International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

# Multi-Model Analysis on Author Attribution Detection on Assamese Text

Ms. Smriti Priya Medhi[*1], Prof. Shikhar Kr. Sarma[2]

**Submitted**:13/03/2024 **Revised**: 28/04/2024 **Accepted**: 05/05/2024

**Abstract:** Author attribution detection is a crucial task in the field of forensic linguistics and computational stylometry, aiming to identify the author of a given text based on linguistic features. This study focuses on the application of multi-model analysis for author attribution detection specifically in the context of Assamese text, which is a less explored area compared to other languages. The proposed approach is a first ever attempt for Assamese language, and involves the integration of multiple traditional machine learning models, like Support Vector Machines (SVM), Multinomial Naïve Bayes (MNB) etc. These models are trained on a dataset consisting of a diverse collection of Assamese texts authored by different individual authors. A structured and sizable dataset has been created as part of the current work. Key linguistic features, including word n-grams, character n-grams, and part-of-speech tags, are extracted from the text to represent the writing styles of each author. These features are then used as inputs to the multi-model framework, which combine the predictions of individual models to make a final author attribution decision. Experimental results demonstrate the effectiveness of the proposed multi-model approach in author attribution detection on Assamese text. The study contributes to the Assamese Natural Language Processing, by adding a novel work on authorship detection for these low resources and underrepresented language- Assamese, and highlights the importance of using multiple models for improved performance in computational stylometric analysis.

***Keywords***: *Assamese, Author Attributes, Automatic Authorship Detection, Low Resource NLP*

## 1. Introduction

Assamese is an Indo-Aryan language spoken primarily in the Indian state of Assam and neighbouring regions. It has a rich literary tradition, yet there is a paucity of research on author attribution in Assamese text. This study seeks to address this gap by applying a multi-model analysis approach to author attribution detection specifically in the context of Assamese text.

### 1.1 What is Author Attribution?

Author attribution, also known as authorship attribution or authorship identification, is a field within forensic linguistics and stylometry that aims to identify the author of a text based on linguistic features. This task has important applications in literary studies, plagiarism detection, forensic investigations, historical and cultural studies and stylometric analysis. Overall author detection procedure can be a valuable tool in various fields for understanding texts, identifying authors and preserving the integrity of literary works. While author attribution has been extensively studied for many languages, there is a lack of research focusing on underrepresented languages such as Assamese.

### 1.2 Role of Stylometry

In the context of author attribution, stylometry plays a crucial role in distinguishing between different authors based on their writing styles. By comparing the stylometric features of a disputed text with those of known authors, researchers can determine the most likely author of the text. Some common stylometric features used in author attribution include word usage, sentence structure, punctuations used, the lexical diversity seen in the vocabulary, the percentage of function words etc. By combining these and other stylometric features, researchers can create a profile of an author's writing style and use this information for author attribution. Stylometry is particularly useful in cases where other forms of evidence are lacking, such as anonymous or disputed texts.

In this paper, we first provide an overview of the existing literature on author attribution and discuss the challenges specific to author attribution in Assamese text. We then describe our proposed multi-model analysis approach, detailing the various machine learning models used and the linguistic features extracted from the text. Subsequently, we present the experimental setup and discuss the results obtained, highlighting the performance improvements achieved through the multi-model approach.

The primary objective of this study is to explore the performance of multiple machine learning models in the task of author attribution detection for Assamese text. By the integration of machine learning into this task, we aim to improve the accuracy and reliability of author attribution in Assamese text.

Overall, this study contributes to the advancement of authorship attribution research in underrepresented

[1] *Dept. of IT, Gauhati University, Guwahati, Assam, India-781017*
*ORCID ID : 0000-0001-5605-8445*
[2] *Dept. of IT, Gauhati University, Guwahati, Assam, India-781017*
*ORCID ID : 0000-0002-9495-1901*
** Corresponding Author Email: sp.medhi26@gmail.com*

languages and demonstrates the effectiveness of multi-model analysis in enhancing author attribution detection on Assamese text. The findings of this research have implications for the development of computational tools for stylometric analysis in Assamese and other less-studied languages.

## 2 Literature Survey

### 2.1 Advent of NLP

Natural Language Processing (NLP) is a field of artificial intelligence (AI) that focuses on the interaction between computers and humans using natural language. NLP enables computers to understand, interpret, and generate human language in a way that is valuable [1].

Rule-based systems laid the foundation stone of NLP tasks which were primarily developed between 1950s to 1980s. These systems relied on a set of handcrafted rules to analyse and generate natural language text. Along with focusing on analysing the syntax and semantics of the natural language sentence, the revolutionary contribution of rule-based systems as parsers can be seen in [2][3][4] and as knowledge representation in [5]. One of the earliest research papers on rule-based NLP systems is [6], published in 1966. This paper introduces the ELIZA program, a simple rule-based chatbot that simulates a conversation with a Rogerian psychotherapist. ELIZA operates by recognizing keywords in a user's input and generating responses based on pre-defined patterns and rules. Despite their early success, rule-based systems had several limitations. They were labor-intensive to develop, requiring linguists and experts to manually create and maintain rules. Additionally, these systems struggled with ambiguity and context-sensitivity in natural language, leading to errors in analysis and generation.

This process has been greatly improved by the advancements in machine learning and deep learning, which have not only positively impacted various sectors of businesses but also enhanced the efficiency of other subfields of artificial intelligence and NLP.

### 2.2 NLP Tasks in Assamese Language

NLP involves working with the analysis and information retrieval of both spoken and written languages. As far as language is concerned, English language has always been the taking the first place. Since the focus of this paper is Assamese language, we also tried reviewing the different NLP research works attempted by different academicians. Among them, some of the noteworthy and revolutionary ones were developing the POS and UPoS tagger [7][8] and [9] in Bodo Language, Assamese-English translator [10][11][12], Assamese-Bodo translator [13] Word-Sense disambiguation in Assamese [14][15], English-Assamese NMT System [16], Assamese Spelling Corrector [17]. We also found some pretty interesting work on Assamese Stemmer [18], Wordnet [19][20] and Bodo Wordnet [21], Text Summarizer [22]. We could see from here that along with Assamese, the language Bodo has also been started to explore in developing compatible NLP solutions [23]

### 2.3 Global Picture of Authorship Attribution

Authorship attribution is also an important field in natural language processing that aims to determine the author of a given text based on stylistic and linguistic features and numerous techniques and algorithms have been developed to address this use case [24]. Globally this task has been experimented with a lot of languages. Text analysis has been gaining attention in the era of big data, and one crucial aspect of text classification is selecting effective features from datasets. Several feature selection methods have been proposed, but determining the most effective method remains a challenge.

