An Empirical Framework to Identify Authorship from Bengali Literary Works

Chapter · July 2020

DOI: 10.1007/978-3-030-52856-0_37

3 authors:

Sumnoon Ibn Ahmad

1 PUBLICATION **0** CITATIONS

SEE PROFILE

Lamia Alam

Chittagong University of Engineering & Technology

14 PUBLICATIONS **39** CITATIONS

SEE PROFILE

Moshiul Hoque

Chittagong University of Engineering & Technology

72 PUBLICATIONS **210** CITATIONS

SEE PROFILE

An Empirical Framework to Identify Authorship from Bengali Literary Works

Sumnoon Ibn Ahmad, Lamia Alam, and Mohammed Moshiul Hoque ✉

Department of Computer Science and Engineering (CSE)
Chittagong University of Engineering & Technology (CUET)
Chattogram-4349, Bangladesh
{sumnoon52, lamiacse09, mmoshiulh}@gmail.com

Abstract. Authorship attribution is the process of identifying the probable author of an unknown document. This paper proposes a neural network based framework, which identifies the authorship from Bengali literary documents. For this purpose, a corpus consisting of 12,142 text documents of 23 writers/bloggers is built. A static dictionary is used to count vectorization and important features are selected using information gain. The proposed system is trained with 9099 documents and tested with 3043 documents. The experimental result shows that neural network with n-gram and parts of speech (PoS) features achieved 94% accuracy on developed corpus.

Keywords: Bangla language processing · Authorship attribution · Feature extraction · Machine learning.

1 Introduction

Due to the rapid growth in the use of internet and its effortless access via digital devices a substantial contents are uploaded enormously and quickly on the web as digital form. Also increasing popularity of text digitization and online documentation has made it very difficult to detect the authorship of a digital text. Therefore, automatic authorship detection or attribution has gained much attention in recent years to identify the original author from a huge amount of digital contents. Authorship detection is conducted mainly to verify authorship of a particular text. It is conducted by comparing works of other authors with the author in question. There are many application of authorship attribution such as plagiarism detection, resolving ownership dispute of unknown text, forensic linguistics etc. [14, 18, 6].

Authorship detection is one of applied field of Natural Language Processing (NLP), which utilizes the stylometric approach to determine authorship of unknown text. The stylometric approach refers to the statistical approach to differentiate writing styles of different authors [9]. Most of the authors tend to follow unique behavior in their text whether it is the use of certain word or collection of words or sometimes it maybe certain style of writing. Stylometry helps in determining these behaviors of the author. In order to do that, multiple texts with

known authors are used to extract stylometric features and using these features the text of unknown author is compared with text of known authors.
Although a substantial amount of works have been conducted on authorship attribution in English and other European languages, no remarkable work has been done yet on authorship attribution for text written in Bengali language. The major barrier of performing research on authorship attribution in Bengali due to the lack of linguistic resources in digital form and inadequate corpora. There are many well-known writers in Bengali literature and important properties can be discovered from their writing variations. These properties can be useful for literary, history, social and cultural studies respectively.
An author usually follows an unique writing style or feature which may be utilized to identify authorship of a particular writing. Stylometry concerns the writing style and it investigates the writing to find the specific pattern or characteristics of that writer. The major contributions of this work is that, we proposed a neural network based authorship identification system for Bengali texts using feature extraction method to extract n-gram and parts of speech (PoS) features to improve accuracy. In order to train and test our system we developed a Bengali text corpora including 23 authored texts which contains about 12,142 texts files. Also, we evaluated the proposed framework against two other algorithms-Random-forest and Support Vector Machine (SVM) are implemented and tested on our developed dataset. The experimental finding reveals that, the proposed neural-network based framework with PoS features achieved the higher accuracy than other algorithms in detecting authorship.

