Automatic Authorship Detection from Bengali Text using Stylometric Approach

Nazmul Islam, Mohammed Moshiul Hoque and Mohammad Rajib Hossain
Chittagong University of Engineering and Technology
Chittagong, Bangladesh
e-mail: nazmulcuet11@gmail.com, moshiul_240@cuet.ac.bd, rajcsecuet@gmail.com

***Abstract*—Authorship detection is the process of predicting authorship of an unknown text. Every writer has a different style of writing of their own. Detecting authorship from text by analyzing writing style of an author is known as stylometry. In this paper, we propose a stylometric feature based approach for detecting authorship from Bengali texts. The system the classify authorship using n-grams, a feature ranking and selection system using information gain (IG). We used 3125 passages written by 10 Bengali authors for evaluating performance. The evaluation result shows that the propose system achieved 96% accuracy in authorship detection using random forest classifier and also reveal that n-gram features are very good discriminators among linguistic style of different Bengali authors.**

Keywords—natural language processing; authorship detection; stylometry; stylometric feature, normalization

I. INTRODUCTION

Authorship detection is a computational process that deals with the testimony of the author in a particular text [1]. This is one of the well-studied problems in the field of Natural Language Processing. The main goal of the authorship detection is to classify documents according to their authorship. There are several approaches of authorship detection problem. In one approach, the goal is to verify a given document has written by a particular author or not. In another approach is to identify the author of a particular given document from predefined set of authors. There are various application areas of authorship detection such as author recognition, plagiarism detection, author profiling, and detection of unlikeness in the writings of authors [1, 2].

Authorship detection belongs to the subtask of Stylometry detection where a resemblance between the predefined writers and the unknown articles has to be established with consideration of various stylistic features of the documents. Stylometry is the application of the study of linguistic style, usually to written language which concerns the writing style or behaviors rather than its contents [2]. It also investigates the writing and finds specific pattern or characteristics of that writing. A writer may use one or two consecutive words (bigrams) more frequently than others, or may have tendency to use specific phrase, specific tense, particular sentence structure or start and ends sentence with specific parts of speech. These are the properties/features that can be used detect the authorship of a writing. Recently, learning techniques are being used to infer attributes that discriminate the styles of authors. In order to identify a particular author's writing, most distinguishing and appropriate features should extract to represent the style of that author. In this context, the Stylometry offers a strong support to define a discriminative feature set [3].

Although authorship detection in English and any other language processing problem, has received a lot of attention since mid of nineteen century, there has been a very few work in Bengali language-one of the most widely spoken South Asian languages. Due to this lack of research progress in Bengali authorship detection could be attributed to shortage of adequate corpora and tools. There are many prominent and skillful writers in Bengali literatures and we often find interesting style in their writings. Analyzing their writing we can identify distinguishable features among them and outcomes may apply in literary studies, historical studies, social studies, gender studies, and many forensic cases.

In this paper, we investigated the writing behaviors of ten prominent Bengali writers. We considered writers such as Rabindranath Tagore, Sarat Chandra Chattopadhyay form golden age Bengali as well as blog writers from different blogs. We collected the blogs and writings in a random way to minimize biasness of data. Through our analysis we found some n-gram features are very useful to detect author while some are not. We use unigram, bigram, trigram and parts of speech features like conjunction and pronoun. In addition to that a set of stop word features also uses to eliminate unnecessary unigrams. Three machine learning algorithms are investigated for classification and evaluate relative performance among these algorithms.

II. RELATED WORK

Authorship detection is a long studied problem for the highly resourced languages like English. A general overview of the topic of authorship detection techniques proposed by different researchers [4, 5]. The instigating study on authorship attributes identification using word-length histograms appeared at the very end of 19th century [6]. Then a number of researches are carried based on content analysis [7], computational stylistic approach [8], exponential gradient learning algorithm [9], Winnow regularized algorithm [10], SVM based approach [11]. Some recent studies, including Bogdanova et al. [12] experimented with cross-language authorship attribution and Nasir et al. [13] framed authorship

978-1-5386-1150-0/17/$31.00 ©2017 IEEE

attribution as semi-supervised anomaly detection via multiple kernel learning.