Taking dataset as the centroid, we come across books, scientific journals, news articles, emails, chats from forums etc. were reported under the task of AA. The work in [25] [[26] applied stylometric analysis to email messages written in English language which were taken from the publicly available corpuses and forums. Eventually, one hundred writers were chosen from each of the datasets mentioned above. Finding people based on their writing style was the goal. The accuracy of the expanded feature set was 94%. SVM was used to recognize and categorize the writers, and PCA and KL Transformers were used to improve similarity detection. When it comes to [27][28] the information was gathered from newsgroup discussion threads, emails and online forums. A combination of lexical, syntactic, structural, and content-specific features made up the feature set that was taken into consideration here. Usually, the writers had thought about Arabic and English. They had chosen C4.5, an SVM and decision-tree based methodology. For both datasets, there were differences in the accuracy scores. They obtained an average score of 87.8% for the approach C4.5 and 93% for SVM for the English dataset; in contrast, they obtained an average score of 67.5% for the former and 91.9% for the latter for the Arabic dataset. In [29], sentences of length not longer than 50 were extracted from books. The authors prepared a feature set of seven different types. For the classification task, linear SVM was employed. They divided their experiments into two types wherein one considered standalone features and the other considered an ensemble. The prior achieved an accuracy of 54% while the later technique gave an accuracy of 97.57%.

Considering literary and scientific works as the test bed, as described in [30], SVM classifier was used to attempt author identification when examining literary short texts of well-known English-language novels written by authors Anne and Charlotte Brontë. Using the same dataset, they presented bigrams of syntactic labels as a novel classification feature that outperformed all other prior approaches. They have also noted that the accuracy of these classification tasks can be raised by using brief texts—no more than 200 words—and taking a variety of features into account. Authors' linguistic styles were represented in [31] by de-signing a probabilistic context-free grammar. The classifiers were trained to identify a specific author using the grammar as a language model.

In [32] newspaper articles were leveraged as a dataset to detect the writing style of authors. They created a set 22 style markers to train the classification model. From this experiment, they achieved an accuracy of 81% which outperformed the lexical model. However, upon increasing the number of style markers to 72, there was not an equivalent rise in the accuracy score. They concluded that to extract deep level style markers, the number of words should not be greater than 1000. Their proposed technique will work best to understand the linguistic style of a single author and is not meant to extend the study to multiple authors.

A few lexical n-grams are used by the model in [33] as the differentiating element. In addition to 3,000 passages from three Bengali authors, the corpus featured an end-to-end authorship classification system based on character n-grams, feature selection for authorship attribution, feature ranking and analysis, and a learning curve to assess the relationship between test accuracy and training data volume.

Author classification techniques in Hindi literature are examined in [34]'s work, both supervised and unsupervised. The corpus was self-curated and mostly comprised of novels written in Hindi by well-known authors, including Sarat Chandra Chattopadhyay, Dhamarvir Bharati, Premchand, Vibhuti Narayan, and Rabindranath Tagore. Their method involved building the feature vector using lexical bigrams, trigrams, and unigrams. The concatenation of the bigram and trigram feature vectors was another way that the MDA concept was utilized. The observations presented in this work indicate that the classification scores are highly dependent on the granularity of a dataset. The accuracy improves with the number of different feature vectors.

In order to categorize authors of Kannada texts, the authors in [35] focused on a machine learning algorithm based on the author's profile. They were able to obtain an accuracy score of 88%. The task of authorship attribution was also attempted for Brazilian language as described in [36]. The corpus here was composed of 20 authors who wrote different contents on economics, sports, literature etc. A similar effort was also made for Nigerian language [37] where the news articles written by 13 authors was taken as the dataset to study the stylistic characteristics.

Another study conducted by Mihalcea and Radev focused specifically on authorship attribution in the Romanian language [38]. They used a combination of lexical, syntactic, and semantic features, as well as machine learning algorithms, such as Naive Bayes and Decision Trees to determine the authors of Romanian texts with high accuracy. In [39] authorship attribution in the Telugu language was focused. They used a combination of word n-grams, character n-grams, and syntactic features, along with machine learning algorithms like Support Vector Machines and Decision Trees, to accurately identify the authors of Telugu texts.

Authorship attribution research has been a global endeavor in the field of linguistics. Discovering the real owners of certain early literary works that are still relevant has always captivated people. One such study is described in [40], where the authors were identified through the use of modern Hebrew passages. Along with Phoenician and a few other lesser-known languages, Hebrew is one of the few Northwest Semitic languages that has persisted to the present day. Nine distinct authors of early modern Hebrew literary works and web blog posts were used to create the diverse corpus that makes up the dataset in the work mentioned above. They used two different methods, SVM and Markov chains, respectively. The literary corpus that used SVM showed a higher level of accuracy of 98%, compared to the blog corpus's 75% accuracy. Few other works were also attempted in Persian [41][42], Russian language [43] and Chinese Language [44]. Using a TF-IDF scheme, the authors in [41] have tried to propose a computational method to identify disputed persian authors. They were able to generate an accuracy of 93.1%. On the other hand, the study in [42] considered four different authors with two each from different periods and tried to explore the contribution of function words and author specific style markers in the use case of author attribution. The work in [43] exploited short comments and posts from social media handles of different people. Mentioning about their methodology, they implemented both stand-alone and ensemble models of machine learning. This study generated an accuracy of 73.2%. Across the globe, around 95% of inhabitants speak the language Chinese. In [44] the corpus consisted of 20 poets of the Tang Dynasty. Here, they experimented with CNN and other related Transformer models on classical Chinese texts.

The literature review mentioned above demonstrates that while the task of authorship attribution has been applied to many globally spoken languages, particularly English, very little or no work has been reported to date for some languages, such as Assamese, Sindhi, and Marathi. Figure 1 further illustrates the fact that, despite Assamese being the third most spoken dialect, no authorship attribution system has yet been implemented. This paper represents the first attempt to test the effectiveness of various machine learning models on texts written in Assamese.