2 Related Work

Automatic identification of authorship is a long studied research issue for well resourced languages like, English. However, it is in preliminary stage till now with respect to Bengali literature.A character-level CNN method was proposed in [15], which identifies authorship and achieved 96% accuracy for 6 authors and 69% accuracy for 14 authors respectively . Marouf et al. proposed a technique that used BanglaMusicStylo dataset and gained 86.29% accuracy on 1470 Bengali songs of Rabindranath Tagore and Kazi Nazrul Islam [17, 11]. A hierarchical classifier based method was developed to detect authorship of unknown text [7]. A neural network based approach was proposed, which achieved 85% accuracy for 5 writers in Bengali languages. They used word length, and Wh words as features with small dataset [12]. Islam et al. is used n-grams, conjunction, pronoun features to detect authorship of 10 authors, which gained 96% accuracy [13]. Hossain et al. used word frequency, modified word frequency, spelling of word features and gained 90.67% accuracy 6 Bangladeshi writers [10]. Chakraborty et al. investigated the ten-fold cross-validation and concluded that SVM is better than decision tree and neural network for small dataset [6]. Phani et al. [18] had devised a process with character bi-grams and tri-grams and word uni, bi and tri-grams. They have also used a corpus of three thousands text from three prominent Bengali authors. Instead of using literature of Bengali authors as cor-

pus, Das et al. [8] have used text from four Bengali blog writers. They have used different feature count, such as length of different word, sentence, number of parts of speech used in sentence and number of words used in a certain position of the sentence. Saha et al. used multi-layer perceptron to correctly attribute short text to their authors using a twitter dataset of four authors and 400 tweets for each author with accuracy of 96.44% [21].
Most of the works stated above had very small dataset and less variation in author categories or limited writing styles. In contrast to these, we developed a neural network based system for Bengali authorship attribution that is trained and tested with larger dataset.

3 Proposed Methodology

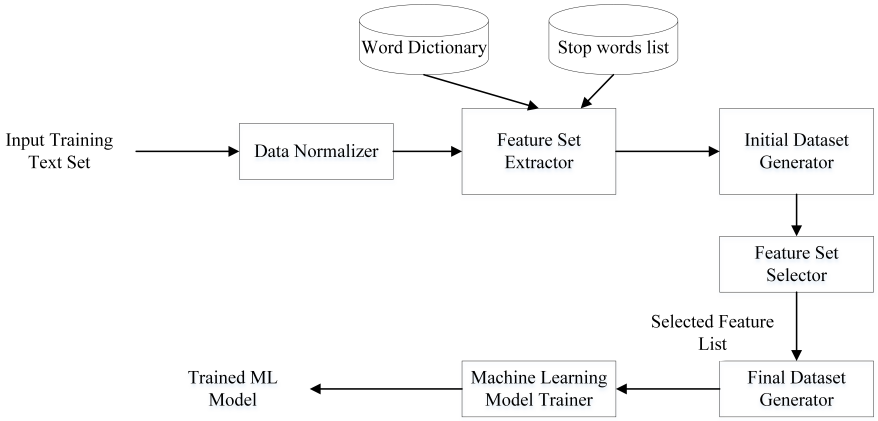
Proposed authorship detection system is divided into two phases- training phase and testing phase. At first, machine learning model has been trained using the training dataset and classification accuracy of the model is evaluated using the testing dataset in testing phase. Around 75% of prepared dataset is used in training and 25% is used in testing. As our primary dataset was raw and full of noises, we have to perform some data cleaning and remove noises. Then, the normalized data is used to extract features. After extracting the features, most useful features are selected using information gain (IG) value of the features. Then, final dataset was prepared and used to train classifier model. We used three classification algorithms and prepared four models. Neural network model was prepared in two ways, in one model we does not use parts of speech features and in other model we have used parts of speech features. Then, we have compared both model with our test set which was unknown to our models during training period. A schematic representation of our proposed authorship detection system is illustrated in Fig. 1.

