Journal of Information Systems Engineering and Management
2025, 10(40s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Ensemble Deep Learning for Author Attribution in Assamese Text Documents: A Hybrid Approach

Smriti Priya Medhi[12], Prof. Shikhar Kumar Sarma[1]
[1]*Department of Information Technology, Gauhati University*
[2]*Department of Computer Science and Engineering, Assam Don Bosco University*

ARTICLE INFO

Received: 24 Dec 2024

Revised: 12 Feb 2025

Accepted: 26 Feb 2025

ABSTRACT

Introduction: Author attribution is a critical task in the domain of computational linguistics, particularly when dealing with low-resourced languages. These languages often lack sufficient annotated datasets and robust linguistic tools, making natural language processing (NLP) applications challenging. A piece of text generally reflects the unique stylistic traits of its author, including their use of vocabulary, punctuation, and sentence structures. Identifying the correct author based on these textual cues is the central aim of author attribution. Despite its importance, this area remains relatively unexplored for low-resourced languages due to the inherent data and resource limitations.

Objectives: This study aims to address the problem of author attribution in low-resourced languages by leveraging deep learning techniques. Specifically, the objectives are to investigate how neural network models like Recurrent Neural Networks (RNN) and hybrid Convolutional Neural Network with Long Short-Term Memory (CNN-LSTM) can effectively capture stylistic features of text, and to design an ensemble model that combines these approaches for enhanced performance in multi-author scenarios.

Methods: To achieve the stated objectives, we employed RNN and CNN-LSTM architectures to model the stylistic nuances present in textual data. These models were trained and tested on the AAALC dataset, which contains writings from multiple authors in a low-resourced language. The performance of each model was evaluated using key classification metrics such as accuracy and F1-score. Additionally, an ensemble model combining the outputs of RNN and CNN-LSTM was proposed to further boost classification performance by leveraging the strengths of both architectures.

Results: The results of our experiments demonstrated that deep learning models are highly effective in the context of author attribution for low-resourced languages. Among the evaluated models, the ensemble approach achieved the best results, with an accuracy of 84.38% and an F1-score of 79.0%. These results indicate that combining neural network architectures can significantly improve classification accuracy by capturing a richer representation of stylistic features.

Conclusions: In conclusion, this study validates the potential of neural network-based models for tackling the author attribution task in low-resourced languages. The proposed ensemble model not only achieved strong performance metrics but also demonstrated the practical viability of deep learning approaches in linguistic contexts where traditional resources are scarce. These findings contribute to the advancement of inclusive NLP systems and offer valuable insights for further research in multilingual and resource-constrained settings.

Keywords: Author Attribution, Low-Resourced Languages, Neural Networks, Ensemble Learning

INTRODUCTION

Today in the 21st century, digital literacy has become a part and parcel of our lives. We can see how the world is shifting towards becoming a digital native gradually. The World Economic Forum states that digital literacy has become a vital life skill and is now a part of the 21st-century toolkit, even though numeracy and basic literacy are still essential for learning. Every global citizen now needs digital skills beyond basic literacy, whether they are for

Journal of Information Systems Engineering and Management
2025, 10(40s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

socializing, communication, job searching, or receiving a thorough education[1]. In urban India, casual workers have the lowest level of digital literacy 30%, while regular wage or salaried workers have the highest 73% [2]. With everything available to access in a digital bulb called internet, it is very easy to understand that there are probabilities of people misusing those information or data for their own selfish purpose. Re-using images, videos or texts which carries the ownership of some other individual, without giving due credit to the original creator is one such angle of crime that has recently seen coming into light. In formal terms, it is defined as digital plagiarism. Plagiarism may take numerous forms, such as explicitly copying and pasting text from a source, paraphrasing someone else's work without giving adequate credit, or using someone else's research or ideas without giving credit. Images, videos, and other media can also be plagiarized in the digital age [3]. With digital plagiarism in the rise, research to devise techniques and methods either to prevent or detect such tasks came into light. These techniques involved tasks such as Author Attribution, Author Verification, Author Profiling etc. The work reported in this paper involves the study of Author Attribution of literary texts using deep learning models.