Authorship detection in Bengali language is a relatively new problem. Very few research activities have been conducted so far in this topic. Among them, Chakraborty [14] used a 10-fold cross-validation on three classes and his results revealed that SVM surpasses decision tree and neural network classifiers. Das et al. proposed an authorship identification technique in Bengali literature using simple n-gram token counts [15]. They expostulated that simple unigram and bi-gram features along with vocabulary richness are better to discriminate amongst authors. However, this approach is restrictive when considering authors of the same period and same variety. Phani et al. represented a system for authorship detection using n-gram features in Bengali texts for three authors [16]. A stylometric study on Bengali literature was conducted by Das et al. [17]. They have analyzed features like unique words used by an author, word length, sentence length, number of parts of speech, number of question mark etc. They conducted analysis only for four different authors with blog texts and considered less variation on the linguistic styles of these authors.

III. PROPOSED METHODOLOGY

There are two major phases in our proposed authorship detection system: training phase and testing phase. A schematic representation of our proposed authorship detection system is illustrated in Fig. 1.

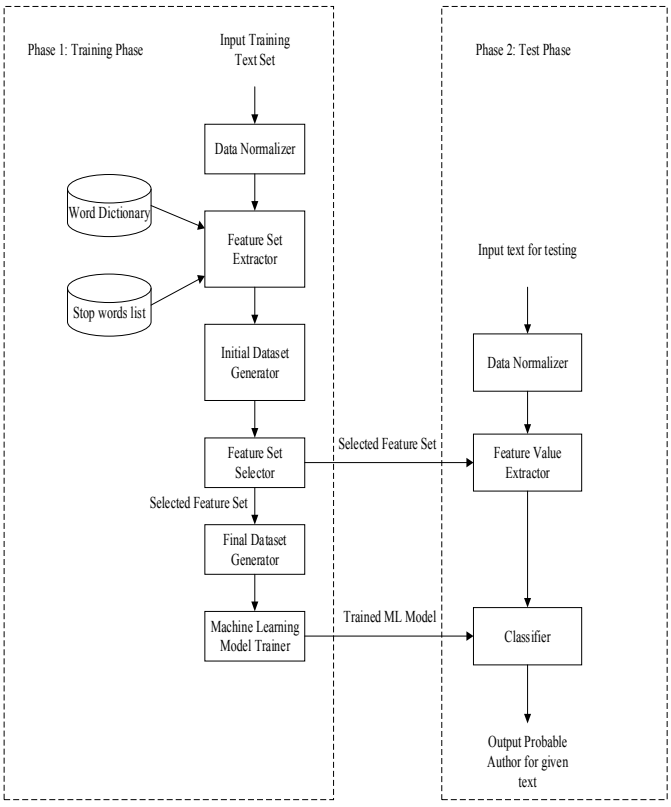

Fig. 1. Proposed authorship detection system

A. Input

Training phase needs a set of texts along with class name that is actual author name for each text as input. Therefore, a set of text file for an author is put together in directory and then the whole directory is compressed along with the author name and used as training set for an author. The unknown author's text will use as input to test module to predict the actual author for that unknown text. A simple text file of Rabindranath-Tagore is shown in Fig. 2 as an example.

মানভঞ্জন

প্রথম পরিচ্ছেদ

রমানাথ শীলের ত্রিতল অট্টালিকায় সর্ব্বোচ্চ তলের ঘরে গোপীনাথ শীলের স্ত্রী গিরিবালা বাস করে। শয়নকক্ষের দক্ষিণ দ্বারের সম্মুখে ফুলের টবে গুটিকতক বেলফুল এবং গোলাপফুলের গাছ; ছাতটি উচ্চ প্রাচীর দিয়া ঘেরা —— বহির্দৃশ্য দেখিবার জন্য প্রাচীরের মাঝে মাঝে একটি করিয়া ইট ফাঁক দেওয়া আছে।. . .

Fig. 2. Sample input text

B. Data Normalizer

Data normalizer takes a text as input. Decompose the whole text as a list of sentences. For each sentence remove extra white space character, punctuation marks. Generate a sequence of tokens. For every text rewrite the sentences in normalized form into another text file for further processing. After normalizing the processed file is shown in Fig. 3.