3.1 Input

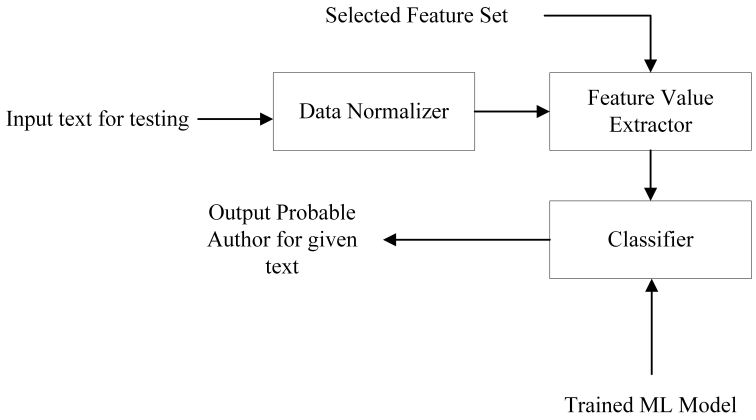
We have collected text from 23 writers which includes various writing styles. We have used hold-out method for training and testing our model as it is very good on large dataset and needs less computational power. For training, text of a certain writer was stored in a folder with his name and compressed for training set. For testing, a collection of text which was unknown to the model during training is used and the authorship detection was done with the help of previous knowledge and characterising the writing style of the text. Fig. 2 shows an sample of raw data.

3.2 Pre-processing of Raw Data

Raw data is not suitable for training purpose due to noises. Sometimes words from foreign languages are introduced into writings and these words help to detect authorship of particular text (e.g. literature of Kazi Nazrul Islam used Urdu



(a) Training Phase



(b) Testing Phase

Fig. 1: Proposed authorship detection system

আজ মহাবিশ্বে মহাজাগরণ, আজ মহামাতার মহা আনন্দের দিন, আজ মহামানবতার মধ্যযুগের মহা উদ্‌বোধন! আজ নারায়ণ আর ক্ষীরোদসাগরে নিদ্রিত নন। নরের মাঝে আজ তাঁহার অপূর্ব মুক্তি-কাঙাল বেশ। ওই শোনো, শৃঙ্খলিত নিপীড়িত বন্দিদের শৃঙ্খলের ঝনৎকার।

Fig. 2: Sample Text

words in his literature) or produces unwanted noise in other cases. Therefore, pre-processing is used to to reduce error rate. Each text can be divided into multiple sentences. The end or beginning of a sentence is determined with punc-

tuation. A collection of Bengali and English punctuation mark has been used to decompose the text into sentence. After the decomposition, the punctuation are removed as they don't have any significance. A dictionary of stop words are used to remove unrelated words from the text. A pre-processed text is shown in Fig 3 as an example.

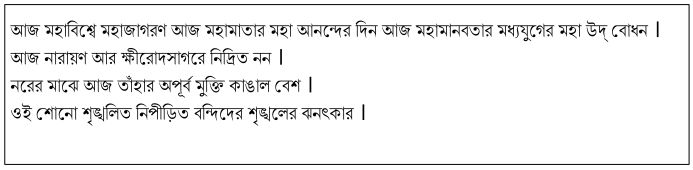

আজ মহাবিশ্বে মহাজাগরণ আজ মহামাতার মহা আনন্দের দিন আজ মহামানবতার মধ্যযুগের মহা উদ্ বোধন ।
আজ নারায়ণ আর ক্ষীরোদসাগরে নিদ্রিত নন ।
নরের মাঝে আজ তাঁহার অপূর্ব মুক্তি কাঙাল বেশ ।
ওই শোনো শৃঙ্খলিত নিপীড়িত বন্দিদের শৃঙ্খলের ঝনৎকার ।

Fig. 3: Pre-processed sample text

3.3 Feature Extraction

N-gram and PoS features are used to observe in the corpus. We have verified with uni-gram, bi-gram and tri-gram of word and found that more than tri-gram because increased grams do not give any significant information about authors. The training set is tokenized into uni-grams. Then we combined them to create word bi-gram and tri-gram. A PoS tagger is used to identify detects token from each sentence. This tagger takes text as input and detects each word from the text and assign them with relative parts of speech. A modified PoS tagger is used to tag other words outside of parts of speech. Due to lack of proper dynamic PoS tagger in Bengali, we had to create our own static PoS tagger which can detect nouns, pronouns, adjectives, verbs, adverbs and conjunctions. The Pos tagger utilizes a dictionary of words which is consist of more than 50 conjunctions, 30 pronouns, 23000 nouns, 1100 adjectives, 70000 verbs and 16000 adverbs. Conjunctions and Pronouns were collected from [5]. Frequency of each features is calculated from the text, which is used to find the important features. Fig 4 shows a set of sample features. With the help of word dictionary and n-gram ex-