Over the last 25 years, around 18,100 papers were reported as per google scholar database [4] on using ensemble model for authorship attribution. Out of these, 17,200 papers were done majorly on English datasets. From these around 346 papers were found to be drafted on Indic Languages and around 13 were found to be reported on low-resource languages. The fact that very less research was done related to low-resource languages is because, firstly, there are no standard datasets available to work upon. Secondly, absence of standard libraries and pre-trained models to embed the data points for a machine to learn. Whatever research has been reported so far regarding low-resource languages on various NLP tasks, involves a rigorous process starting with data collection to training models and evaluating results.

LITERATURE REVIEW

In this section we discuss the existing work on authorship attribution reported in the last couple of years with specific reference to high resource and low resource languages. The writing style of an author can be studied via both as a supervised and unsupervised problem. But according to [5], they proposed a new taxonomy of author attribution models with sub-divides into five namely, stylistic, statistical, language models, machine learning models and deep learning models.

Stylistic models typically study the patterns underlying in the word usage by the different authors. They consider these patterns can be considered as differentiating markers between the writing style of the authors under the scanner.

A first ever attempt to study this factor was taken up by[6]. The Federalist problem is used in this paper to evaluate the usefulness of three new stylometric approaches. With the help of genetic algorithm, they studied the importance of most frequent words as a significant factor in authorial style.

After them, many researchers took up the work of studying the federalist papers using different techniques and methodologies. Apart from using word frequencies as stylistic markers, an attempt was made to study the contribution of punctuation symbols as a unique marker by [7].

Apart from this case study, there were many other reported works that tried studying the stylometric features of many other datasets. Among them some significant ones were [8] whose study shows that Thomas Middleton's plays can be identified based on common word use, helping confirm his authorship of some disputed works and highlighting his unique writing style.

[9] on the other hand performed authorship attribution studies on the scripts of Historiae Augustae and argues that stylometric analysis of text subsets yields different results than analyzing entire texts, indicating that individual texts are not internally homogeneous.

Another work by [10] however proposed a novel technique where style markers were extracted using NLP analysis and they proved that their technique gave better results rather than using traditional lexical features.

Considering the second category where authorship attribution is established using statistical techniques, many approaches were proposed which involves usage of statistical assumptions to find the outcome of the said task. Popular statistical models that are commonly utilized for such tasks Support Vector Machines (SVM), Logistic

Journal of Information Systems Engineering and Management
2025, 10(40s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Regression, Random Forest, Decision Trees, Naïve Bayes etc. A paper that has studied the performance of all these models in order to investigate the results of author attribution is outlined in [11]. The authors here have reported an elaborate study on the literary works of 9 notable authors of assamese community. The dataset they utilized was manually curated by the authors. Few other works were [12] which explores the usage of SVM for author identification through text-mining, achieving 60–80% accuracy with German newspaper texts, and showing robust performance even with topic variations, [13] proposes a fully automated author attribution approach that could secure a better accuracy than the lexical based method. They also stated that the combination of the traditional and the proposed method gave the maximum accuracy of 87% overall.

Language Models are probabilistic and work by analysing the grammatical content like characters, words, word sequences, POS tag sequences. They play a pivotal role in modern NLP applications. There are various types of language models like unigrams, bigrams, trigrams and another category based on neural network based. There were many significant works reported using this. A work in [14] highlights the study of authorship attribution using POS tags as features for language modelling. They used movie tweets and reviews as the corpus. They could achieve highest accuracy of 96% for movie reviews. Another work by [15]explores neural authorship attribution to identify the origin of AI-generated text. It compares proprietary and open-source LLMs. It analyses writing signatures using stylometric features to enhance classifier performance and address AI-generated misinformation risks. In [16], the authors have proposed a paragraph vector which overcomes the disadvantages of word order and context dependency issues of BoW models. Going through all these studies, the significance of language models in authorship attribution tasks is well understood.

