An Ensemble Learning Based Author Identification System

Ankita Dhar[1(✉)], Himadri Mukherjee[1], Sk. Md. Obaidullah[2], and Kaushik Roy[1]

[1] Department of Computer Science, West Bengal State University, Kolkata, India
ankita.ankie@gmail.com, himadrim027@gmail.com, kaushik.mrg@gmail.com
[2] Department of Computer Science and Engineering, Aliah University, Kolkata, India
sk.obaidullah@gmail.com

Abstract. Author identification is an emerging domain in the area of Natural Language Processing (NLP) that allows us to identify the respective author of a particular piece of text. Every author had some unique characteristics of writing that involves their signature style of applying specific terms, making their piece of art distinct and noticeable and also there exists an extended story behind the linguistic and stylistic analysis in the identification of authors. In this paper we aim to produce a content resemblance based author identification system (AIS) using ensemble learning model for identification of author for a given piece of text. The experiment was tested on approximately 6,000 passages from 26 authors obtained from Bangla literature. Experimental results reported that the proposed technique performed better compare to the state-of-the art methods.

Keywords: Author identification · Content resemblance · Rotation forest · Ensemble learning

1 Introduction

Authors basically possess a unique writing style that is quite noticeable in their articles. The writing style is not always relevant with the topic of the article and can be identified by the readers. The task of identifying the author from a set of different authors based on a piece of text that may vary from short to big articles is author identification. Author identification has various applications associated with it such as intelligence-detaining the messages connecting the terrorists; civil law-copyright and estate disagreement; criminal law-identification of writers of provoking and ransom letters; cyber security-finding out the authors behind the virus source code; detection of plagiarism and identifying ghost writer. Earlier, author identification problem generally relied on huge texts whereas the recent works shows the focus is on short texts because of the blooming trends of social media that uses short text messages [14]. Identification of authors can be classified in two means: closed category and open category [11]. The closed category

B. Deka et al. (Eds.): PReMI 2019, LNCS 11941, pp. 65–73, 2019.
https://doi.org/10.1007/978-3-030-34869-4_8

means represents the identifying author in a group while the open category problem leaves the author unspecified in the series. Most of the works followed closed category procedure that are said to be easier compared to open category means.

While dealing with text documents in Bangla, a lot of challenges need to be faced because of the unavailability of standard resources and tools. Bangla being the 6^{th} most popular language in whole world [6], demands the development of the technologies that help the users in predicting and choosing the respective author for a piece of text from a set of literature articles being considered based on various lexical, syntactic and analytical features. Also, Bangla based system can assist the users who are not well accomplished in western languages to utilize the information technology efficiently. These facts motivated us to devote our time and knowledge in this area of interest. Thus, development of an automatic author identification system in Bangla will have a great social impact especially in south Asia region. In this study, author identification has been considered as a supervised learning procedure, especially a task of multiple domain categorization problems. However, unlike traditional text representation schemes, we introduce a content based resemblance method for extraction of features and using rotation forest-an ensemble learning algorithm, the proposed model outperformed the existing models by obtaining an average accuracy of 98.75%. Also, we have encountered various articles from different authors having similar styles of writing their piece of art which made our task more difficult as well as challenging.

The rest of the paper is structured as follows: Sect. 2 provides a brief literature survey followed by Sect. 3 illustrating the proposed methodology; in Sect. 4, the experimental results have been analysed and comparative study with the existing works have been provided; and Sect. 5 concludes the paper showing some future works in this area.