সাথে	কোন	কিন্তু	না	করিয়া
0	1	12	28	1

Fig. 4: Sample text after feature extraction

tractor, a large number of feature words are found from the training data. These

features can be reduced by the information gain (IG). IG is used to determine how much information can be extracted using a feature and how important is the feature to contribute in overall prediction system. The information gain (IG) is calculated by Eq. 1.

$$IG(S,T) = E(S) - \sum_{t \in T} p(t) \times E(t) \tag{1}$$

where, $E(S)$ is defined as entropy which is the opposite of probability and it is directly related to the information gain in such a way that the more the entropy of an event the more information can be gained from that event. Entropy is calculated by Eq. 2.

$$E(S) = -\sum_{x \in S} p(x) \times \log_2 p(x) \tag{2}$$

3.4 Final dataset Generation

With the help of information gain calculation and stop word dictionary we have selected most important features from our primary dataset and removed unnecessary words and stop words from the text and prepared our final dataset. The final dataset is generated in .csv format.

3.5 Classifier

A neural network model is trained with developed dataset [22]. The propose neural network consist of three hidden layer with 128, 64 and 32 nodes in each layer. An activation function [20] is used to find the output from a node. In the proposed model, the rectified linear unit function a.k.a *ReLU* is used. Fig. 5 illustrates the neural network model. The proposed neural network model have used three hidden layer and one input and one output layer. In each hidden layer number of nodes or neurons were 128, 64 and 32 respectively. Each neurons, also known as perceptron[19], acts as a simple learner that takes one or multiple inputs and process them with a weight given on each of the input. Than it generates a binary decision. Using multiple similar neurons a layer of multi-layer perceptron is created. For weight optimization we have used Adam stochastic gradient-based optimization [16] and number of epoch was 3000.

The training procedure consists of three major steps:

- **Step 1: Forward Pass** In forward pass we run the sample vector from input layer to output layer through multiple hidden layers. The input value is multiplied with weight and a bias is added. Then the output is applied through a activation function in our case ReLU function. Suppose, w denotes the vector of weights, x is the vector of inputs, b is the bias and ϕ is the activation function, then for the i_th neuron the output y would be given by Eq. 3.

$$y = \sum_{i=1}^{n} w_i x_i + b = \phi(w^x + b) \tag{3}$$

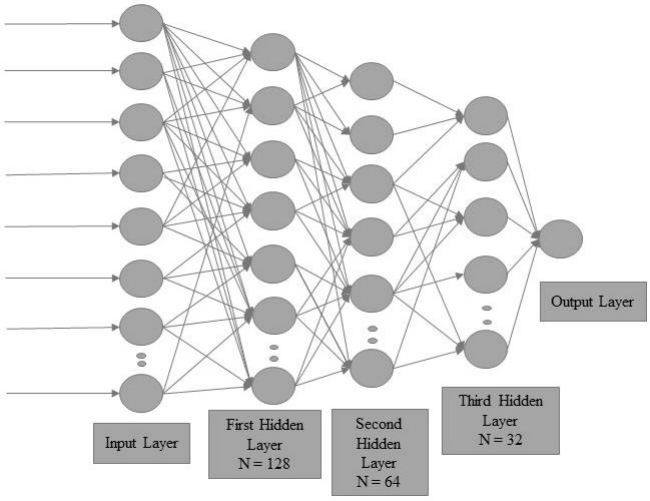


Fig. 5: Multilayer perceptron model having three hidden layers

Activation Function: It is used to determine the output of neural network like yes or no. Depending on the function the value results from 0 to -1 or -1 to 1 etc. In our system, we have used Rectified Linear Unit ($ReLU$) activation function, which is given by Eq. 4.

$$\phi(x) = max(0, x) \tag{4}$$

It is one of the most simplest and popular activation function. The biggest advantage of $ReLU$ is the non-saturation of its gradient, which improves the acceleration of Adam stochastic optimization more than any other activation function.

