International Journal of Engineering & Technology, 7 (4.19) (2018) 17-21

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Author Identification of Bengali Poems

A.Pandian[1*], K.Manikandan[2], V.Ramalingam[3], Payal Bhowmick[4], Sree Vaishnavi[5]

[1] Associate Professor, [2,3] Asst. Prof.(S.G), [4,5]B.Tech Student, [1,3,4,5]Dept. of CSE, [2]Dept. of IT, SRMIST
**Corresponding Author Email: [1]pandian.a@ktr.srmuniv.ac.in*

Abstract

Author identification of Bengali poems is a project mainly focusing on identification of an author of a poem. We train the system using a dataset consisting of features extracted from poems by various authors. Features like count of characters, words, spaces, vowels and consonants of Bengali poems are considered. The training algorithm used is J48 decision tree. It has additional features of J48 like it accounts for missing values, prunes decision trees, and derives rules from the data, etc. All of this is helpful when we want to classify with larger datasets.

Keywords: *Authorship identification; Bengali poems; J48 decision tree authorship identification; bengali poems; J48 decision tree.*

1. Introduction

Author Identification is used in application of forensic analysis, electronic commerce, and to compute solutions to address various problems in the world today. Some authors write anonymously. It could be any kind of content; innocuous or detrimental. In some situations, to prevent harm, it is important to identify the author of a particular work. The data might be simple text or images from which the text needs to be extracted. In earlier days people used to identify a document which was very time consuming and also expensive whereas in the present, electronic documents can be identified by using modern and automatic techniques in very less time. In this way, we are continuously trying to automatize each and every process including author identification. Certain properties need to be identifiable in the poems to classify them. These are called features. Features need not be visible to the human eye; but the trained model should be able to make out the difference. There are many ways to extract features like character count, word count, whitespace count, count of vowels and consonants of Bengali literature. These features are used to train a model using a proper algorithm until efficiency is achieved. The techniques of data mining are used for many purposes and tools such as Weka provide simple GUI that assist us to apply advanced machine learning algorithms to the datasets.

2. Literature Review

2.1. Authorship Attribution in Bengali Language

Authors: Shanta Phani, Arindam Biswas from Infor-mation Technology IIEST, Shibpur Howrah 711103, West Bengal, India and Shibamouli Lahiri from Computer Science and Engineering University of Michigan Ann Arbor, MI 48109
Their method which is similar to the above, but uses corpus of 3,000 passages which is the work of three Bengali authors. Their authorship classification uses n-gram feature extraction technique on Bengali characters, the best features are selected, and feature ranking is done before analyzing. Hence it is indicated that lexical n-grams are the best features for author identification.

2.2. Automated Analysis of Bangla Poetry for Classification and Poet Identification

Authors: Geetanjali Rakshit, Anupam Ghosh, Push-pak Bhattacharyya, Gholamreza Haffari IITB-Monash Research Academy, India, IIT Bombay, India Monash University, Australia. In this method, again, the most useful features were identified. Their achieved accuracy was around 57%. On not being satisfying with the results, they used stylistic features like syntactic, orthographic and phonemic along with the word features and gained 92.3% accuracy by using a multiclass SVM Classifier.

2.3. Authorship Analysis and Identification Tech-niques: a Review

Authors: Mubin Shaukat Tamboli Department of Computer Engg, Amrutvahini COE, Sangamner, India and Rajesh S. Prasad, Ph.D Department of Computer Engg. Zeal Education Society, Narhe Pune, India.
This review paper involves an extensive research on multiple techniques used for author attribution on a sample test case. Here they put their foresight of future of being able to identify authors of literary text. They wanted to achieve identification with great accuracy and to find the solution for behavioral feature extraction from literary text.
They have used two different approaches:
1. writer dependent
2. writer independent

Their aim was to develop a powerful method for au-thor identification. This paper describes various methods like the Naive Bayes algorithm to select features and words which was used to classify text.