মানভঞ্জন প্রথম পরিচ্ছেদ রমানাথ শীলের ত্রিতল অট্টালিকায় সর্ব্বোচ্চ তলের ঘরে গোপীনাথ শীলের স্ত্রী গিরিবালা বাস করে।
শয়নকক্ষের দক্ষিণ দ্বারের সম্মুখে ফুলের টবে গুটিকতক বেলফুল এব গোলাপফুলের গাছ ছাতটি উচ্চ প্রাচীর দিয়া ঘেরা বহির্দৃশ্য দেখিবার জন্য প্রাচীরের মাঝে মাঝে একটি করিয়া ইট ফাঁক দেওয়া আছে।. . . .

Fig. 3. Normalized text from input text

C. Word Dictionary

Parts of speech include nouns, verbs, adverbs, adjectives, pronouns, conjunction and their sub-categories. Due to lack of good POS tagger for Bangla language we have used our own version of static POS tagger that can only tag pronouns and conjunction. We have used a dictionary consists of pronouns and conjunction form the database of Society for Natural Language Technology Research [18]. A sample word dictionary is shown in Table I.

TABLE I. SAMPLE WORD DICTIONARY

Word	POS
আর	CONJ
আমি	PRON
কেউ	PRON
কিংবা	CONJ
একে	PRON

D. Feature Set Extractor

Feature extractor generates all possible unigram, bigram, trigram, parts of speech and stop words features. This process uses an n-Gram generator for generating all possible n-Grams. We have only used unigram, bigram and trigram and exclude larger grams due to fewer contributions in information about the linguistic style of an author. Parts of speech features are generated by first tokenizing every sentences of a given text and then getting tags for every token. In the current implementation, we used a list of Bengali stop words provided by Society for Natural Language Technology Research [18]. A list of sample unigram, bigram, and trigram features is shown in Table II.

TABLE II. SAMPLE FEATURES OF UNIGRAM, BIGRAM AND TRIGRAM

Feature types	Sample features
Unigram	সাথে, কোন, কিন্তু, না, করিয়া, হইয়া, মত, আমি, হইতে, একটা, করিতে, আমার, তাহার, নিয়ে, হয়, কোনো, কী, তাঁহার, শুধু, ...
Bigram	সেই সাথে, না না, না কিন্তু, হইল না, ছিল না, হইয়া গেল, এই বলিয়া, করিতে লাগিল, হইতে লাগিল, হয় নাই , হইয়া উঠিল, মনে হয়, ...
Trigram	বাহির হইয়া গেল, করিতে পারে না, চুপ করিয়া রহিল, মনে হইতে লাগিল, সে যাই হোক, আসিয়া উপস্থিত হইল, বলিতে পারি না, ...

E. Feature Set Selector

Feature extractor generates a huge set of n-Gram based features. Feature selector selects important features form the extracted feature set and generates a comparatively reduced size feature set. To select important feature set we used two methods: (i) calculate Information Gain (IG) for every feature and discard features for which information gain is very low in our case we considered 0.05 and (ii) generate correlation matrix and discard correlated features to reduce biasness.

Entropy is a measure of unpredictability of the state, or equivalently, of its average information content. It represents how 'mixed up' an attribute is. Higher entropy denotes the higher rate of predictability of a feature. Entropy and information gain can be calculated using the Eqs. (1) and (2) respectively.

$$E(S) = -\sum_{x \in S} p(x) \times \log_2 (p(x)) \quad (1)$$

$$IG(S,T) = E(S) - \sum_{t \in T} p(t) \times E(t) \quad (2)$$

Feature ranking based on information gain of first 20 unigram features is shown in Table III.

TABLE III. SAMPLE WORD DICTIONARY

Rank	Words	IG
1	সাথে	0.598
2	কোন	0.572
3	কিন্তু	0.482
4	না	0.463
5	করিয়া	0.418
6	হইয়া	0.368
7	মত	0.353
8	আমি	0.347
9	হইতে	0.340
10	একটা	0.340
11	করিতে	0.339
12	আমার	0.334
13	তাহার	0.326
14	নিয়ে	0.325
15	হয়	0.325
16	কোন	0.320
17	কি	0.315
18	তাঁহার	0.314
19	শুধু	0.309
20	করে	0.309

F. Correlation Matrix

In statistics, dependence or association is any statistical relationship, whether causal or not, between two random variables or two sets of data. If there is a correlation between two features that denotes that one feature is strongly dependent on other. So if we use both the feature at a time for training there will be bias. We have plotted the correlation matrix as heat map for better representation and understanding of correlation (as shown in Fig. 4). Correlation between two different features X and Y is given by Pearson correlation coefficient matrix as in Eq. (3).

$$corr(X,Y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}} \quad (3)$$

Where $\bar{x}$ and $\bar{y}$ are the sample means of X and Y respectively.