Narrowing down our search to works of authorship attribution using ensemble learning, it was observed that here too majority of the work have been performed in English language datasets. And that very few works have been reported so far for other under-resourced languages. We include here some important work related to both the categories such as in [17], the authors propose an authorship attribution approach using ensemble learning, DistilBERT, and traditional machine learning techniques with count vectorizer and bi-gram TF-IDF features. Experiments on the "All the News" dataset show that the proposed methods outperform previous state-of-the-art approaches, achieving accuracy gains of up to 5.25% (ensemble) and 7.17% (DistilBERT) for 20 authors. The authors in [18] addresses the challenge of author attribution in Internet Relay Chat (IRC), a platform frequently used for cybercrime. It introduces a novel deep forest (DF) model, an ensemble-based method, for identifying authors of threat messages. The proposed approach, supported by autonomic IRC monitoring, demonstrates high accuracy even with a large number of candidates and limited training data. The paper in [19] proposes an ensemble approach for cross-domain authorship attribution by combining predictions from three classifiers using variable-length n-grams and multinomial logistic regression. The method outperforms baseline systems, demonstrating effectiveness on PAN-CLEF 2018 test data and a new corpus of English and Portuguese song lyrics. Meanwhile another work by [20]. Apart from few other relevant works, we failed to find out any significant work which reports the work of authorship attribution on low-resource languages. The cloud of low-resource languages was earlier limited to their community only. However, with the advent of digitization, it has become a mandatory work to build applications inclusive of these languages.

The table below summarizes the findings on authorship attribution task related to some low-resource languages.

Table 1: List of few low-resource languages and their findings w.r.t authorship attribution

Language	Findings
Hindi	A few papers on authorship attribution using machine learning methods, but limited studies using ensemble models specifically
Assamese	Limited work; Only one study on Assamese language documents for authorship attribution has been identified
Urdu	A study on hybrid ensemble model for authorship attribution in Urdu with reported accuracy of 92%
Kannada	One study on profile-based authorship attribution in Kannada with 88% accuracy using ensemble models

Journal of Information Systems Engineering and Management
2025, 10(40s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Odia	Minimal research available; some authorship attribution work using machine learning but not ensemble models specifically
Telegu	Few studies on authorship attribution; limited focus on ensemble models

OBJECTIVES

This paper highlights on the study of the performance of ensemble learning methodology on low-resource languages specifically Assamese, which is a language spoken by almost 15 million across the state of Assam as per census 2011 [21]. This constitutes 1.26 percent of the country's total population. Also recently in the year 2024, the Government of India conferred Assamese language as one of the classical languages in India [22].

METHODS

3.1 DATA LOADING AND PRE-PROCESSING

The manually curated dataset comprising of 9730 data points where each datapoint reflects a short story, excerpts of novels, articles written by 16 famous authors of the Assamese literary community. The dataset is constructed having two columns, a 'Text' field and a corresponding 'Author' field. The data from the 'Text' field is extracted as features and the one from the 'Author' field as labels. With the help of label encoding technique, the categorical author names are converted to numerical labels. The data is then split into training and testing considering a ratio of 80:20.

3.2 TOKENIZATION AND TEXT PROCESSING

Throughout our study, we had experimented with quite a number of tokenizers. But we could conclude with only "Indic Tokenizer" module from IndicNLP to handle Assamese text. The table below outlines the statistical data with respect to "No. of Text Files", No. of Sentences", "No. of Tokens", and "Average length of sentences".

3.3 MODEL TRAINING

Using TensorFlow and Keras models, an ensemble learning environment was crated consisting of three models namely RNN, CNN and LSTM. Text sequences were padded and tokenized before feeding them into the models. The final prediction is obtained by aggregating predictions from individual models, selecting the one with the highest probability. The model architecture can be found in the fig 1.