– **Step 2: Calculation of loss function** After the forward pass we would get some output from our model which refers as predicted output. Using predicted output and real output we calculate loss that we need to propagate using back-propagation algorithm. We have used cross-entropy as loss function in our system. The calculation of loss function is performed by calculating loss for each label separately and then summing the result [Eq. 5].

$$loss = -\sum_{c=1}^{M} y_{o,c} \log(p_{0,c}) \tag{5}$$

where, M is the number of classes, y is the binary indicator (0 or 1) of classification and $p_{0,c}$ is the predicted probability observation (o) of class c

– **Step 3: Backward Pass** After calculation of loss function, we back-propagate the loss and update the model by using gradient. In this step, weights would

adjust according to the gradient flow in that direction. The process is repeated until the final error is minimum.

4 Experimental Results

We used corpus of 12,142 literary passages written in Bengali language. We chose 8 eminent Bengali writer and 15 famous bloggers. For collecting data, we have scraped online websites and blog sites using custom web scraper and saved them in doc file. The proposed neural network model is experimented in two types of datasets: with PoS features and without PoS features. As writings of literature writers is not available in proper format we had to collect them from books, online portals [3, 1]. Moreover, some texts are collected manually. The data in later stages was converted to .txt files and stored in folder of the respective author. In order to collect data from bloggers, we scraped writings of numerous bloggers from [2], [4]. Then some texts were left out due to lack of information and volumes of text. Also, we have collected data from [5] which have a good collection of writings from various bloggers. Table 1 represents the summary of dataset.

Table 1: Data Summary

Number of documents	12142
Number of sentences(approx.)	607050
Number of words(approx.)	1214100
Total unique words(approx.)	29000

In order to classify the texts, we have to feed our collected documents to our classifier model. Table 2 shows the summary of dataset used for our classification process.

Table 2: Data Summary for train and test phase

	Training	Testing
Number of class	23	23
Number of documents	9099	3043
Average word per documents	50	52

4.1 Evaluation Measures

Confusion matrix is used to evaluate our model against test data. Confusion matrix of proposed approach with parts of speech feature is shown in Fig. 6.

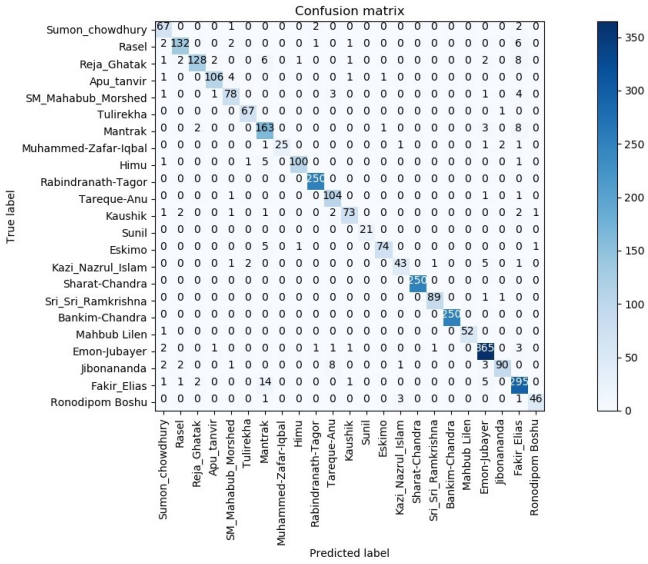


Fig. 6: Confusion matrix of proposed approach with PoS features

Fig. 6 shows that 250 texts of Rabindranath Tagor, 250 text of Sarat Chandra and 250 text of Bankim Chandra are detected correctly. Precision, recall, F_1 score and accuracy measures are used as per Eq. 6 - Eq. 9 respectively.

$$\texttt{Precision} = \frac{TP}{TP + FP} \tag{6}$$

$$\texttt{Recall} = \frac{TP}{TP + FN} \tag{7}$$

$$F_1 = \frac{2 * precision * recall}{precision + recall} \tag{8}$$

$$\texttt{Accuracy} \ = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

Here, TP, TN, FP and FN stands for true positive, true negative, false positive and false negative respectively. In Table 3 shows the precision, recall and F_1 and accuracy of different classification algorithms used based on our dataset.