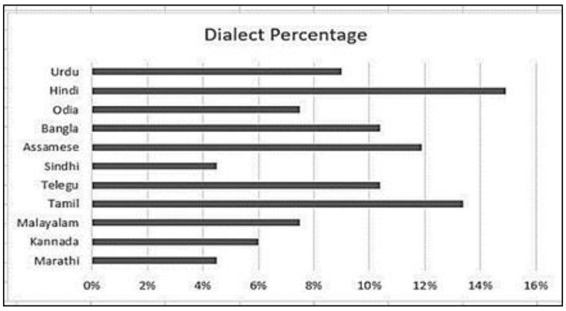


**Fig.1.** Dialect percentage of Indo-Aryan and Dravidian Languages in India

## 3 Methodology

Authorship attribution in NLP using machine learning generally involves the process of feature extraction, feature selection, model selection, and finally training, testing and evaluating the model. However, when it comes to parsing Assamese language, there are a few challenges involved. It is a morphologically rich free-order language belonging to the Indo-Aryan roots. It is considered as a bridge language of the eastern-most region of India spoken by 23 million speakers [45]. Its origin dates back to the 7th century [46] and its writings are based on assamese alphabet or Oxomiya bornomala, a child script known to be evolved from the Kamarupi Script. In 1917 the Assamese literary society known as the "Assam Sahitya Sabha" was established with the goal of promoting Assamese literature and laying the groundwork for its ongoing growth. Parsing an assamese text is still considered an open problem as stated in [47]. It comprises of 41 consonant and 11 vowel letters and unlike the English language, it is a one-case alphabetic system. Lately, as mentioned in the literature review, researchers have started working on the upliftment of the language. However, the libraries and modules required for 360-degree parsing and understanding the language is still in its developing phase.

The first step in the approach used in this work was loading and reading the dataset using basic Python libraries. In NLP, tokenization is a crucial step toward comprehending the distribution, count, and range of vocabulary. To do the same, we have used indic NLP libraries. To extract features, we used the Bag of Words method. It is necessary to comprehend the author's word choice and frequency in order to analyse their stylistic pattern. Therefore, the BOW model aids in our information gathering. We tested the performance of six models, namely Multinomial Naïve Bayes, Gaussian Naïve Bayes, Logistic Regression, Support Vector Machine, K-Nearest Neighbor, and Random Forest Classifiers, after dividing the data into an 80:20 ratio. Eventually, the performance metrics of all the classifiers were analyzed.

## 4. Dataset Statistics

The dataset utilized for this work has been manually curated. We went through the famous literary works of 9 most famous assamese poets/novelists namely Banikanta Kakati, Dandinath Kalita, Hemaprava Das, Kirtinath Hazarika, Lakhinandan Bora, Lakshminath Bezbarua, Nilima Bora, Nilima Dutta, Rajani Kanta

**Table 1.** Brief Biodata of the Authors

| SL. No. | Author Name | Presence | Penname (if any) | Famous As |
|---|---|---|---|---|
| 1 | Banikanta Kakati | 15 November 1894- 15 November 1952 | Bhabananda Pathak | Prominent linguist, literary figure, critic and scholar in Assamese language |
| 2 | Dandinath Kalita | June 30, 1890- May 15, 1955 | None | Assamese poet, short story writer, novelist, playwright and satirical writer |
| 3 | Hemaprava Das | 1886-1945 | None | Assamese Literary Writer |
| 4 | Kirtinath Hazarika | 22 August 1922- May 3, 2002 | (Dadai) দদাই | Literary Figure |
| 5 | Lakhinandan Bora | 15 June 1932- 3 June 2021 | None | Short Story Writer, Scientist, Novelist of |
| 6 | Lakshminath Bezbarua | 14 October 1864- March 26, 1938 | Roxoraj, Sahityarathi | Writer, Novelist, Dramatist, Poet, Editor, Satirist |
| 7 | Nilima Bora | Data Not Available | None | Assamese Writer (Specialist in Children Short Stories) |
| 8 | Nilima Dutta | | | |
| 9 | Rajani Kanta Bordoloi | 24 November 1867- 25 March 1940 | Bholai Sharma | Noted writer, Journalist |

Bordoloi. The details of them are included in Table 1.

With the help of digital media, we were successful in retrieving the e-books of their work. The next herculean task was to perform the OCR on the collected e-books and convert them into readable text files. With the help of the online utilities, we could complete this task. Finally, a CSV file was generated out of all the data collected against each author. Figure 2 describes the count of text files collected against all the authors. The data in Table 2 describes the count of words, tokens and average length of sentences in the corpus. Further analysis of the sentence length of the corpus on the basis of number of words is illustrated in Table 3.

### 4.1 Statistics of the corpus

**Table 2.** Count of Tokens, Words and Sentences

| Author Name | No. of Text Files | No. of Sentences | No. of Tokens | Average Length of Sentences |
|---|---|---|---|---|
| Banikanta Kakati | 159 | 3507 | 49135 | 11 |
| Dandinath Kalita | 503 | 8182 | 110158 | 11 |
| Hemaprava Das | 37 | 555 | 8428 | 13 |
| Kirtinath Hazarika | 114 | 6804 | 81011 | 9 |
| Lakhinandan Bora | 165 | 9065 | 102848 | 9 |
| Lakhminath Bezbarua | 201 | 10579 | 138896 | 11 |
| Nilima Bora | 170 | 4091 | 38937 | 7 |
| Nilima Dutta | 138 | 5991 | 64532 | 8 |
| Rajani Kanta Bordoloi | 218 | 6970 | 83855 | 9 |
| Total | 1705 | 55,744 | 6,77,800 | |

**Table 3.** Bucket-wise Sentence Count of the Corpus

| Author Name | Bucket Range | Quantity |
|---|---|---|
| Banikanta Kakati | B1(<10 words) | 30 |
| | B2(10-20 words) | 147 |
| | B3(20-30 words) | 318 |
| | B4(30-40 words) | 383 |
| | B5(>40 words) | 2620 |
| Dandinath Kalita | B1(<10 words) | 287 |
| | B2(10-20 words) | 463 |
| | B3(20-30 words) | 748 |
| | B4(30-40 words) | 898 |
| | B5(>40 words) | 5715 |
| Hemaprava Das | B1(<10 words) | 23 |
| | B2(10-20 words) | 19 |
| | B3(20-30 words) | 23 |
| | B4(30-40 words) | 54 |
| | B5(>40 words) | 435 |
| Kirtinath Hazarika | B1(<10 words) | 344 |
| | B2(10-20 words) | 761 |
| | B3(20-30 words) | 1019 |
| | B4(30-40 words) | 952 |
| | B5(>40 words) | 3684 |
| Lakhinandan Bora | B1(<10 words) | 170 |