2 Related Works

A general overview on author identification problem from the literature study have been observed that author identification has not widely been explored in Bangla and it is relatively an innovative field. We came across few works, for instance, Das and Mitra [4] have also worked with 36 text documents from 3 authors: Rabindranath Tagore, Sukanta Bhattacharya and Bankim Chandra Chattopadhyay. Uni-gram and bi-gram features were used and reported an accuracy of 90% for uni-gram and 100% for bi-gram feature. However, the database was too small to produce such genuine conclusions. Further, the authors being considered had different writing styles which made their study quite easy. Chakraborty and Choudhury [3] proposed 3 graph based methods by considering the association of series of character, use of certain expressions and formation of the sentences in a text document for the generation of individual graphs for an author. For each method, the graphs for different authors were clustered for developing a weighted graph and a simple graph traversal algorithm was used for author identification task. The experiment was tested on the articles of 6 authors and reported a maximum accuracy of 94.98% which they claimed to be 9.89%

higher than the standard method even for short texts. Phani et al. [10] worked with 3,000 passages, 1000 passages each from 3 Bengali authors: Sarat Chandra Chattopadhyay, Rabindranath Tagore and Bankim Chandra Chattopadhyay based on character n-grams, feature selection, ranking and analysis, and learning curve to evaluate the connection between the train and test accuracy. The experimental results show their proposed model obtained maximum accuracy of 98% based on character bi-gram feature. Rakshit et al. [12] delved into the articles for automatic extraction of linguistic and stylistic knowledge which are essential for the interpretation and differentiation of poems. They have worked with semantic based features for categorization of Bangla poems into 4 classes: devotional, love, nature and nationalism using SVM classifier and obtained an accuracy of 56.8% but they concluded that for poem categorization, word features are not sufficient and thus used semantic based features together with stylistic features and obtained an accuracy of 92.3%. Anisuzzaman and Salam [2] tested a hybrid approach by fusing n-gram and Naïve Bayes and achieved accuracy of 95%.

However, a number of similar works have been carried out for English, Marathi, Arabic and others. For instance, Madigan et al. [8] introduced scattered Bayesian logistic regression along with bag of words and bi-gram POS based features for identification of 27,342 articles from 114 authors. Alharthi et al. [1] proposed CNN based approach using different feature vectors for identifying the documents of Litrec dataset and obtained an average accuracy around 60%. Digamberrao and Prasad [5] developed two methods based on lexical and stylistic feature extraction schemes for author identification of 15 philosophical documents in Marathi from 5 authors and obtained 80% accuracy based on optimization with J48 algorithm. Otoom et al. [9] worked on a database comprises of 12 features and 456 articles from 7 authors. They combined the features with machine learning algorithms and achieved an accuracy of 82% for identification of the respective authors.

3 Proposed Methodology

The literature articles were first tokenized and then stopwords were removed followed by content resemblance measure and thereafter rotation forest algorithm was used to identify the author of the given piece of texts. The overview of the proposed system is diagrammatically demonstrated in Fig. 1.

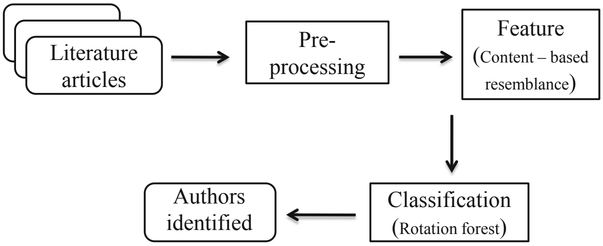
Literature articles
Pre-processing
Feature (Content – based resemblance)
Classification (Rotation forest)
Authors identified

Fig. 1. The overview of the proposed system.

3.1 Dataset

Data is the most important element of an experiment. It is very essential for the database to incorporate real world properties and disparity in order to develop robust systems. From the literature study it has been found that there is no standard author identification database is available in Bangla consisting of sufficient amount of text documents from adequate amount of authors and thus we had to build our own database. The present work includes approximately 6,000 passages written by 26 authors such as Rabindranath Tagore (1), Kazi Nazrul Islam (2), Sarat Chandra Chattopadhyay (3), Bankim Chandra Chattopadhyay (4), Jibananda Das (5), Sunil Gangopadhyay (6), Humayun Ahmed (7), Vivekanada (8), Sukanta Bhattacharya (9), Sukumar Ray (10) and others in Bangla literature from www.ebanglalibrary.com. The database consists of male and female authors of various ages with some similarity in the writing styles and also has an ample amount of works such as novels, poems, essays, songs, dramas, short and big stories and others thoroughly been digitized to be used for author identification tasks.