2.4. Author Identification in Bengali Literary Works

Authors: Suprabhat Das and Pabitra Mitra , Depat-ment of Computer Science and Engineering from Indian Institute of Technology Kharagpur, West Bengal.
In this paper, they study the problem of author identi-fication in Bengali literary works. They have taken three authors into consideration. The features extraction is based on simple unigram and bi-gram features along with the quality of vocabulary. They have got 90% accuracy from unigram feature and almost 100% from bi-gram features. They have used machine learning techniques as classification algorithms. The techniques like SVM and PCA.

2.5. Authorship Attribution in Tamil Classical Poem (Agananooru): a Mathematical Model

Authors: Dr.A.Pandian, Dr.V. V. Ramalingam, R. P. Vishnu Preet from Department of CSE, SRM Univer-sity,Kattankulathur and Dr.R.Varadharajan from De-partment of Maths, SRM University,.Kattankulathur.
In this paper, the main attention of this paper is to briefly perceive the authors of unidentified Tamil dataset in perspective of the work of recognized authors. The writer employs mukkoodarpallu dataset which consists of 800 poems. The classification of the features is done by using support vector device and bi-gram to attain an accuracy of 83%. The features used in this paper are Lexical character-based, Ratio of digits to Character count, Character count, Ratio of letters to character count, Ratio of uppercase letters to character count, Ratio of tabs to Character count and the occurrences of special characters.

2.6. Identification of Authorship in Tamil Classical Poem (Paripadal)

Authors: Dr.A.Pandian, V.V.Ramalingam and R.P.Vishnu Preet from Department of Computer Science and Engineering, SRM University
In this paper, the main attention of this paper is to briefly perceive the authors of unidentified Tamil dataset in perspective of the work of recognized authors. The writer employs paipadal dataset. The features were extracted by using a decision tree with the help of J48 algorithm. The accuracy of 82.6% has been incurred on this paripadal dataset.

2.7. Author Identification based on Word Distribu-tion in Word Space

Authors: Barathi Ganesh H B, Reshma U and Anand Kumar M Centre for Excellence in Computational Engineering and Networking Amrita Vishwa Vidyapeetham, Coimbatore, India. In this paper, the unigram, bigram and latent semantic features were considered and the similarity of texts was tested. The algorithms used in this project are Random forest tree, Logistic Regression and SVM. It has been observed that by using these algorithms, the proposed model gives the accuracy of 80%. The datasets used in this are essays, novels, reviews, articles of Dutch, English, Greek and Spanish languages.

2.8. Multi-Lingual Author Identification and Lin-guistic Feature Extraction– a Machine Learning Approach

Authors: Hassan Alam and Aman Kumar BCL Tech-nologies San Jose, CA, USA
In this paper, they have developed state-of-the-art semantic features to aid in author attribution for the Arabic language. To work with the rich structure of Arabic, they have intelligently used parse tree for features. They involved NLP parsers specific to Arabic; used lexicons, semantic heuristics, and semantic processing. Also, Machine Learning was used to train predictive models. This led to the formation of a system that identifies authors by their style of writing and also finds similarity with other works of the author and unidentifiable literary works found online. They have used Support Vector Machine or SVM for this task. They have yielded a staggering accuracy of 98%.

2.9. Author Identification for Digitized Paintings Collections

Authors: Razvan Condorovici, Corneliu Florea and Constantin Vertan, from the Image Processing and Analysis Laboratory, LAPI, University Politehnica of Bucharest, Romania.
This paper, presents an automatic system for the painter recognition from digital representations of different paintings. These paintings are described with low-level features like 3D RGB Histograms and Gabor Energy Features. They have used the possible eight classifiers and have achieved the best performance by using a Multi Class Classifier with 52.67% accuracy. The performance of the system has been evaluated by using database which contains 1800 paintings belonging to 15 different painters.