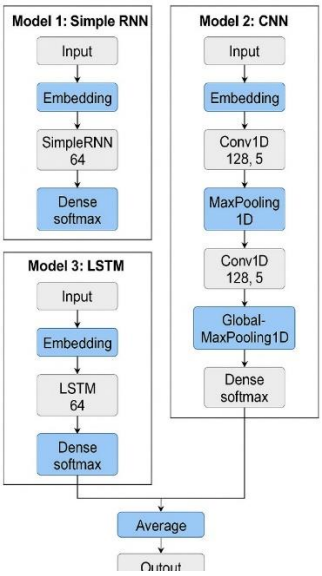

Fig.1: *Ensemble Model Architecture*

Journal of Information Systems Engineering and Management
2025, 10(40s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

3.4 EVALUATION

The ensemble model's performance is evaluated using metrics such as accuracy and a classification report and is listed in Table 2. The ensemble outperforms individual models by utilising diverse predictions.

Table 2: *Statistical data related to the data collected and analysed*

Author Name/Class Label	No. of Text Files	No. of Sentences	No. of Tokens	Average Length of Sentences
BK	189	4091	58699	11
DK	503	8182	110158	11
HD	37	555	8428	13
KH	524	6804	346722	11
LNB	823	9065	460945	8
LB	574	10579	199052	12
NB	567	4091	62477	8
ND	639	5991	183632	8
RB	594	6970	118206	10
BS	892	17598	364992	16
LG	784	14836	249470	13
ACH	774	21870	351370	13
SAM	755	25183	357900	11
HB	722	25595	459593	14
MRG	700	14424	195051	11
BNS	653	23125	333568	11
Total	9730	55,744	38,60,263	

RESULTS AND DISCUSSION

The following table 3 lists the performance metric scores for the ensemble learning implemented. The model shows high consistent performance across most classes, with an average AUC score of 0.9199. This indicates that the

Journal of Information Systems Engineering and Management
2025, 10(40s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

ensemble model effectively distinguishes between the majority of classes. The highest AUC score of 0.9618 is observed for Class 15, which showcases excellent classification accuracy. The lowest score of 0.4984 for Class 5 indicates performance close to random guessing. From this factor we can say that the model struggled in distinguishing the content of this class with the rest of the classes. The range of AUC scores is 0.4634, and the median AUC is 0.93185, further emphasizing the model's strong overall performance. Primarily, three classes (11, 13, and 15) demonstrate exceptional discrimination with AUC scores exceeding 0.95, while most other classes (0, 2, 3, 4, 8, 10, 12, and 14) fall in the good performance category with AUC values between 0.90 and 0.95. Classes 6, 7, and 9 show moderate performance with AUC scores ranging from 0.85 to 0.90. The classes 1 and 5 gives very poor AUC score which indicates that additional techniques hyper-parameter tuning and data augmentation techniques will have to implemented to improve its performance.

Table 3: *Epoch-wise performance metric values*

Model	Epoch	Accuracy	F1-Score	Precision	Recall
RNN +CNN +LSTM	10	84.38	77.0	77.0	78.0
	50	84.33	79.0	84.0	78.0

The Overall AUC-ROC Score (Macro-Average) is 0.8835. The class wise AUC score is further listed in the following Table 4.

Table 4: *Class-wise AUC Scores*

Class	AUC Score
0	AUC = 0.9131
1	AUC = 0.7128
2	AUC = 0.9315
3	AUC = 0.9382
4	AUC = 0.9377
5	AUC = 0.4984
6	AUC = 0.8750
7	AUC = 0.8596
8	AUC = 0.9406
9	AUC = 0.8824
10	AUC = 0.9183
11	AUC = 0.9503
12	AUC = 0.9322
13	AUC = 0.9527
14	AUC = 0.9323
15	AUC = 0.9618

Journal of Information Systems Engineering and Management
2025, 10(40s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