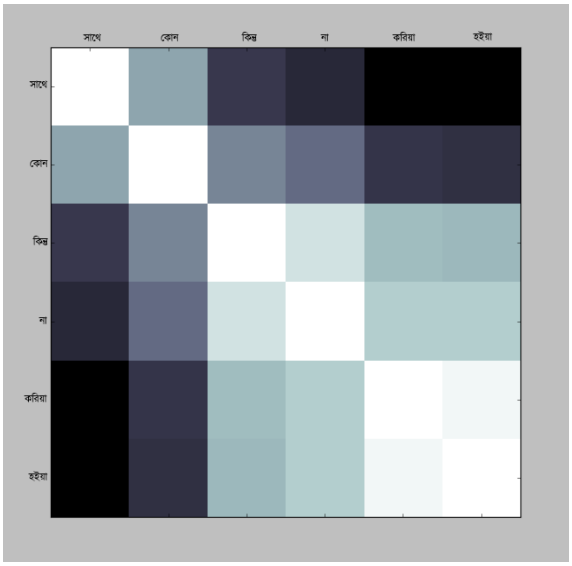

Fig. 4. A sample correlation heat map for 6 features

G. Final Dataset Generator

Incorrect or inconsistent data can lead to false conclusions. After filtering out correlated column in sample dataset, a sample of our cleaned dataset is shown in Fig. 5.

	A	B	C	D	E	F	G
1		সাথে	কোন	কিন্তু	না	করিয়া	Author Name
2	Text1	0	0	29	91	81	Rabindranth-Tagore
3	Text2	0	28	75	180	89	Sharat-Chandra
4	Text3	0	4	4	43	15	Bankim-Chandra
5	Text4	10	1	11	28	0	Muhammed-Zafar-Iqbal
6	Text5	7	0	10	31	0	Muhammed-Zafar-Iqbal
7	Text6	0	0	7	17	0	Anisul-Haque
8	Text7	0	0	4	32	0	Emon-Jubayer
9	Text8	2	2	2	8	0	Tareq-Anu
10	Text9	11	7	8	41	0	Rabindranth-Tagore
11	Text10	11	7	8	41	0	Hassan-Mahbub
12	Text11	0	2	3	9	0	Nir-Sondhani
13	Text12	31	36	25	43	0	Kandari-Athorbo
14	Text13	0	30	82	214	110	Sharat-Chandra
15	Text14	0	3	5	37	16	Bankim-Chandra
16	Text15	1	0	4	12	0	Anisul-Haque
17	Text16	35	44	22	52	0	Kandari-Athorbo
18	Text17	0	0	2	35	0	Emon-Jubayer
19	Text18	2	1	2	9	0	Tareq-Anu
20	Text19	10	5	8	47	0	Hassan-Mahbub
21	Text20	1	4	4	11	0	Nir-Sondhani

Fig. 5. Final dataset with reduced feature set

H. Feature Value Extractor

Feature value extracts the number of times each of the selected features occurs in a document. For our given input text 'সাথে' occurs 0 times, 'কোন' appears 0 times, 'কিন্তু' appears 27 times and so on. Now for the sample input text we have a feature vector like shown in Table IV.

TABLE IV. FEATURE VECTOR FOR GIVEN INPUT TEXT

সাথে	কোন	কিন্তু	না	করিয়া
0	0	27	112	79

I. Machine Learning Model Trainer

We used three classifiers (Naïve Bayes, Decision tree and Random forest) and compared their performance. We trained the above three classifier using final dataset, then for a new document *d* we calculated feature vectors for that document. These feature vectors are passed through the classifier to get a predicted author name. We will explain random forest classifier in below.