Table 3: Comparison of Results

	Precision	Recall	F_1	Accuracy
Neural Network (without PoS feature)	0.92	0.90	0.91	91.62
Neural Network (with PoS feature)	0.94	0.94	0.94	94.25

Sample Input	Real Author	Probable Author
তৃতীয় পরিচ্ছেদ বিজয় — বদ্ধজীবের মনের কি অবস্থা হলে মুক্তি হতে পারে? শ্রীরামকৃষ্ণ — ঈশ্বরের কৃপায় তীব্র বৈরাগ্য হলে, এই কামিনী-কাঞ্চনে আসক্তি থেকে নিস্তার হতে পারে। তীব্র বৈরাগ্য কাকে বলে? হচ্ছে হবে, ঈশ্বরের নাম করা যাক — এ-সব মন্দ বৈরাগ্য। যার তীব্র বৈরাগ্য, তার প্রাণ ভগবানের জন্য ব্যাকুল; মার প্রাণ যেমন পেটের ছেলের জন্য ব্যাকুল। যার তীব্র বৈরাগ্য, সে ভগবান ভিন্ন আর কিছু চায় না। সংসারকে গাঙকুয়া দেখে; তার মনে হয়, বুঝি ডুবে গেলুম। আত্মীয়দের কাল সাপ দেখে, তাদের কাছ থেকে পালাতে ইচ্ছা হয়; আর পালায়ও। বাড়ির বন্দোবস্ত করি, তারপর ঈশ্বরচিন্তা করব — এ-কথা ভাবেই না। ভিতরে খুব রোখ।.....	Sri Sri Ramakrishna	Jibonananda
অষ্টম পরিচ্ছেদ সকলে আনন্দ করিতেছেন। ঠাকুর কেশবকে বলিতেছেন, "তুমি প্রকৃতি দেখে শিষ্য কর না, তাই এইরূপ ভেঙে ভেঙে যায়। "মানুষগুলি দেখতে সব একরকম, কিন্তু ভিন্ন প্রকৃতি। কারু ভিতর সত্ত্বগুণ বেশি, কারু রজোগুণ বেশি, কারু তমোগুণ। পুলিগুলি দেখতে সব একরকম। কিন্তু কারু ভিতর ক্ষীরের পোর, কারু ভিতর নারিকেলের ঘুই, কারু ভিতর কলায়ের পোর। (সকলের হাস্য) "আমার কি ভাব জানো? আমি খাই-দাই থাকি, আর সব মা জানে। আমার ভিন কথাতে গায়ে কাঁটা বেঁধে। গুরু, কর্তা আর বাবা।	Sri Sri Ramakrishna	Sri Sri Ramakrishna

Fig. 7: Sample input-output

4.2 Sample Input and Output

Fig. 7 shows the sample input and corresponding output as examples. First example indicates the incorrect prediction of author and second shows the correct prediction of author. The reason behind the incorrect prediction is that certain text of the author Sri Sri Ramakrishna and the author Jibonanada are almost similar and frequency of PoS features are also similar.

4.3 Comparison with Existing Techniques

In order to measure the effectiveness, we compare the proposed method with the available techniques. Table 4 shows the summary of the comparison.

Table 4 reveals that the proposed system has performed very well compared to other systems. The previous approaches used their own dataset. A recent method proposed by Khatun et al. achieved the higher accuracy (96%) than others [15]. However, they used only 6600 text documents written by 6 authors. Another method [13] also found the 96% accuracy for 10 authors with very small text documents (only 3125). Accuracy may vary due to the writing styles. Therefore, accuracy may comes naturally higher for small dataset and limited

Table 4: Comparison with previous approaches

	Total Authors	No. of documents	Accuracy (%)
Khatun et al. [15]	6	6600	96
Chowdhury et al. [7]	6	2,400	92.9
Islam et al. [12]	5	1,973	85
Islam et al. [13]	10	3,125	96
Hossain et al. [10]	6	2,764	90.5
Proposed System	23	12,142	94.12

number of authors due to less variation of writing styles. The proposed system considered the larger number of text documents (12,142) and authors (23) than the existing approaches. The system achieved a reasonably good accuracy which amount to 94% in terms of number of documents and authors.