2.10. Author Identification by Automatic Learning

Authors: Jordan Frery, Christine Largeron Labora-toire Hubert Curien, Universitie Jean Monnet, Saint Etienne, France Mihaela Juganaru-Mathieu Institut H. Fayol, Ecole Nationale Superieure des Mines de St Etienne, France.
In this method of author identification problem, they have proposed three methods (DCM-Voting, DCM-classifier and DCM). DCM is a counting method that gives good results but its effectiveness depends on the documents containing at least two known documents from the same author. The second method is DCM-voting, which overcomes one of the problems of DCM because it uses several representation spaces but it cannot handle problems with only a single known document. The extension of DCM voting is DCM classifier which is based on decision trees which gives the results of DCM to build the input attributes and solves all the problems of DCM. The accuracy they obtained by using DCM classifier is 70.7%. The datasets in this paper consists of documents of authors in English, Spanish and Greek languages.

2.11. Authorship Identification and Author Fuzzy Fingerprints

Authors: Nuno Homem, Joao Paulo Carvalho INESC-ˇID TULisbon âAS¸ Instituto Superior TÃl’cnico Lisbon, Portugal.
In this paper, it acknowledges the problem of extracting fingerprints from texts and differentiate them with those known set of authors. It presents an innovative fuzzy fingerprint algorithm based on vector valued fuzzy sets. Words and other stylometric features are used to create the fingerprint. The implementation is based on an approximated fast and compact algorithm. The accuracy obtained from this proposed system is around 60%.

2.12. Author Identification using Sequential Minimal Optimization

Authors: John Jenkins, William Nick, Kaushik Roy, Albert Esterline Computer Science NC A and T SU Greensboro, USA and Joel Bloch Computer Science UNC Wilmington Wilmington, USA.
In this paper, they have proposed the system by combining unigram features and different types of stylometric features that involves n-grams and part-of-speech. Using a Corpus dataset of 2,500 different articles, they had effectively captured a "non-topic sensitive sample". By using Weka tool for machine learning, they successfully produced an accuracy ranging from 76 to 85 percent

using classification techniques such as Random Forest and Sequential Minimal Optimization (SMO).

2.13. Towards Author Identification of Arabic Text Articles

Authors: Ahmed Fawzi Otoom, Emad E. Abdullah, Shifaa Jaafer, Aseel Hamdallh, Dana Amer, The Faculty of Prince Al-Hussein Bin Abdullah II for Information Technology, The Hashemite University Zarqa, Jordan.
In this method, they have targeted the problem on identification of author of an Arabic text article. Their main motive is to develop an intelligent system. They have used a proposed novel dataset consisting of 12 features and 456 instances belonging to the 7 authors. The classification algorithms used in this paper are SVM and Functional trees. By using these algorithms they have obtained an accuracy of 82%.

2.14. Author identification in Albanian language

Authors: Hakik PACI, Elinda Kajo, Evis Trandafili Information Engineering Department, Information Tech-nology Faculty, Polytechnic University of Tirana Tirane, Albania Igli TAFA, Denisa Salillari Department of Math-ematic Engineering Polytechnic University of Tirana, Tirane, Albania.
In this paper, they have used the datasets of Albanian language for author identification. Their previous work was on the adoption of Dmitri Khmelev algorithm for identifying the authorship of Albanian texts. They have improved the systems on the same datasets that were used in their previous paper by considering the syntactic structure of Albanian sentences and also by adding specific linguistic elements. By using this method they have got better results than the previous results.

3. Experimental Procedures

Rabindranath Tagore is the most globally renowned figure of Bengali literature. His notable Bengali work in poetry is Geetanjali, a book consisting of beautiful poems for which he received the Nobel Prize in Literature. He acts as a bridge between early writers and writers of the modern age in terms of content and in his style of writing.
All the poets in bengali literature have their own unique style that many other languages do not have.The vowels in Bengali called shoroborno and the consonants are called benjonborno. There are some special characters called juktokhor. Juktokhor are the alphabet which are the combination of any two letters in Bengali literature. We have used the alphabet and some of the statistical features in making this project. Other elaborate information about the feature extraction, datasets, creating database and the process of training are given below.