3.2 Pre-processing

Digital text documents are generated using a linear sequence of characters, words and phrases. Prior to the further processing or analysis on the data, the raw texts were segregated into linguistic element called 'token'. The process of segregation of sentences into such tokens is termed as 'tokenization'. To extract the tokens, they were split depending on 'space' delimiter in our experiment. After the task of segregation of sentences, the total count of tokens becomes 6,79,343. Since each and every token do not contribute enough useful information that can be prove to be useful in further classification task, thus there is no need to retain those elements in the feature set and hence stopwords were removed from the set of tokens to clean and filter the dataset. Since the selection of stopwords is problem specific, therefore in the present work, 355 Bangla stopwords have been used provided in [15]. The total count of tokens results in 5,57,214 after removal of stopwords.

3.3 Feature

In the experiment a feature extraction technique has been proposed based on the resemblance of the contents between two text documents and named it as 'content resemblance measure' feature. The feature has been proposed due to the fact that in some cases it has been observed that the contents of the articles written by two different authors have similarities which lead to the misclassification. Thus we focus on extracting those relevant features that can be used for reducing the change of these ambiguities in the contents. The features based on content resemblance measure $Con\_Sim(x, y)$ are determined by the Eq. 1.

$$Con\_Sim(x, y) = \frac{\sum_{p \in D(x|y), q \in D(y|x)} S\_Cos(p, q)}{|D(x|y)| * |D(y|x)|} \tag{1}$$

where $D(x|y)$ denotes the set of contents for document x but not for document y, $D(y|x)$ denotes the set of contents for document y but not for document x and $S\_Cos(p,q)$ is the similarity between the contents of two text documents calculated based on soft cosine concept. A soft cosine measures the similarities between a pair of features. The standard cosine similarity metric treated the features as individual feature, whereas the soft cosine considers the similarity of features between two vectors, that helps to establish the concept of soft cosine as well as the similarity between two features.

$$S\_Cos(p,q) = \frac{\sum_{m,n}^{L} sim_{mn} a_m b_n}{\sqrt{\sum_{m,n}^{L} sim_{mn} a_m a_n}\sqrt{\sum_{m,n}^{L} sim_{mn} b_m b_n}} \tag{2}$$

where sim_{mn} is the similarity between feature m and n and L is the total data.

3.4 Ensemble Learning Model

Rotation forest classifier is a tree-based ensemble learning model with some major differences to random forest classifier and implements transformation on the subsets of attributes before the construction of each tree. The two major differences of rotation forest lies with the transformation of the attributes into sets of principle components; and secondly it utilizes the algorithm of decision tree (J48) by default. However, random forest and bagging provides encouraging outcomes while dealing with large number of ensembles whereas rotation forest is particularly developed to deal with a small number of ensembles. In rotation forest algorithm, the number of trees needed in the forest is determined by the user. It was proved to be useful compare to other ensembles such as AdaBoost, bagging and random forest on a number of standard databases [13].

4 Results and Discussion

4.1 Our Results

An experimental set up was made in order to compare rotation forest algorithm with three other ensemble models mentioned above in Sect. 3.4 on the dataset consisting of around 6,000 passages with 5,57,214 tokens after pre-processing. In all ensemble models, J48 was used as the default classifier, an extended version of C4.5, except in the random forest algorithm that generates the tree in such a way that random choice of a feature can be made at each node. As PCA is being used as default parameter for projection filter, hence the discrete features were changed into numeric features for rotation forest classifier. The decision tree classification algorithm implements an error-based pruning model. The confidence value required while pruning the tree can be given by the user which is by default set as 0.25 and in most of the cases provides better outcomes, thus, it was left unchanged in our present work. We have evaluated the experiment using 5-fold cross validations, keeping J48 classifier fixed along with changing projection filter and the results are provided in Table 1.