- **Random Forest Classifier**: This classifier grows many classification trees. To classify a new object from an input vector, put the input vector down each of the trees in the forest. Each tree gives a classification, and count tree 'votes' for that class. The forest chooses the classification having the most votes. Each tree is grown according to the following steps.

  Step1: If the number of documents in the training set is D, sample D documents at random. This sample will be the training set for growing the tree.

  Step2: If there are X input variables(i.e. features), a number $x << X$ is specified such that at each node, x variables are selected at random out of the X and the best split on these x is used to split the node. The value of x is held constant during the forest growing. For our case we have selected the value of x as $\sqrt{|x|}$.

  Step3: Each tree is grown to the largest extent possible. That is a node is divided until the node contains 1 sample or the gini-index of the node is 0.0 that is splitting is pure.

We have grown $\sqrt{|x|}$ decision tree with $\sqrt{|x|}$ random feature in each. For our sample dataset we have X = 5 features. Therefore, we have generated $ceil(\sqrt{|5|}) = 3$ different tree with three different feature set. Fig. 6 shows the Tree 1 with feature set {'**সাথে**', '**কোন**', '**কিন্তু**'}as an example.

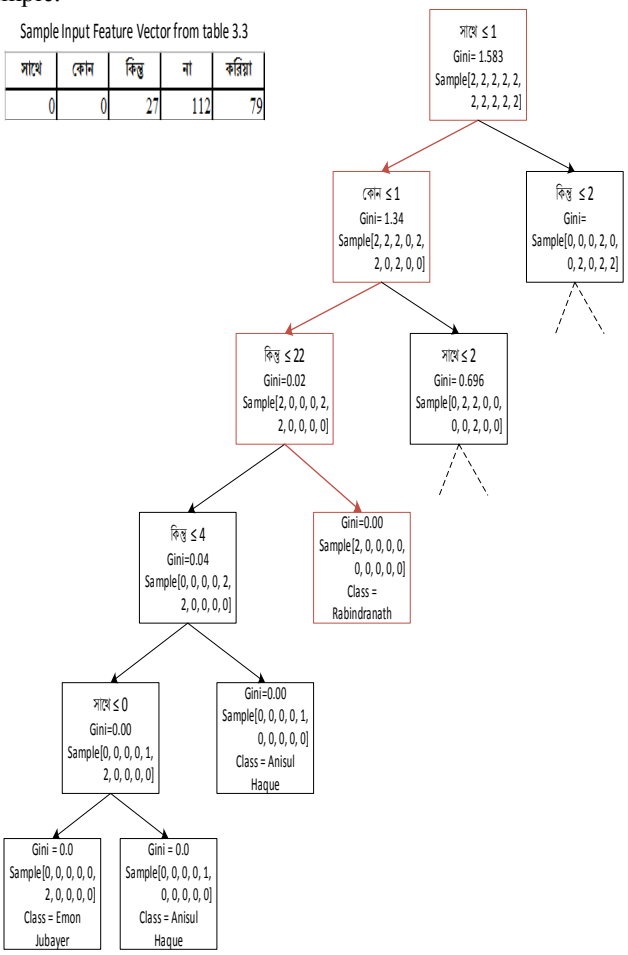

Fig. 6. Tree 1 of random forest model using feature set {'সাথে', 'কোন', 'কিন্তু'}.

From Tree 1, it is shown that the decision path leads to leaf labeled with class name Rabindranath. As in input vector '**সাথে**' = 0 which is less than 1 so go to the left child. Again as '**কোন**' = 0 <1 so we go to the left then '**কিন্তু**' = 27 > 22 so we go to right and finally reach a leaf labeled as 'Rabindranath'. Therefore, the predicted author using Tree 1 of random forest classifier is Rabindranath Tagore (RNT). The decision path is shown in red color.

IV. EXPERIMENTAL RESULTS

In our experiment, we have taken new corpus of 3125 literary passages texts of 10 different writers collected from books [18] and blogs [19, 20, 21, 22, 23]. We use 2032 texts (66.03%) for training and 1093 texts (34.97%) for testing.

We used texts written by of five eminent Bengali writers and five famous blog writers. The reason to choosing these specific writers is the availability and popularity of their writings.

A. Dataset Distribution

We have tried to collect at least 100 passages for each of the authors. In case of Anisul Haque (AH), we had to take 35 passages due to lack of online available literary passage written by him. A small fragment of data set distribution is illustrated in Table V.

