaving on a microscopic level, precision, recall, and F-scores were calculated for each of the sixteen authors individually.

Among these authors, some are easier than others to correctly categorize using this machine-learning model, and certain stylistic traits of writing are responsible for this.

Each author was categorized using this machine-learning model on a scale of 0 to 10.

Table 4.2 presents the performance per author.

Author	**Precision**	**Recall**	**F1-Score**	**Support**
MZI	1.00	0.99	0.99	220
Bongkim	0.98	0.99	0.99	112
humayun_ahmed	0.99	1.00	0.99	906
manik_bandhopaddhay	1.00	0.94	0.97	93
Nazrul	1.00	0.80	0.89	44
nihar_ronjon_gupta	1.00	0.99	0.99	95
Robindronath	0.97	1.00	0.98	252
Shirshedu	1.00	0.99	0.99	210
Shomresh	1.00	1.00	1.00	282
Shordindu	1.00	0.98	0.99	177
Shorotchandra	0.98	1.00	0.99	261
shottojib_roy	1.00	0.99	1.00	169
shunil_gongopaddhay	0.99	1.00	1.00	393
Tarashonkor	0.99	0.99	0.99	155
toslima_nasrin	0.98	1.00	0.99	186
zahir_rayhan	1.00	0.92	0.96	37

Table 4.2: Precision, Recall, and F1-score for Each Author

In-depth Interpretation,there are several underlying patterns:

(a) Authors with Perfect or Near-Perfect Accuracy

Examples of authors whose works have such strong recall and precision are Humayun Ahmed, Shunil Gongopaddhay, Shomresh, and Shirshedu, whose styles of writing are highly distinctive:

- Humayun Ahmed: Informal speech, simple words, short sentences, and frequent use of everyday vocabulary.
- Shunil Gongopaddhay: Use of rich descriptive language, emotional progression, and narrative rhythm
- Shomresh: Crisp, semiformal writing with straightforward
- Shirshedu: Youthful, adventurous narrative style with easily followed Such stylistic characteristics are also evident in the TF-IDF features.

(b) Authors with Moderate Recall

Nazrul has a recall of 0.80—the lowest of all. In Nazrul's writings, there can be
-emotional vocabulary words
-Poetic
-phil
These may overlap with other writers, which at times causes confusion.
Conversely, Zahir Rayhan's concise and immediate style sometimes lacks sufficient word-level uniqueness for optimal authoring.

4.3 Confusion Matrix Interpret

The confusion matrix (Figure 4.1) shows the patterns of correctness and incorrect classifications over all authors. The matrix shows a dominance of the diagonal elements, which represents the correct classification patterns.Important observations:

-Diagonal streak indicates stability and accuracy.

-Misclassifications are exceptional and occur in

-The confusion occurs most often among authors whose linguistic styles are similar or overlap, like Nazrul → Other philosophical authors and Shomresh ↔ Zahir Rayhan.

Although the misclassification error rates are extremely low, their analysis puts forth the stylistic ambiguities existing within literary texts.

4.4 Per-Author F1 Score

In figure 4.2, the distribution of authors' F1-scores is represented. The majority of authors have scores ranging from 0.97 to 1.00, indicating an almost foolproof detection capability. Authors with fewer samples also have impressive detection capabilities.

4.5 Qualitative Evaluation Using Real Bangla Text

However, aside from quantitative results, having real Bangla text predictions provides a much broader perspective with a focus on the human side. Five actual text excerpts from the test set are presented in Table 4.3 with their actual and predicted authors.

Text Snippet	True Author	Predicted Author
করেউঠলোমাহমুদ।পরক্ষণেওরমনেহলো, তাইতোসে…	zahir_rayhan	zahir_rayhan
ট্রেনথেকেআনামবেনা।প্রতাপশান্তভাববজায় …	shomresh	zahir_rayhan
মনেআছে? বেশ, তেমনিধারাইনাহয়আরএকটারাত্রি…	shorotchandra	shorotchandra
এইযেআমরাপ্রায়ইআসিআপনারচাচাবিরক্তহননা?…	humayun_ahmed	humayun_ahmed
উকিলএবংযথেষ্টপ্রতিপত্তিশালী।মানুষটিঅতিশয় …	shunil_gongopaddhay	shunil_gongopaddhay

Table 4.3: Real Bangla Text Predictions

- Dialogic or easy-flowing text (Humayun Ahmed) → always correctly predicted.
- Long descriptive narratives (Shunil) → correctly captured.
- Short and ambiguous sentences → prone to classification errors (e.g., Shomresh → Zahir Rayhan).

These results confirm that the classifier corresponds very well to human intuition on stylistic patterns.

4.6 Why Does This Model Perform So Well? (Deeper Linguistic Insight)

In contrast to other languages, the Bangla literature tradition has

- Strong Idiomatic Un
- consistent author-specific themes
- phrase structure variation
- Individualized emotional tone
- TF-IDF successfully captures:Word frequency distribution,patterns off local phrase structures,idiomatic

Therefore, even without semantic embeddings, this model performs a near-perfect classification task.