Table 1. Accuracy obtained using J48 classifier along with changing projection filter.

Classifier	Projection filter	No. of leaves	Size of tree	Accuracy (%)
J48	**Principal component analysis**	41	81	**96.75**
	Random projection	59	117	93.50
	StringtoWord Vector	53	105	76.75
	Standardize	53	105	75.50
	Cartesian product	53	105	76.75

Since maximum accuracy of 96.75% was obtained using PCA as projection filter, therefore, we have further studied the experiment keeping PCA constant and changing the classification algorithm to best first tree (BFTree). Based on the way of pruning, the results are provided in Table 2.

Table 2. Accuracy obtained using BFTree classifier and PCA projection filter.

Classifier	Projection filter	Pruning	No. of leaves	Size of tree	Accuracy (%)
Best first tree	Principal Component Analysis	**Post-pruned**	53	27	**98.75**
		Un-pruned	55	28	97.75
		Pre-pruned	53	27	98.00

The maximum accuracy of 98.75% is obtained using BFTree classifier for post-pruned scenario along with PCA as projection filter and the number of leaves and size of tree being 53 and 27. The average percentage of correct identification for all the authors is shown in Fig. 2 where it can be observed that there is a fall of accuracy for identification of Rabindranath Tagore, Sarat Chandra Chattopadhyay and Bankim Chandra Chattopadhyay due to the reason of their presence during the golden era of Bengali Renaissance and were having similar style of writing. Another confusion has been observed between the writings of Syed Shamsul Haq and Shamsur Rahman and between Syed Shamsul Haq and Samares Mazumder.

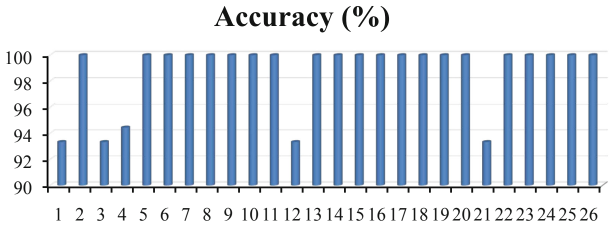

Fig. 2. The percentages of correct identification of all the authors.

We have compared the performance of the system with bagging, AdaBoost and random forest ensembles as well using WEKA [7] and the obtained accuracies together with precision, recall, F1 measure and Kappa statistics (statistic for measuring the score of homogeneity in the ratings provided by the judges and considers the similarity arises by chance) given for all the classifiers are presented in Table 3. We have considered BFTree classifier for bagging and AdaBoost ensembles as we have obtained maximum accuracy of 98.75% using BFTree for rotation forest. It can be observed from the Table that the performance of rotation forest is better among all the ensemble models being considered for the present experiment.

Table 3. Comparison among various ensemble learning models.

Ensembles	Precision	Recall	F1 measure	Kappa statistics	Accuracy (in %)
Rotation forest	**0.988**	**0.988**	**0.988**	**0.987**	**98.75**
AdaBoost	0.933	0.930	0.930	0.927	93.00
Bagging	0.886	0.880	0.880	0.875	88.00
Random forest	0.699	0.700	0.690	0.688	70.00

4.2 Comparison with Existing Methods

The working methodology have been compared with some of the recent works performed in Bangla. Our proposed approach outperformed all other existing methods for Bangla in terms of recognition accuracy. Also the dataset being developed for the present experiment is larger compare to all other works. The accuracies obtained for the existing methods along with our proposed model is demonstrated in Table 4.

Table 4. Comparison of the working methodology with other existing approaches.