TABLE. V. RAGMENT OF DATASET DISTRIBUTION

Author Name		Overall	Train Set	Testing Set
Rabindranath Tagore	Mean #Sentence	229.6147	228.4828	181.3333
	Total Files	109	76	33
Sarat Chandra	Mean #Sentence	305.1100	271.7821	329.8667
	Total Files	100	70	30
Bankim Chandra	Mean #Sentence	78.8034	78.9756	78.4000
	Total Files	117	82	35
Muhammed Zafar Iqbal	Mean #Sentence	104.2314	101.9111	99.4872
	Total Files	121	82	39
Anisul Haque	Mean #Sentence	25.3714	26.0882	16.1538
	Total Files	35	22	13
Emon Jubayer	Mean #Sentence	117.9200	144.9945	75.4867
	Total Files	1500	900	600
Tareque Anu	Mean #Sentence	56.7864	57.1605	55.4922
	Total Files	426	298	128
Hasan Mahbub	Mean #Sentence	148.4050	150.8000	29.3288
	Total Files	242	169	73
Nir Sondhani	Mean #Sentence	83.0844	83.9632	75.5970
	Total Files	225	158	67
Kandari Athorbo	Mean #Sentence	104.2320	103.2928	85.1067
	Total Files	250	175	75

B. Evaluation Measures

In order to evaluate the performance of the proposed system uses confusion matrix, precision, recall, F_1-score and accuracy measures. Confusion matrix also known as an error matrix is a specific table layout that allows visualization of the performance of an algorithm. Each column of the matrix represents the instances in a predicted class while each row represents the instances in an actual class. Confusion matrix for random forest classifier is shown in Fig. 7. Figure shows that out of 35 files of Bankim Chandra (BC) is all are detected as BC's, 36 files of Muhammed Zafar Iqbal (MZI) is detected as MZI's, 1 files of MZI is detected as AH's. One file of MZI is detected as Kandari Athorbo's (KA) and 1 as Emon Jubayer's (EJ) and so on. The precision, recall, F_1-score and accuracy are measured according to the Eqs. (4), (5), (6) and (7) respectively.

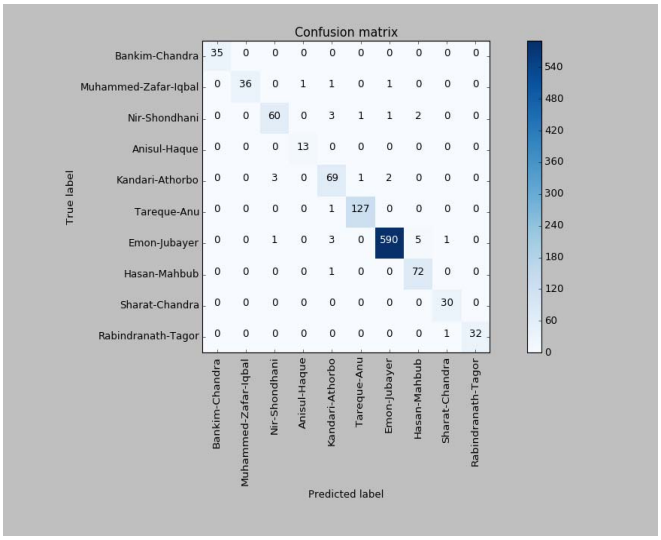

Fig. 7. Confusion matrix for random forest classifier

$$p = \frac{ture\ positive\ (TP)}{true\ positive\ (TP)\ +\ false\ positive\ (FP)} \tag{4}$$

$$r = \frac{ture\ positive\ (TP)}{true\ positive\ (TP)\ +\ false\ negative\ (FN)} \tag{5}$$

$$F_1 = 2 \times \frac{p \times r}{p + r} \tag{6}$$

$$accuracy = \frac{TP + \mathrm{TN}}{TP + TN + \mathrm{FP} + FN} \tag{7}$$

Table VI shows the performance measure of the Naïve Bayes, decision tree and random forest classifiers.