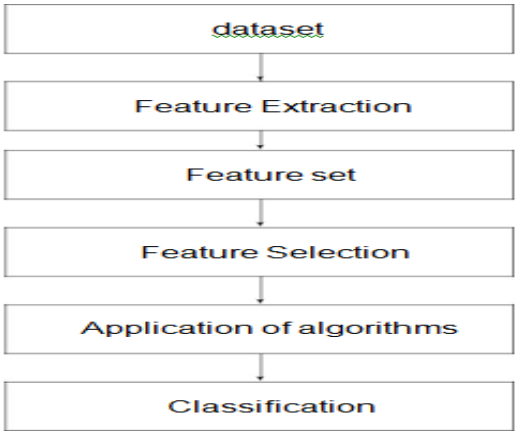

Fig. 1: Flowchart of the model we proposed

3.1. Collection of Poems by Various Authors

The poems we have used here are called datasets. We have manually collected the poems of different authors from different sites. The datasets are of 100 poems of each 5 authors which was collected from a popular website for Bengali poems.[15]

3.2. Feature Generation

Feature generation has been done by taking 100 poems of each author and extracting various features from the poems using a java program. The number of bengali word and the number of spaces between the word were counted initially. This was a rather simple task.
Other features that involved simple counting were sentence count and paragraph count. Once we had counted the number of words, whitespaces, lines and paragraphs, we could derive many other features from them using mathematical formulae of mean, median and mode. We also calculated the average number of characters in each word or the average word length by counting the total number of characters and dividing it by the total number of words. Similarly we found the average number of words in a sentence and the average number of characters in a paragraph. We then tried to train the model using these features, but it was not enough. We needed to find features that were more relevant to the bengali text. At this point, our accuracy was very low. The addition of slightly complicated features like Swaroborno count (count of bengali vowels in each poem), Benjonborno count(count of bengali consonants in each poem), Juktokhor count and matra count improved the accuracy greatly. Juktokhor are special characters in bengali that are formed by joining two other normal characters that may be consonants or vowels. Matra are special characters in bengali that signify the sound of each of the vowels (Swaroborno). These are referred to as Independent features in the list shown below. In addition to these, the ratio between the number of characters and whitespaces and such other dependent ratios were also calculated. The extracted features of each of the poems are then listed in a .txt file with a class label that refers to the author of the poem. All the features, that we have used in this project are listed below.
Statistical features:
Word (Count)
Character (Count)
Sentence/Line (Count) Average Word Length Paragraph (count)
Whitespace (count)
Mean of word (count)
Median of word (count) Mode of word (count) Average Word length
Average character in sentence
Average character in paragraph Average words in sentence
Average character in word
Average words in paragraph Ratio of Whitespace to words
Ratio of Whitespace to character
Independent features:
vowels count(Swaroborno count)
consonants count(Benjonborno count)
special character count(Juktokhor count) matra (count)

3.3. Creating Database with Generated Features

We use Microsoft Excel to create a database containing the various features that were generated in the previous step. Each set of features correspond to each of the five authors. Hence with this dataset we can train a model.

3.4. J48 Algorithm

Our particular problem can be best solved using the decision tree. ID3 (Iterative Dichotomiser 3),an algorithm proposed by Ross Quinlan, generates a decision tree from a dataset and can be easily

used in machine learning. J48 is an extension of ID3, also invented by Ross Quinlan. The algorithm was optimized to account for missing values, prune the decision tree, facilitate the use of continuous attribute values, and also derive rules. J48 is a Java implementation of the C4.5 algorithm and is available for use in WEKA.

3.5. Training using Weka Tool

Weka contains machine learning algorithms that can be used for data processing, classification, clustering and other such tasks. We have applied the j48 algorithm directly to the dataset created in the previous step.J48 algorithm.
Download and install weka tool.
Start weka tool. The Weka GUI Chooser lets us choose one of the options: Explorer, Experimenter, Knowledge Explorer and the Simple CLI (command line interface). The Explorer option allows us to load datasets and run classification algorithms.
Load the dataset.
Select j48 and click Start button to run this algorithm.
After running the J48 algorithm, we find the results in the ”Classifier” output section. The algorithm was run with 10-fold cross-validation: this means it made a prediction for each instance of the dataset.
The results are calculated with True-Positive, False-Positive, True-Negative, False-Negative and other values.