|  |  |  |
|---|---|---|
|  | B2(10-20 words) | 953 |
|  | B3(20-30 words) | 1341 |
|  | B4(30-40 words) | 1411 |
|  | B5(>40 words) | 5151 |
| Lakhminath Bezbarua | B1(<10 words) | 1083 |
|  | B2(10-20 words) | 838 |
|  | B3(20-30 words) | 1772 |
|  | B4(30-40 words) | 1421 |
|  | B5(>40 words) | 5384 |
| Nilima Bora | B1(<10 words) | 55 |
|  | B2(10-20 words) | 500 |
|  | B3(20-30 words) | 736 |
|  | B4(30-40 words) | 759 |
|  | B5(>40 words) | 2018 |
| Nilima Dutta | B1(<10 words) | 591 |
|  | B2(10-20 words) | 785 |
|  | B3(20-30 words) | 774 |
|  | B4(30-40 words) | 711 |
|  | B5(>40 words) | 3032 |
| Rajani Kanta Bordoloi | B1(<10 words) | 208 |
|  | B2(10-20 words) | 799 |
|  | B3(20-30 words) | 1088 |
|  | B4(30-40 words) | 941 |
|  | B5(>40 words) | 3901 |

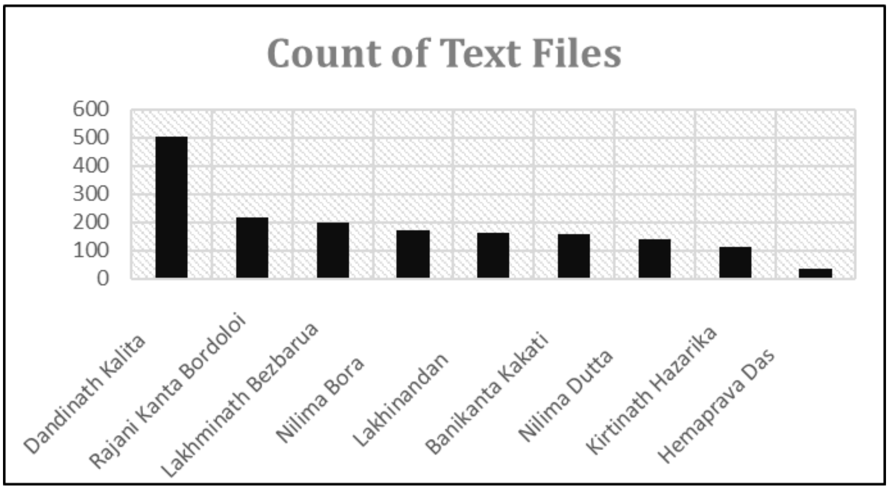
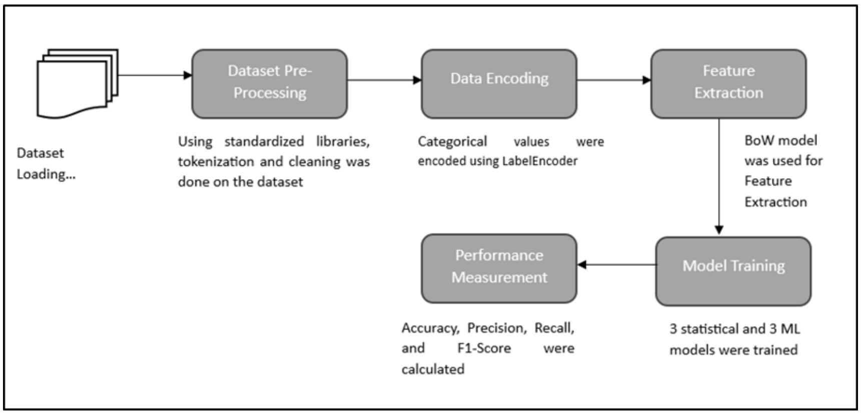


**Fig.2.** Distribution of text corpora collated against each author

## 5. Experiment Details:



Using standardized libraries, tokenization and cleaning was done on the dataset

Categorical values were encoded using LabelEncoder

BoW model was used for Feature Extraction

3 statistical and 3 ML models were trained

Accuracy, Precision, Recall, and F1-Score were calculated

**Fig.3.** Flow diagram of the Experiment

The steps of the experiment of our reported work is described in detail in the following steps:

### 5.1 Dataset Loading

The dataset constructed was loaded with the help of python libraries like panda and were converted to the required Data Frame. Through this we could visualize the layout of the dataset generated. The shape of the dataset was 1705X2.

### 5.2 Dataset Pre-processing

Raw data consists of a lot of noise. These noisy values affect the overall processing and analysis of the models. Hence, data needs to be pre-processed before forwarding it for further analysis. In our work, we have taken up noise removal techniques like tokenization using Indic NLP libraries, a customized function for stop word removal, and BoW model for data vectorization.

### 5.3 Encoding of Categorical Values

Machines can only process numerical values. The centroid of our work is completely textual data. Specifically, we had to deal with the categorical labels to numerical values. We used the LabelEncoder from the sklearn.preprocessing module to fulfil this.

### 5.4 Splitting Data into Train and Test samples

The complete pre-processed data was divided into 80:20 proportions with the former being used for training and the later to be used for testing. With respect to our corpus, among 1705 samples, 1364 samples were taken for training and the rest 341 samples were reserved for the training process.

### 5.5 Feature Extraction and Vectorization

The feature extraction was done using BoW and TF-IDF techniques. For each text file in the corpus, a vector is created where each element corresponds to the frequency of a word in the vocabulary within that document. This means that the length of the vector is equal to the size of the vocabulary.

### 5.6 Training and Testing the Model

Six models were trained on the corpus. Diverse testbeds were designed to test the efficacy of the models. The first two experiments 1(a) and Exp 1(b) were conducted to test the effect of different tokenizers on the accuracy of the models. We had considered two tokenizers, namely, Indic NLP and Moses. The second experiment, Exp 2 was conducted to perform a raw training on the dataset and check on the accuracy of the different models. The numerical values of precision, recall, and F1 score are macro-average values.