TABLE VI. PERFORMANCE MEASURES OF THREE CLASSIFIERS

Author Name	Naïve Bayes			Decision Tree			Random Forest		
	p	r	F_1-score	p	r	F_1-score	p	r	F_1-score
RNT	0.96	0.79	0.87	0.75	**1.00**	0.86	**1.00**	0.97	**0.98**
SCC	0.86	0.83	0.85	0.88	0.93	0.90	**0.94**	**1.00**	**0.97**
BC	0.92	**1.00**	0.96	0.97	0.97	0.97	**1.00**	**1.00**	**1.00**
MZI	0.97	0.90	0.93	0.93	**0.95**	0.94	**1.00**	0.92	**0.96**
AH	0.03	0.69	0.06	0.45	**1.00**	0.62	**0.93**	**1.00**	**0.96**
EJ	**0.99**	0.56	0.71	0.98	0.84	0.91	**0.99**	**0.98**	**0.99**
TA	0.60	0.92	0.72	0.85	0.91	0.88	**0.98**	**0.99**	**0.99**
HM	0.69	0.27	0.39	0.74	0.96	0.83	**0.91**	**0.99**	**0.95**
NS	0.92	0.88	0.89	0.79	0.85	0.82	**0.94**	**0.90**	**0.92**
KA	0.56	0.47	0.51	0.69	0.85	0.76	**0.88**	**0.92**	**0.90**
Avg/ Total	0.87	0.64	0.71	0.90	0.88	0.88	**0.97**	**0.97**	**0.97**

The overall accuracy of each three classifier is shown in Fig. 8. It reveals that the random forest classifier gives better accuracy (96%) than that of the other classifiers.

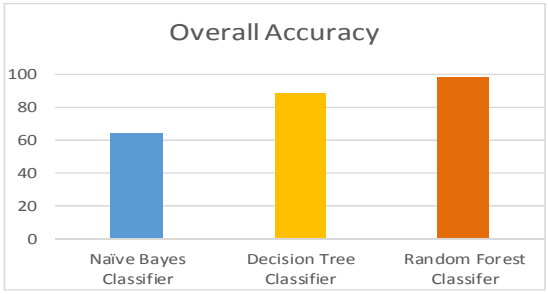

Fig. 8. Overall accuracy of each of the three classifer

C. *Learning Charecteristics*

Learning curve is graphical representation of prediction accuracy/error vs. the training set size. Learning curve for each of the three algorithms is shown in Fig. 9. From the curve we can see that learning rate of random forest model is almost always higher than decision tree and Naïve Bayes Model. Therefore, with the increase number of training samples the accuracy of the system will improve.

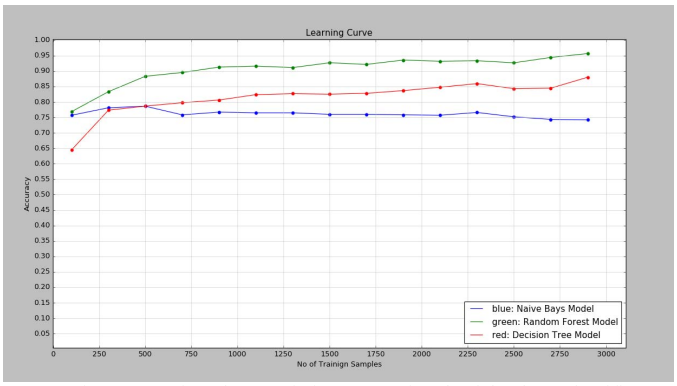

Fig. 9. Learning characteristics curve of each of the three classifier

Authorship detection is in rudimentary stage in Bangla language proceeding. In previous work [17], the authors analyzed 8 Stylometric features and worked with 4 different writers, 50 passages all collected from blogs. They have showed relative differences among writers based on each Stylometric feature. However, But they did not proposed any methodology for detecting author using Stylometric features. On the other hand, the proposed method deals with 10 writers including blog writers and can detect authorship from Bengali text literatures using n-Gram based Stylometric features.