5 Conclusion

This paper introduced a neural network based approach for identifying authorship from Bengali literary or blog texts. The proposed system can identify authorship of 23 authors in Bengali literature. To build the framework a self-developed dataset is used for training and testing with 12,142 text documents. The neural network approach with n-gram and parts of speech features provided the better accuracy than the existing techniques. The proposed system is not tested with standard dataset and not validate with the standard technique which are the main limitations of the system. The accuracy may be improved with more label data. K-fold cross validation technique may be used for training phases for better training accuracy. These are left as future issues.

References

1. Ebanglalibraray, https://www.ebanglalibrary.com
2. Sachalayatan, https://en.sachalayatan.com
3. Society for natural language technology research, https://nltr.org/index.php
4. Somewhere in blog, https://www.somewhereinblog.net
5. Stylogenetics, https://github.com/olee12/Stylogenetics
6. Chakraborty, T.: Authorship identification in bengali literature: a comparative analysis. CoRR **abs/1208.6268** (2012), http://arxiv.org/abs/1208.6268
7. Chowdhury, H.A., Imon, M.A.H., Islam, M.S.: Authorship attribution in bengali literature using fasttext's hierarchical classifier. In: 2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT). pp. 102–106. IEEE (2018)
8. Das, P., Tasmim, R., Ismail, S.: An experimental study of stylometry in bangla literature. In: 2015 2nd International Conference on Electrical Information and Communication Technologies (EICT). pp. 575–580. IEEE (2015)

9. Holmes, D.I.: The evolution of stylometry in humanities scholarship. Literary and linguistic computing **13**(3), 111–117 (1998)
10. Hossain, M.T., Rahman, M.M., Ismail, S., Islam, M.S.: A stylometric analysis on bengali literature for authorship attribution. In: 2017 20th International Conference of Computer and Information Technology (ICCIT). pp. 1–5. IEEE (2017)
11. Hossain, R., Al Marouf, A.: Banglamusicstylo: A stylometric dataset of bangla music lyrics. In: 2018 International Conference on Bangla Speech and Language Processing (ICBSLP). pp. 1–5 (2018)
12. Islam, M.A., Kabir, M.M., Islam, M.S., Tasnim, A.: Authorship attribution on bengali literature using stylometric features and neural network. In: 2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT). pp. 360–363. IEEE (2018)
13. Islam, N., Hoque, M.M., Hossain, M.R.: Automatic authorship detection from bengali text using stylometric approach. In: 2017 20th International Conference of Computer and Information Technology (ICCIT). pp. 1–6. IEEE (2017)
14. Juola, P.: Rowling and galbraith: an authorial analysis. Language Blog (2013)
15. Khatun, A., Rahman, A., Islam, M.S., Marium-E-Jannat: Authorship attribution in bangla literature using character-level cnn. arXiv preprint arXiv:2001.05316 (2020)
16. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
17. Marouf, A., Hossain, R.: Lyricist identification using stylometric features utilizing banglamusicstylo dataset. In: 2nd International Conference on Bangla Speech and Language Processing (ICBSLP2019) (2019)
18. Phani, S., Lahiri, S., Biswas, A.: A supervised learning approach for authorship attribution of bengali literary texts. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP) **16**(4), 28 (2017)
19. Rosenblatt, F.: The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review **65**(6), 386 (1958)
20. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning internal representations by error propagation. Tech. rep., California Univ San Diego La Jolla Inst for Cognitive Science (1985)
21. Saha, N., Das, P., Saha, H.N.: Authorship attribution of short texts using multilayer perceptron. International Journal of Applied Pattern Recognition **5**(3), 251–259 (2018)
22. Wilson, E., Tufts, D.W.: Multilayer perceptron design algorithm. In: Proceedings of IEEE Workshop on Neural Networks for Signal Processing. pp. 61–68. IEEE (1994)