**Table 4.** Results from Experiment 1(a)

| *Model* | *Accuracy* | *Precision* | *Recall* | *F1 Score* |
|---|---|---|---|---|
| MNB | 76.8 | 77.0 | 82.0 | 78.0 |
| GNB | 48 | 65.0 | 64.0 | 55.0 |
| LR | 89.4 | 89.0 | 88.0 | 89.0 |
| SVC | 85.0 | 83.0 | 84.0 | 83.0 |
| KNN | 69.7 | 73.0 | 66.0 | 66.0 |
| RF | 88.2 | 92.0 | 87.0 | 89.0 |

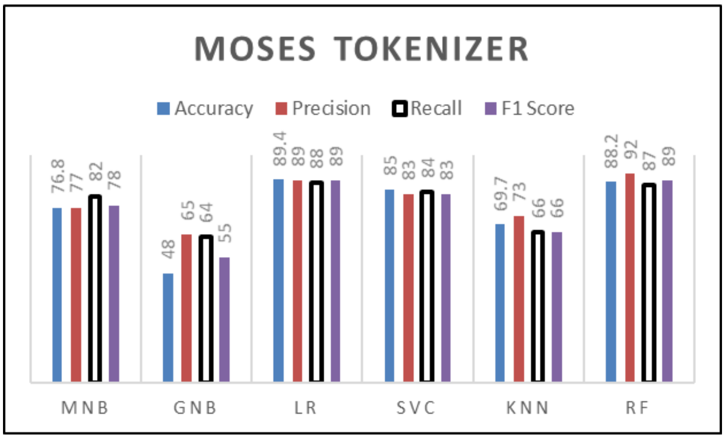


**Fig.4.** Evaluation Metric Statistics from Experiment 1(a)

**Table 5.** Results from Experiment 1(b)

| ***Model*** | ***Accuracy*** | ***Precision*** | ***Recall*** | ***F1 Score*** |
|---|---|---|---|---|
| MNB | 78.5 | 81.0 | 83.0 | 82.0 |
| GNB | 50.7 | 68.0 | 67.0 | 58.0 |
| LR | 89.1 | 89.0 | 88.0 | 89.0 |
| SVC | 89.1 | 87.0 | 88.0 | 87.0 |
| KNN | 72.1 | 79.0 | 67.0 | 69.0 |
| RF | 93.2 | 95.0 | 93.0 | 94.0 |

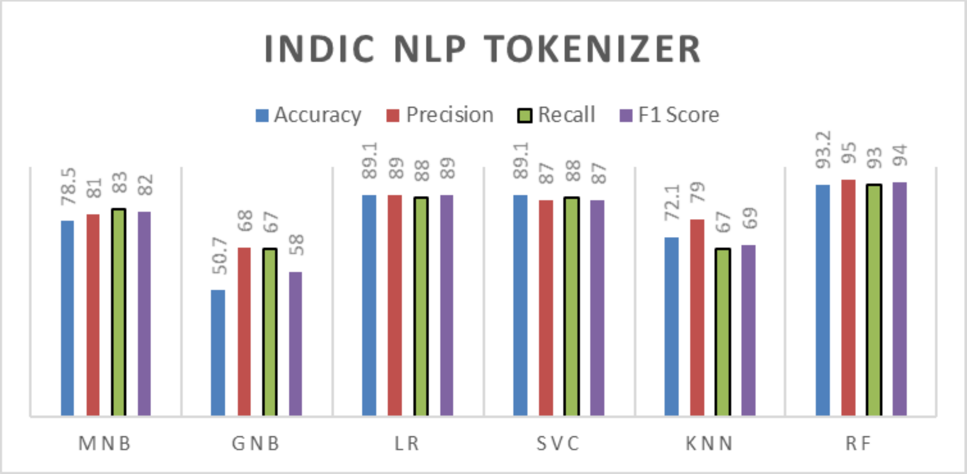


**Fig.5.** Evaluation Metric Statistics from Experiment 1(b)

**Table 6.** Results from Experiment 2

| ***Model*** | ***Accuracy*** | ***Precision*** | ***Recall*** | ***F1 Score*** |
|---|---|---|---|---|
| MNB | 78.5 | 81.0 | 83.0 | 81.0 |
| GNB | 50.7 | 70.0 | 67.0 | 59.0 |
| LR | 89.7 | 89.0 | 89.0 | 89.0 |
| SVC | 90.0 | 88.0 | 89.0 | 88.0 |
| KNN | 72.7 | 81.0 | 67.0 | 69.0 |
| RF | 93.2 | 94.0 | 92.0 | 93.0 |

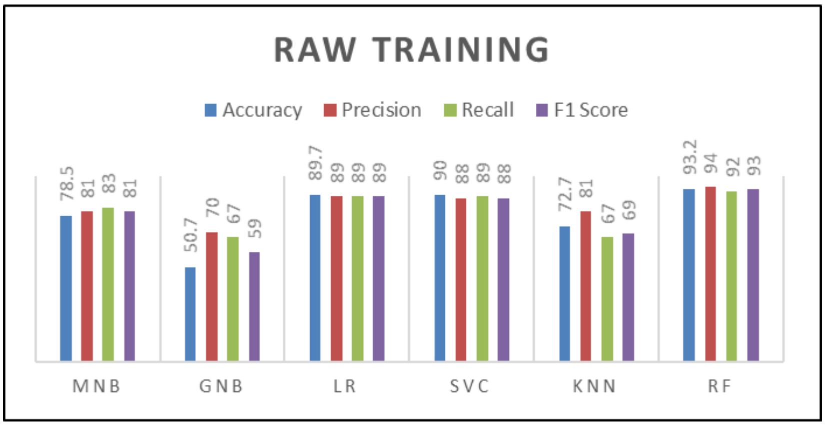


**Fig.6.** Evaluation Metric Statistics from Experiment 2

### 5.7 Learning Curves

Learning curves are a useful tool in machine learning to understand how the performance of a model changes as the amount of training data increases. In the context of text classification, learning curves can help you analyze the impact of dataset size on your model's performance. In our study also we generated learning curves for all the six classifier models. Figures 7 describes their results. We were able to achieve some decent results from Logistic Regression and Random Forest Classifier Model.