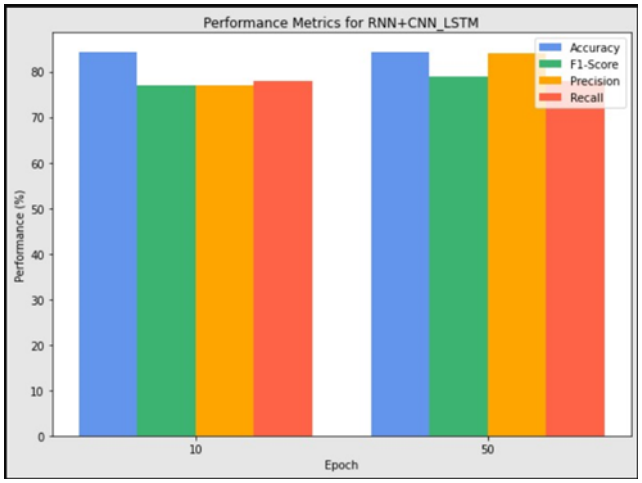

Fig.2: *Epoch vs Performance graph across all the metrics*

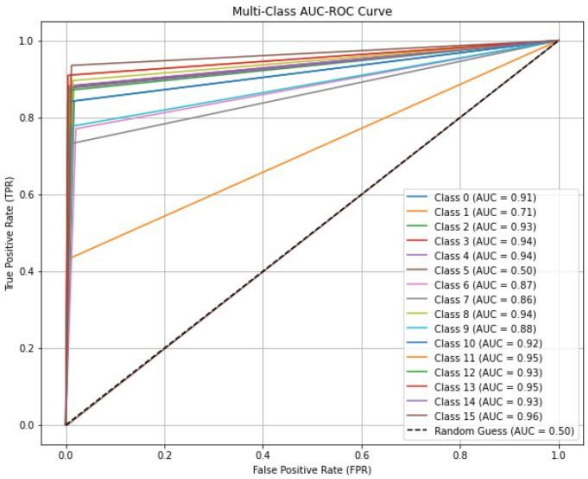

Fig.3: *AUC-ROC Curve across all the classes*

ACKNOWLEDGEMENT

The authors would like to thank the NLP Laboratory of the Department of IT, Gauhati University for extending their technical Support behind the experiments conducted related to this paper.

REFRENCES

[1] R. Bandura and E. I. M. Leal, "The Digital Literacy Imperative," Jul. 2022, Accessed: Feb. 24, 2025. [Online]. Available: https://www.csis.org/analysis/digital-literacy-imperative

[2] "CBWE." Accessed: Feb. 24, 2025. [Online]. Available: https://dtnbwed.cbwe.gov.in/

[3] "Understanding Digital Plagiarism," Research Experts. Accessed: Feb. 24, 2025. [Online]. Available: https://www.researchexperts.in/understanding-digital-plagiarism/

[4] "Google Scholar." Accessed: Mar. 12, 2025. [Online]. Available: https://scholar.google.com/schhp?hl=en&as_sdt=0,5

[5] X. He, A. H. Lashkari, N. Vombatkere, and D. P. Sharma, "Authorship Attribution Methods, Challenges, and Future Research Directions: A Comprehensive Survey," Information, vol. 15, no. 3, Art. no. 3, Mar. 2024, doi: 10.3390/info15030131.

[6] "Federalist Revisited: New Directions in Authorship Attribution | Digital Scholarship in the Humanities | Oxford Academic." Accessed: Mar. 13, 2025. [Online]. Available: https://academic.oup.com/dsh/article-abstract/10/2/111/956290