3.6. Figures and Tables

The below table shows the accuracy of different algorithms.

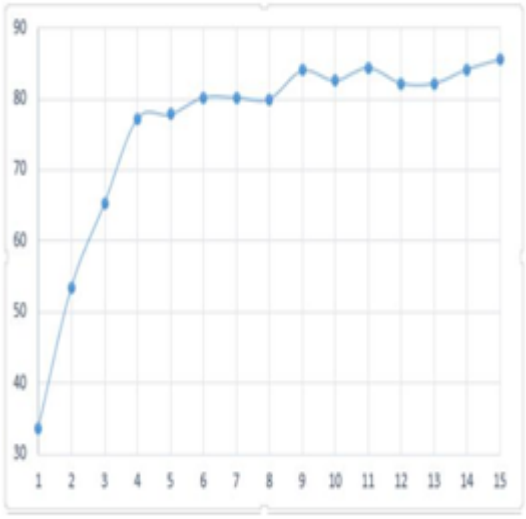

Fig. 2: The chart plots the achieved accuracies in percentage while the J48 decision tree is pruned with the minimum number of objects ranging from 1 to 15

3.7. The Best 15 Attributes are

Average words in sentence
special character count(Juktokhor count) vowels count(Swaroborno count)
Average words in paragraph Mean of word (count)
matra (count)
Ratio of Whitespace to character
consonants count(Benjonborno count) Average Word Length Character (Count)
Mode of word (count) Word (Count)
Average character in paragraph Average character in sentence Paragraph (count)

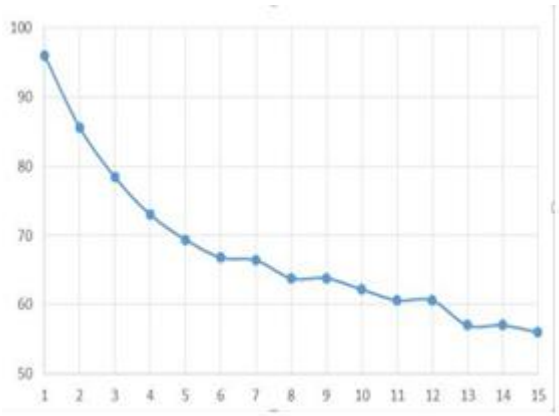

Fig. 3: The chart plots the achieved accuracies in percentage while the number of best attributes used in the training model ranges from 1 to 15

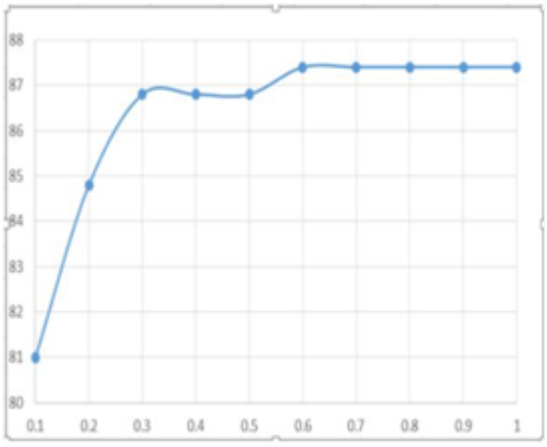

Fig. 4: The chart plots the achieved accuracies in percentage while the confidence factor ranges from 0.1 to 1

4. Results and Discussions

In our proposal, we have considered 100 poems of each five Bengali authors. We have achieved a higher score than the average accuracy which was 70-80% in most of the research papers. We achieved 87.4% using the j48 algorithm.
The identification of authors based on their work can be done through supervised learning techniques to achieve better accuracy. This process is much simpler when the author identification is done in English. The major task was to extract the various features of the literary works done by the author in vernacular language.
We have taken minimum number of objects and confidence factor as two important parameters. These two parameters helps us to obtain the best accuracy by using J48 algorithm. In fig.3 we have taken the accuracies of 15 best features by taking a default confidence factor value of 0.25. The best accuracy we have obtained by performing this is 85.6% and in fig.4 we have taken 2 as the default minimum number of object as we have obtained the highest frequency in fig.3 and the confidence factor in fig.4 is varied from 0.1 to 1. By performing this we have obtained the highest accuracy of 87.4%.