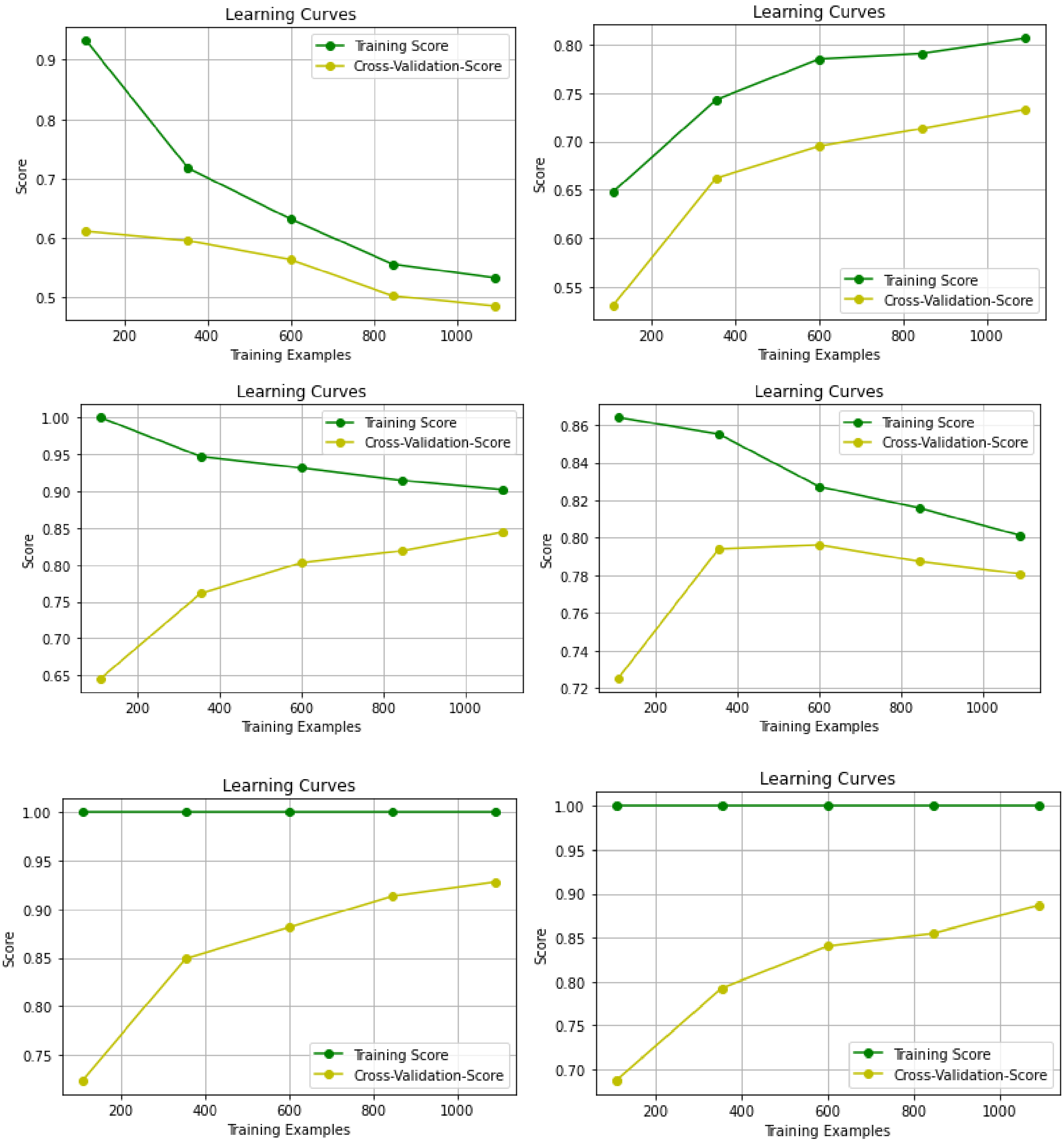


**Fig.7.** Learning Curves of all the six classifier models

## 6 Results and Analysis

From the above experiments, we can conclude that using Indic NLP Tokenizer, the models portrayed elevated levels of accuracy in comparison to Moses Tokenizer. However, irrespective of the type of tokenizer, model Random Forest Classifier outperformed the other models. However, from the learning curve point of view, Logistic Regression Model also showcased some promising results. From the experiment 2, also we can put forward the insight that the Random Forest model gave the best performance with accuracy 93.2%. The lowest performance in all the scenarios was shown by Gaussian Naïve Bayes Classifier. In experiment 1(a) and 1(b), the accuracy values were 45.7% and 48.6% respectively. And in experiment 2, it gave a bit better accuracy with 50.7%.

Further the confusion matrices of all the experiments can be seen in figure 8.

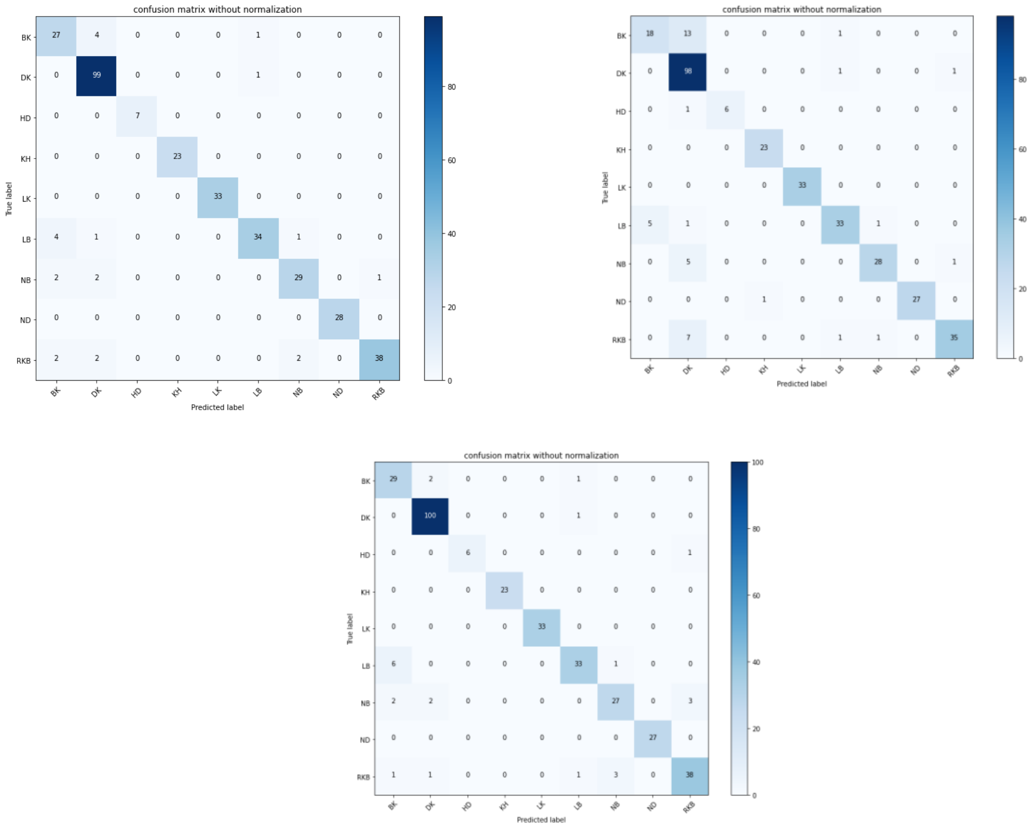


**Fig.7.** Confusion Metric Results of Experiment 1(a), 1(b) and 2

## 7 Conclusion

In conclusion, our study demonstrates the effectiveness of various machine learning techniques in authorship attribution of Assamese text. Through feature engineering and selection, we were able to achieve a high level of accuracy in identifying the authors of Assamese texts. Our experiments with different classifiers highlight the importance of choosing the right algorithm for the task.