V. CONCLUSION

Due to the abrupt growth of text in electronic form in webs, blogs, social media, forums etc. creates anonymously or under unverified names. Therefore, it is necessary to group texts written by the same author or track texts written under different names but belonging to the same person. Authorship detection developed by the computational analysis of texts attracts increasing attention because it may offer quick answers to these problems. This paper attempts to demonstrate the mechanism to recognize ten authors in Bengali literature based on their style of writing. We showed that n-gram based stylometric analysis along with parts of speech based feature gives higher accuracy. As part of our study we constructed a corpus of 3125 of literary passage of 10 prominent Bengali writers including blog writers. We performed classification experiments based on our generated dataset, and found accuracy of 96% using random forest classifier. We found that random forest classifier outperforms both Naïve Bayes and Decision tree classifier. We will extend our approach to other forms of text, such as news articles, tweets, and online forum threads in future. Structural testing feature and semantically correctness checking feature may be included in future for better performance.

References

[1] S. Phani, S. Lahiri, and A. Biswas, "A supervised learning approach for authorship attribution of Bengali literary texts", ACM Trans. Asian Low-Resour. Lang. Inf. Process., vol. 16, no. 4, August 2017.

[2] T. Chakraborty, "Authorship identification in Bengali literature: A comparative analysis", in Proc. of Computational Linguistics (COLING), pp. 41-50, 2012.

[3] D. J. Croft, Book of Mormon word prints reexamined, Sun Stone Publishers, 6, pp. 15-22, 1981.

[4] P. Juola, Authorship attribution, J. Found. & Trends in Inf. Retr., vol. 1, no. 3, pp. 233–334, Dec. 2006.

[5] M. Koppel, J. Schler, and S. Argamon, Computational methods in authorship attribution, J. Am. Soc. Inf. Sci. Technol., vol. 60, no. 1, pp. 9-16Jan. 2009.

[6] M. B. Malyutov, Authorship attribution of texts: A review. Lecture Notes in Computer Science, vol. 4123, pp. 362-380, 2006.

[7] K. H. Krippendorf, Content analysis- An introduction to its methodology, 2nd Ed., Sage Publications Inc., 2003.

[8] E. Stamatatos, N. Fakotakis, and G. Kokkinakis, Automatic authorship attribution, in proc. of Conf. on European Chapter of the ACL, pp.158-164, 1999.

[9] S. Argamon, M. Saric, and S. S. Stien, Style mining of electronic messages for multiple authorship discrimination: First results, in Proc. 9th ACM SIGKDD, pp 475-480, 2003.

[10] T. Zhang, F. Damerau, and D. Johnson, Text chunking using regularized winnow, in proc. 39th Annual Meeting on ACL, pp. 539-546, 2002.

[11] D. Pavelec, E. Justino, and L. S. Oliveira, Author Identification using Stylometric Features, Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial, vol. 11, no. 36, pp. 59-66, 2007.

[12] A. J. Nasir, N. G¨ornitz, and U. Brefeld, An off-the-shelf approach to authorship attribution, in Proc. 25th Int. Conf. on Computational Linguistics, (COLING), pp. 895-904, 2014.

[13] D. Bogdanova, and A. Lazaridou, "Cross-language authorship attribution", in Proc. 9th Int. Conf. on LRE, pp. 2015-2020, 2014.

[14] T. Chakraborty, "Authorship Identification Using Stylometry Analysis in Bengali Literature", CoRR, abs/1208.6268, 2012.

[15] S. Das, P. Mitra, "Author identification in Bengali literary works", in Proc. Int. Conf. on Pattern Recognition & Machine Intelligence, LNCS, vol. 6744, Springer-Verlag Berlin Heidelberg, pp. 220-226, 2011.

[16] S. Phani, S. Lahiri, and A. Biswas, "Authorship attribution in Bengali language", in Proc. 12th Int. Conf. on NLP, pp. 100-105, 2015.

[17] P. Das, R. Tasmim and S. Ismail, "An Experimental Study of Stylometry in Bangla Literature", in Proc. Int. Conf. on EICT, pp. 575-580, 2015.

[18] Society of Natural Language Technology Research (SNLTR), http://www.nltr.org/.

[19] Stylogenetics, http://github.com/olee12/Stylogenetics.

[20] Anisul Haque, http://www.facebook.com/pg/AnisulHoque71/notes/

[21] Somewhere in blog, "http://www.somewhereinblog.net/"

[22] Sachalayatan, "সচলায়তন" , "http://en.sachalayatan.com"

[23] Amra bandhu, "আমরা বন্ধু" , "http://www.amrabondhu.com/"