5. Conclusions

In this paper, using J48 the result is more accurate as compared to other algorithms. The quality and quantity of features generated directly affect the success of the classifying algorithm. On using J48 decision tree algorithm, we have achieved an accuracy of 87.4% after decision tree pruning and taking into account the confidence factor that gives us the best result.
For the future work, the research should be open to the scope of higher quality of features and better accuracy. Extensive research

needs to be done on using advanced methods such as deep learning to achieve tasks such as author identification by determining the style of writing of an author.

References

[1] Authorship Attribution in Bengali Language,Shanta Phani, Shibamouli Lahiri, Arindam Biswas, ltrc:iiit:ac:in=icon2015 =icon2015proceedings=P DF= 37rp:pdf

[2] Automated Analysis of Bangla Poetry for Classification and Poet Identification using SVM classifier ,2015, Gee-tanjali Rakshit, Anupam Ghosh ,Pushpak Bhattacharyya, Gholamreza Haffari.

[3] Authorship Analysis and Identification Techniques: A Review , International Journal of Computer Applications (0975 – 8887), Mubin Shaukat Tamboli , Rajesh S. Prasad, Ph.D, Volume 77 – No.16, September 2013

[4] Author Identification in Bengali Literary Works using probabilistic classification method. , S.O. Kuznetsov et al. (Eds.): PReMI 2011, LNCS 6744, pp. 220–226, 2011. Suprabhat Das and Pabitra Mitra ,Department of Com-puter Science and Engineering

[5] AUTHORSHIP ATTRIBUTION IN TAMIL CLASSI-CAL POEM (AGANANOORU): A MATHEMATICAL MODEL, Dr.A.Pandian ,V.V.Ramalingam and R.P.Vishnu Preet , 2016.

[6] IDENTIFICATION OF AUTHORSHIP IN TAMIL CLASSICAL POEM (PARIPADAL) USING J48 ALGO-RITHM Dr.A.Pandian ,V.V.Ramalingam and R.P.Vishnu Preet , 2016.

[7] Author Identification based on Word Distribution in Word Space, 978-1-4799-8792-4/15/$31.00 c 2015 IEEE Barathi Ganesh H B*, Reshma U* and Anand Kumar M.

[8] Multi-Lingual Author Identification and Linguistic Fea-ture Extraction — a Machine Learning Approach, 978-1-4799-1535-4/13/$31 c 2013 IEEE , Hassan Alam, Aman Kumar.

[9] Author Identification for Digitized Paintings Collections, 978-1-4673-6143-9/13/$31.00 c 2013 IEEE, Razvan Con-dorovici, Corneliu Florea and Constantin Vertan

[10] Author Identification by Automatic Learning, 2015 13th International Conference on Document Analysis and Recognition (ICDAR), 978-1-4799-1805-8/15/$31.00 c 2015 IEEE,Jordan Frery, Christine Largeron Labora-toire Hubert Curien

[11] Authorship Identification and Author Fuzzy ”Finger-prints”, 978-1-61284-968-3/11/$26.00 c 2011 IEEE, Nuno Homem , Joao Paulo Carvalho

[12] Author Identification using Sequential Minimal Opti-mization, 978-1-5090-2246-5/16/$31.00 c 2016 IEEE, John Jenkins, William Nick, Kaushik Roy, Albert Esterline, Joel Bloch

[13] Towards Author Identification of Arabic Text Articles, 2014 5th International Conference on Information and Communication Systems (ICICS), Ahmed Fawzi Otoom, Emad E. Abdullah, Shifaa Jaafer, Aseel Hamdallh, Dana Amez

[14] Author identification in Albanian language, 2011 In-ternational Conference on Network-Based Information Systems, Hakik PACI, Elinda Kajo, Evis Trandafili, Igli TAFA, Denisa Salillari.