[7] M. Jin and M. Jiang, "Text clustering on authorship attribution based on the features of punctuations usage," in 2012 IEEE 11th International Conference on Signal Processing, Oct. 2012, pp. 2175–2178. doi: 10.1109/ICoSP.2012.6492012.
[8] H. Craig, "Authorial attribution and computational stylistics: if you can tell authors apart, have you learned anything about them?," Literary and Linguistic Computing, vol. 14, no. 1, pp. 103–113, Apr. 1999, doi: 10.1093/llc/14.1.103.
[9] "Subsets and Homogeneity: Authorship Attribution in the Scriptories Historiae Augustae | Digital Scholarship in the Humanities | Oxford Academic." Accessed: Mar. 13, 2025. [Online]. Available: https://academic.oup.com/dsh/article-abstract/13/3/133/933280
[10] E. Stamatatos, N. Fakotakis, and G. Kokkinakis, "Automatic Text Categorization in Terms of Genre and Author," Computational Linguistics, vol. 26, no. 4, pp. 471–495, Dec. 2000, doi: 10.1162/089120100750105920.
[11] S. P. Medhi, "Multi-Model Analysis on Author Attribution Detection on Assamese Text," International Journal of Intelligent Systems and Applications in Engineering, vol. 12, no. 4, Art. no. 4, Jun. 2024.
[12] J. Diederich, J. Kindermann, E. Leopold, and G. Paass, "Authorship Attribution with Support Vector Machines," Applied Intelligence, vol. 19, no. 1, pp. 109–123, Jul. 2003, doi: 10.1023/A:1023824908771.
[13] E. Stamatatos, N. Fakotakis, and G. Kokkinakis, "Computer-Based Authorship Attribution Without Lexical Measures," Computers and the Humanities, vol. 35, no. 2, pp. 193–214, May 2001, doi: 10.1023/A:1002681919510.
[14] O. Fourkioti, S. Symeonidis, and A. Arampatzis, "Language models and fusion for authorship attribution," Information Processing & Management, vol. 56, no. 6, p. 102061, Nov. 2019, doi: 10.1016/j.ipm.2019.102061.
[15] T. Kumarage and H. Liu, "Neural Authorship Attribution: Stylometric Analysis on Large Language Models," in 2023 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), Nov. 2023, pp. 51–54. doi: 10.1109/CyberC58899.2023.00019.
[16] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," Sep. 06, 2013, arXiv: arXiv:1301.3781. doi: 10.48550/arXiv.1301.3781.
[17] A. Abbasi, A. R. Javed, F. Iqbal, Z. Jalil, T. R. Gadekallu, and N. Kryvinska, "Authorship identification using ensemble learning," Sci Rep, vol. 12, no. 1, p. 9537, Jun. 2022, doi: 10.1038/s41598-022-13690-4.
[18] S. Shao, C. Tunc, A. Al-Shawi, and S. Hariri, "An Ensemble of Ensembles Approach to Author Attribution for Internet Relay Chat Forensics," ACM Trans. Manage. Inf. Syst., vol. 11, no. 4, p. 24:1-24:25, Oct. 2020, doi: 10.1145/3409455.
[19] J. E. Custódio and I. Paraboni, "An Ensemble Approach to Cross-Domain Authorship Attribution," in Experimental IR Meets Multilinguality, Multimodality, and Interaction, F. Crestani, M. Braschler, J. Savoy, A. Rauber, H. Müller, D. E. Losada, G. Heinatz Bürki, L. Cappellato, and N. Ferro, Eds., Cham: Springer International Publishing, 2019, pp. 201–212. doi: 10.1007/978-3-030-28577-7_17.
[20] M. Al-Sarem, F. Saeed, A. Alsaeedi, W. Boulila, and T. Al-Hadhrami, "Ensemble Methods for Instance-Based Arabic Language Authorship Attribution," IEEE Access, vol. 8, pp. 17331–17345, 2020, doi: 10.1109/ACCESS.2020.2964952.
[21] "India - Census of India 2011 - LANGUAGE ATLAS - INDIA." Accessed: May 03, 2024. [Online]. Available: https://censusindia.gov.in/nada/index.php/catalog/42561
[22] "'Classical language' status for Assamese," The Times of India, Oct. 04, 2024. Accessed: Mar. 28, 2025. [Online]. Available: https://timesofindia.indiatimes.com/city/guwahati/assamese-language-granted-classical-status-by-union-cabinet/articleshow/113920542.cms