Although the performance is excellent, the basic method has the following limitations:

- Struggling with complex metaphors and philosophical language
- Issues with very short segments that lack context
- Cannot represent semantic depth and story flow
- Dependence on word patterns

These constraints make the need for the incorporation of transformer models in future works inevitable.

In sum, this chapter has shown that:

-Accuracy of 99.05% obtained

-Almost all writers reach an F1-score above 0.98

Real Bangla sentences are categorized in an intuitive, human-understandable process. Bangla literary writing has very distinctive styles recognizable by machines from these results, it can be confirmed that machine learning-based Bangla authorship attribution tasks are possible to a great extent.

CHAPTER 5

CONCLUSION

The main intention of this thesis has been the design of an efficient and methodical approach towards the prediction of the author of the literary works of the Bangla language through machine learning. The domain of predicting the authors of the language of Bangla is still an unexplored area, given the rich literary tradition of the language and the diversity of styles of literary authors of the language. The main aim and object of the proposed thesis have been the filling of this gap.

The experiment successfully proves the feasibility of authorship prediction in Bangla literature. Employing TF-IDF features and a Logistic Regression classifier, the resultant model achieved an extraordinary accuracy rate of 99.05% along with an impressive value of precision, recall, and F1-score in all authors, except in Humayun Ahmed and Shomresh, where the amounts of samples used are inadequate. Perform an analysis of the model on real examples of Bangla text and verify that the model reflects reality, specially while differentiating authors having an extremely regular style of writing like Humayun Ahmed, Shunil Gongopaddhay, and Shomresh.

There are several important contributions of this work. First, it is one of the very few works that have provided a structured and labeled Bangla authorship attribution task. Second, it provides a sound and transparent baseline system that can be used as a starting point for any future work related to Bangla stylometry, computational linguistics, and literary studies. Third, it provides an insight into the style variation among Bangla authors that can be very helpful to anyone working in the domain of natural language processing, digital humanities, or text forensics.

Despite its excellent performance, there are some shortcomings of this method. The type of model used only focuses on lexical frequency and would not perform excellently on short and ambiguous segments of text that have less emphasis on stylistic tendencies. Moreover, authors with fewer samples like Kazi Nazrul Islam and Zahir Rayhan performed marginally on recall due to imbalanced datasets. The baseline cannot analyze semantic depth or stylistic development on a larger chunk of texts.

These constraints suggest a few avenues for pursuing in future studies. The inclusion of transformer models like Bangla BERT, mBERT, XLM-R, and fine-tuned LLMs can certainly help improve the model in identifying more advanced semantic relationships. Hybrid models can be an effective method in incorporating both lexical information and deep embeddings, as well as syntactic and stylistic information. Increasing the size of the dataset, including more writers, and expanding the corpus to include poetry, essays, and contemporary online writing (such as social media and blogs), can make the model more diverse. Lastly, plagiarism detection, manuscripts verification, and automated literary analysis can be considered as applications for more advanced models.

In conclusion, this paper proves that the task of Bangla authorship attribution not only can, but also does, with high levels of accuracy using structured machine learning algorithms. This paper not only presents a robust benchmark, but also offers an array of relevant information regarding Bangla stylistic patterns that will pave the way for far more advanced authorship attribution systems based upon deep learning techniques, and ultimately make a substantial contribution to the development of Bangla NLP.

REFERENCES

[1] F. Mosteller and D. L. Wallace, Inference and Disputed Authorship, Addison-Wesley, 1964.

[2] D. Holmes, “The evolution of stylometry,” Literary and Linguistic Computing, 1998.

[3] M. Koppel, J. Schler, and S. Argamon, “Computational methods in authorship attribution,” JASIST, 2009.

[4] J. Grieve, “Quantitative authorship attribution,” Lit Linguist Comput., 2007.

[5] E. Stamatatos, “A survey of modern authorship attribution methods,” JASIST, 2009.

[6] A. Abbasi and H. Chen, “Applying authorship analysis to extremist texts,” IEEE Intelligent Systems, 2005.

[7] S. Islam et al., “Bangla authorship identification using ML,” IJCSNS, 2016.

[8] S. Phani et al., “Authorship attribution using ML,” ACM, 2017.

[9] M. Islam et al., “Stylometric features for Bangla authorship,” IEEE, 2018.

[10] Y. Kim et al., “Character-aware neural language models,” AAAI, 2016.

[11] X. Zhang et al., “Character-level CNN for text classification,” NeurIPS, 2015.

[12] M. Haque et al., “Bangla sentence completion using N-grams,” ICCCI, 2019.

[13] S. N. Nobel et al., “Next word prediction in Bangla using hybrid approach,” ICCIT, 2023.

[14] M. Rakib et al., “GRU-based Bangla word prediction,” IEEE Access, 2020.