Despite the promising results, there are several avenues for future work in this area. Firstly, exploring deep learning models, such as LSTM or Transformer-based architectures, could potentially improve the performance of authorship attribution on Assamese text. Additionally, incorporating more advanced linguistic features, such as syntactic and semantic features, may further enhance the accuracy of the models.

Furthermore, expanding the dataset to include a larger and more diverse set of authors and texts could help in building more robust models. Additionally, investigating the impact of different writing styles, genres, and topics on authorship attribution could lead to a deeper understanding of the underlying patterns in the text.

Overall, authorship attribution of Assamese text is a challenging yet promising area of research specifically for low-resource languages, and further exploration of advanced techniques and datasets could lead to significant advancements in this field.

### 7.1 Acknowledgment

The NLP and Language Technology Lab at Gauhati University's Department of IT provided the authors with invaluable guidance during the development of the standalone corpus, for which they are sincerely grateful.

### Author contributions

**Prof. Shikhar Kr. Sarma:** Conceptualization, Methodology, Software, Field study

**Ms. Smriti Priya Medhi:** Data curation, Writing-Original draft preparation, Software, Validation., Field study, Writing-Reviewing and Editing

### Conflicts of interest

The authors declare no conflicts of interest.

## References

[1] S. Garg, D. S. Panwar, A. Gupta, and R. Katarya, “A literature review on sentiment analysis techniques involving social media platforms,” PDGC 2020 - 2020 6th Int. Conf. Parallel, Distrib. Grid Comput., pp. 254–259, Nov. 2020, doi: 10.1109/PDGC50313.2020.9315735.

[2] W. A. Woods, “Transition Network Grammars for

Natural Language Analysis," Commun. ACM, vol. 13, no. 10, pp. 591–606, 1970, doi: 10.1145/355598.362773.

[3] T. Winograd, "Understanding natural language," Cogn. Psychol., vol. 3, no. 1, pp. 1–191, Jan. 1972, doi: 10.1016/0010-0285(72)90002-3.

[4] W. M. Reynolds and G. E. Miller, of of Psychology, vol. 5. 2003.

[5] S. E. Fahlman, "Representing and Using Real-World Knowledge.," Energy Technology Review, vol. 1. pp. 451–470, 1979.

[6] J. Weizenbaum, "ELIZA-A computer program for the study of natural language communication between man and machine," Commun. ACM, vol. 9, no. 1, pp. 36–45, 1966, doi: 10.1145/365153.365168.

[7] K. Talukdar and S. K. Sarma, "Parts of Speech Taggers for Indo Aryan Languages: A critical Review of Approaches and Performances," 2023 4th Int. Conf. Comput. Commun. Syst. I3CS 2023, 2023, doi: 10.1109/I3CS58314.2023.10127336.

[8] Kuwali Talukdar Shikhar Kumar Sarma, "UPoS Tagger for Low Resource Assamese Language: LSTM and BiLSTM based Modelling," 2023 IEEE Int. Conf. Mach. Learn. Appl. Netw. Technol., pp. 1–6, 2023.

[9] B. Basumatary, M. Rahman, and S. K. Sarma, "Deep Learning Based Bodo Parts of Speech Taggere," IEEE Explor. 2023 14th Int. Conf. Comput. Commun. Netw. Technol., pp. 1–5, 2023.

[10] K. Kanchan Baruah, P. Das, A. Hannan, and S. Kr Sarma, "Assamese-English Bilingual Machine Translation," Int. J. Nat. Lang. Comput., vol. 3, no. 3, pp. 73–82, 2014, doi: 10.5121/ijnlc.2014.3307.

[11] M. A. Ahmed, K. Talukdar, P. A. Boruah, S. K. Sarma, and K. Kashyap, "GUIT-NLP's submission to Shared Task: Low Resource Indic Language Translation," Conf. Mach. Transl. - Proc., pp. 933–938, 2023, doi: 10.18653/v1/2023.wmt-1.87.

[12] K. K. Kashyap, S. K. Sarma, and M. A. Ahmed, "Improving translation between English, Assamese bilingual pair with monolingual data, length penalty and model averaging," 2024.

[13] K. Talukdar, S. K. Sarma, F. Naznin, and K. K. Kashyap, "Influence of Data Quality and Quantity on Assamese-Bodo Neural Machine Translation," IEEE Explor. 2023 14th Int. Conf. Comput. Commun. Netw. Technol., pp. 1–5, 2023.

[14] J. Sarmah and S. Kumar Sarma, "Survey on Word Sense Disambiguation: An Initiative towards an Indo-Aryan Language," Int. J. Eng. Manuf., vol. 6, no. 3, pp. 37–52, 2016, doi: 10.5815/ijem.2016.03.04.

[15] J. Sarmah and S. Kr., "Decision Tree based Supervised Word Sense Disambiguation for Assamese," Int. J. Comput. Appl., vol. 141, no. 1, pp. 42–48, 2016, doi: 10.5120/ijca2016909488.

[16] M. A. Ahmed, K. K. Kashyap, and S. K. Sarma, "Pre-processing and Resource Modelling for English-Assamese NMT System," 4th Int. Conf. Comput. Commun. Syst., pp. 1–6, 2023.

[17] M. P. Bhuyan and S. K. Sarma, "Automatic Formation, Termination Correction of Assamese word using Predictive Syntactic NLP," Proc. 3rd Int. Conf. Commun. Electron. Syst. ICCES 2018, pp. 544–548, Oct. 2018, doi: 10.1109/CESYS.2018.8724023.

[18] A. K. Barman, J. Sarmah, and S. K. Sarma, "Development of assamese rule based stemmer using WordNet," Proc. 10th Glob. WordNet Conf., pp. 135–139, 2020.