Reference	Used feature	Accuracy (in %)
Chakraborty and Choudhury [3]	Graph based	94.98
Phani et al. [10]	Character n-gram	98.00
Proposed work	**Content resemblance based**	**98.75**

5 Conclusion

An ensemble learning based author identification system has been proposed in this paper which is capable of capturing the expressive behaviours of 26 authors and efficiently assigns the text pieces belonging to the respective authors using a content resemblance measure feature. The proposed system performed better compare to the existing approaches developed by others for the task. We have also harvested quite a large database in Bengali consists of around 6,000 passages which can be made publicly available on request. In future, we would like to explore some linguistically inspired features that can help in order to develop an automatic author identification system in Bangla. Considering the literature survey, it can be said that this study appears to be the first well grounded experiment for identification of authors in Bangla, especially in terms of number of authors as well as number of texts. Also in future, we would like to study the proposed methodology to other applications of NLP such as analysis of sentiments from blogs and tweets, detection of plagiarism in Bangla and others.

Acknowledgement. One of the authors thank DST for INSPIRE fellowship. We thank Dr. Amitabha Biswas, Asst. Prof. of Bengali Dept, WBSU and Ankita Shaw and Shilpa Roy, students of Comp. Sc. Dept, WBSU for their help while developing the dataset. We also thank Debajyoti Bose, Dept. of Electrical, Power & Energy, University of Petroleum and Energy Studies for his help during the preparation of this manuscript.

References

1. Alharthi, H., Inkpen, D., Szpakowicz, S.: Authorship identification for literary book recommendations. In: Proceedings of the International Conference on Computational Linguistics, pp. 390–400 (2018)
2. Anisuzzaman, D.M., Salam, A.: Authorship attribution for Bengali language using the fusion of N-gram and Naïve bayes algorithms. Int. J. Inf. Technol. Comput. Sci. **10**, 11–21 (2018)
3. Chakraborty, T., Choudhury, P.: Authorship identification in Bengali language: a graph based approach. In: Proceedings of the International Conference on Advances in Social Networks Analysis and Mining, pp. 443–446 (2016)
4. Das, S., Mitra, P.: Author identification in Bengali literary works. In: Kuznetsov, S.O., Mandal, D.P., Kundu, M.K., Pal, S.K. (eds.) PReMI 2011. LNCS, vol. 6744, pp. 220–226. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21786-9_37
5. Digamberrao, K.S., Prasad, R.S.: Author identification using sequential minimal optimization with rule-based decision tree on Indian literature in Marathi. Procedia Comput. Sci. **132**, 1086–1101 (2018)
6. Ethnologue: (2019). https://www.ethnologue.com/language/ben
7. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. ACM SIGKDD Explor. Newslett. **11**(1), 10–18 (2009)
8. Madigan, D., Genkin, A., Lewis, D.D., Argamon, S., Fradkin, D., Ye, L.: Author identification on the large scale. In: Proceedings of the Meeting of the Classification Society of North America (2005)
9. Otoom, A.F., Abdullah, E.E., Jaafer, S., Hamdallh, A., Amer, D.: Towards author identification of Arabic text articles. In: Proceedings of the International Conference on Information and Communication Systems, pp. 1–4 (2014)
10. Phani, S., Lahiri, S., Biswas, A.: Authorship attribution in Bengali language. In: Proceedings of the 12th International Conference on Natural Language Processing, pp. 100–105 (2015)

11. Potha, N., Stamatatos, E.: A profile-based method for authorship verification. In: Likas, A., Blekas, K., Kalles, D. (eds.) SETN 2014. LNCS (LNAI), vol. 8445, pp. 313–326. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07064-3_25
12. Rakshit, G., Ghosh, A., Bhattacharyya, P., Haffari, G.: Automated analysis of Bangla poetry for classification and poet identification. In: Proceedings of the International Conference on Natural Language Processing, pp. 247–253 (2015)
13. Rodriguez, J.J., Kuncheva, L.I., Alonso, C.J.: Rotation forest: a new classifier ensemble method. IEEE Trans. Pattern Anal. Mach. Intell. **28**(10), 1619–1630 (2006)
14. Sapkal, K., Shrawankar, U.: Transliteration of secured SMS to Indian regional language. Procedia Comput. Sci. **78**, 748–755 (2016)
15. Stopwords: (2019). https://www.isical.ac.in/~fire/data/stopwords_list_ben.txt