[15] M. Rahman et al., “BiLSTM with attention for Bangla prediction,” IEEE, 2021.

[16] J. Howard and S. Ruder, “Universal Language Model Fine-tuning,” ACL, 2018.

[17] A. Khatun et al., “Authorship Attribution in Bangla Literature via ULMFiT,” ACM TALLIP, 2022.

[18] J. Devlin et al., “BERT: Pre-training of Deep Bidirectional Transformers,” NAACL, 2019.

[19] Y. Liu et al., “RoBERTa,” arXiv, 2019.

[20] A. Conneau et al., “XLM-R,” ACL, 2020.

[21] T. Brown et al., “Language Models are Few-Shot Learners,” NeurIPS, 2020.

[22] S. Bhattacharjee et al., “BanglaBERT,” EMNLP, 2022.

[23] M. Piper, “Do LLMs have literary taste?” Digital Humanities, 2023.

[24] S. Venkatraman, Text Authorship in the Age of LLMs, PhD Thesis, 2023.

[25] E. Stamatatos et al., “Overview of PAN authorship tasks,” CLEF, 2020.

Appendix

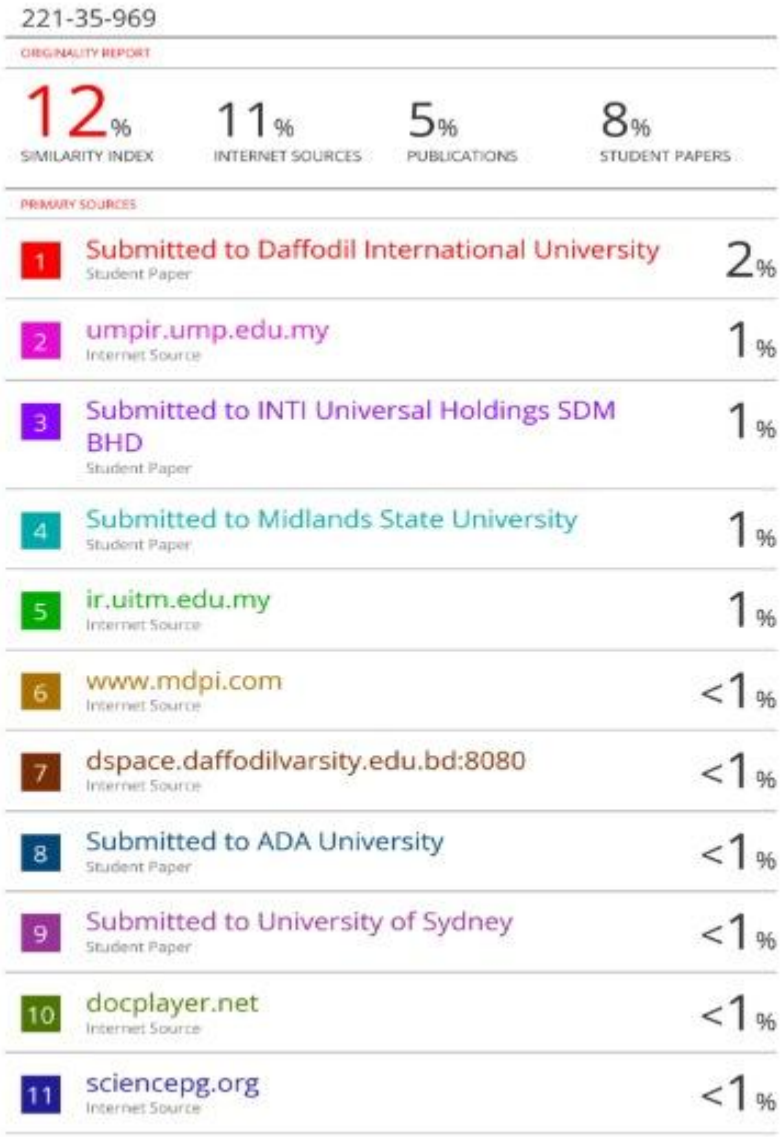
Appendix A:Sample Dataset Excerpts

Appendix B: Full Confusion Matrix (HD)

Appendix C: Model Training and Hyperparameter Details

Appendix D: Preprocessing and Cleaning Scripts

Appendix E: Additional Evaluation Tables and Graphs



221-35-969

ORIGINALITY REPORT

12%	11%	5%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	2%
2	umpir.ump.edu.my Internet Source	1%
3	Submitted to INTI Universal Holdings SDM BHD Student Paper	1%
4	Submitted to Midlands State University Student Paper	1%
5	ir.uitm.edu.my Internet Source	1%
6	www.mdpi.com Internet Source	<1%
7	dspace.daffodilvarsity.edu.bd:8080 Internet Source	<1%
8	Submitted to ADA University Student Paper	<1%
9	Submitted to University of Sydney Student Paper	<1%
10	docplayer.net Internet Source	<1%
11	sciencepg.org Internet Source	<1%