[19] A. K. Barman, J. Sarmah, and S. K. Sarma, "WordNet based information retrieval system for assamese," Proc. - UKSim 15th Int. Conf. Comput. Model. Simulation, UKSim 2013, pp. 480–484, 2013, doi: 10.1109/UKSIM.2013.90.

[20] S. Kr and S. Dibyajyoti, "Building Multilingual Lexical Resources Using Wordnets : Structure , Design and Implementation," vol. 1, no. December, pp. 161–170, 2012.

[21] S. K. Sarma, B. Brahma, M. Gogoi, and M. B. Ramchiary, "A Wordnet for Bodo language: Structure and development," Glob. Wordnet Conf. GWC 2010, 2010.

[22] N. Baruah, S. K. Sarma, and S. Borkotokey, "Evaluation of Content Compaction in Assamese Language," Procedia Comput. Sci., vol. 171, pp. 2275–2284, Jan. 2020, doi: 10.1016/J.PROCS.2020.04.246.

[23] B. Brahma, A. K. Barman, P. Shikhar, K. Sarma, and B. Boro, "Corpus Building of Literary Lesser Rich Language-Bodo: Insights and Challenges," vol. 1, no. December, pp. 29–34, 2012.

[24] S. Swain, G. Mishra, and C. Sindhu, "Recent approaches on authorship attribution techniques-An overview," Proc. Int. Conf. Electron. Commun. Aerosp. Technol. ICECA 2017, vol. 2017-Janua, no. October, pp. 557–566, 2017, doi: 10.1109/ICECA.2017.8203599.

[25] A. Abbasi and H. Chen, "Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace," ACM Trans. Inf. Syst., vol. 26, no. 2, 2008, doi: 10.1145/1344411.1344413.

[26] J. Burrows, "'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship," Lit. Linguist. Comput., vol. 17, no. 3, pp. 267–287, Sep. 2002, doi: 10.1093/LLC/17.3.267.

[27] A. Abbasi and H. Chen, "Analysis to Extremist-," IEEE Intell. Syst., no. October, pp. 67–75, 2005.

[28] S. Argamon, M. Šarić, and S. S. Stein, "Style mining of electronic messages for multiple authorship discrimination: First results," Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., pp. 475–480, 2003, doi: 10.1145/956750.956805.

[29] M. Gamon, "Linguistic correlates of style: Authorship classification with deep linguistic analysis features," COLING 2004 - Proc. 20th Int. Conf. Comput. Linguist., 2004.

[30] G. Hirst and O. Feiguina, "Bigrams of syntactic labels for authorship discrimination of short texts," Lit. Linguist. Comput., vol. 22, no. 4, pp. 405–417, 2007, doi: 10.1093/llc/fqm023.

[31] S. Raghavan, A. Kovashka, and R. Mooney, "Authorship attribution using probabilistic context-free grammars," ACL 2010 - 48th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf., pp. 38–42, 2010.

[32] E. Stamatatos, N. Fakotakis, and G. Kokkinakis, "Computer-based authorship attribution without lexical measures," Lang. Resour. Eval., vol. 35, no. 2, pp. 193–214, 2001.

[33] D. S. Sharma, R. Sangal, S. Proc, S. Phani, S. Lahiri, and A. Biswas, "Authorship Attribution in Bengali Language," NLP Association of India. NLPAI, pp. 100–105, 2015, Accessed: Mar. 21, 2024. [Online]. Available: https://aclanthology.org/W15-5915.

[34] J. S. Kallimani, C. P. Chandrika, A. Singh, and Z. Khan, "Authorship Identification Using Supervised Learning and n-Grams for Hindi Language," J. Comput. Theor. Nanosci., vol. 17, no. 9, pp. 4258–4261, Dec. 2020, doi: 10.1166/JCTN.2020.9058.

[35] C. P. Chandrika and J. S. Kallimani, "Authorship Attribution on Kannada Text using Bi-Directional LSTM Technique," Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 9, pp. 963–971, Dec. 2022, doi: 10.14569/IJACSA.2022.01309111.

[36] W. Oliveira, E. Justino, and L. S. Oliveira, "Comparing compression models for authorship attribution," Forensic Sci. Int., vol. 228, no. 1–3, pp. 100–104, 2013, doi: 10.1016/j.forsciint.2013.02.025.

[37] I. I. Ayogu and V. A. Olutayo, "Authorship Attribution using Rough Sets based Feature Selection Techniques Authorship Attribution using Rough Sets based Feature Selection Techniques," no. May, 2020, doi: 10.5120/ijca2016911889.

[38] S. Avram and M. Oltean, "A Comparison of Several AI Techniques for Authorship Attribution on Romanian Texts," pp. 1–40, 2022.

[39] S. Nagaprasad, N. Krishnaveni, J. K. R. Sastry, and A. Vinayababu, "On authorship attribution of telugu text," Indian J. Sci. Technol., vol. 9, no. 35, pp. 1–7, 2016, doi: 10.17485/ijst/2016/v9i35/98735.

[40] "Faculty of Natural Sciences Department of Computer Sciences Authorship Attribution in Modern Hebrew Presented By David Gabay."

[41] R. Ramezani, "A language-independent authorship attribution approach for author identification of text documents," Expert Syst. Appl., vol. 180, no. May 2021, 2021, doi: 10.1016/j.eswa.2021.115139.

[42] R. Modaber Dabagh, "Authorship attribution and statistical text analysis," Adv. Methodol. Stat., vol. 4, no. 2, pp. 149–163, 2007, doi: 10.51936/uvjx7198.

[43] E. Reisi and H. M. Farimani, "a Uthorship a Ttribution in H Istorical and L Iterary T Exts By," pp. 118–127, 2021, doi: 10.22034/jaisis.2021.269735.1018.

[44] H. Wang and A. Riddell, "CCTAA: A Reproducible Corpus for Chinese Authorship Attribution Research," 2022 Lang. Resour. Eval. Conf. Lr. 2022, no. June, pp. 5889–5893, 2022.

[45] "C-17: Population by bilingualism and trilingualism, India - 2011." https://censusindia.gov.in/nada/index.php/catalog/10262.

[46] W. Bright and R. C. Nigam, "Grammatical Sketches of Indian Languages, with Comparative Vocabulary and Texts (Part I)," Language (Baltim)., vol. 54, no. 1, p. 247, Mar. 1978, doi: 10.2307/413037.

[47] N. Saharia, "A First Step Towards Parsing of Assamese Text," Spec. Vol. Probl. Parsing Indian Lang., vol. 11, no. 5, pp. 30–34, 2011.