TEST Engineering & Management

May – June 2020
ISSN: 0193-4120 Page No. 25467– 25472

Study of Author Identification and Verification Systems Using Statistical and Stylometry Models on Different Languages

Swapnali Balasaheb Ware, Rajesh S. Prasad
Department of Information Technology, Sinhgad Institute of Technology, Lonavala, Pune, India
swapnali25.ware@gmail.com, rajesh.prasad@sinhgad.edu

Article Info
Volume 83
Page Number: 25467– 25472
Publication Issue:
May - June 2020

Article History
Article Received*: 11 May 2020*
Revised*: 19 May 2020*
Accepted*: 29 May 2020*
Publication*: 12 June 2020*

Abstract:
The process of Authorship Identification and Verification is the process of identification and verification who wrote a given piece of text from Anonymous Text Document set written by different writers or authors. On this system, huge work has been already done for the languages such as English, Japanese, Albanian, Indian, Brazilian, Chinese and Russian and so on.
Comparatively, huge research has been done for Author Identification or Attribution only but not Author Verification for many of Indian languages such as, Punjabi, Bengali, Telugu and Tamil etc, whereas very less work has been done on Marathi Literature and Guajarati language. In this paper, methods and experiments on above languages presented by different researchers have been reviewed. We have also studied about different text features extraction techniques and how that can be applied for better result in Author identification and then verifying the same using soft computing algorithm.

Keywords: *Authorship Identification; Authorship Verification; Anonymous Text Document; Text Features.*

I. INTRODUCTION

Authorship Identification and verification system can be used in multiple application where you need to find out who is the suspect of written documents. Significant amount of research work has been done all over the world for different languages regarding this as it has a biggest role in identifying and verifying who has written a particular document and finding out the suspect. It has been observed from study that research work on Indian regional languages, various types of techniques, methodologies and algorithms has been used for a better author identification and verification system, but still the systems are not full proof and there is tremendous scope for improvement in this area.
In this process of Author Identification and Verification general steps are as follows:

Step I: Collection of documents written by different authors;
Step II: Identify different feature set such as length of sentences, paragraph similarity, phrases, formatting of text, numeric value, paragraph-title similarity and so on. These features can be used to identify stylometry of the author;
Step III: Construct the classifier for applying this testing and training datasets using machine learning algorithms; and
Step IV: By using testing parameters such as Precision, Recall and Confusion Matrix finding the similarity score of training and testing text datasets.

May – June 2020
ISSN: 0193-4120 Page No. 25462– 25466

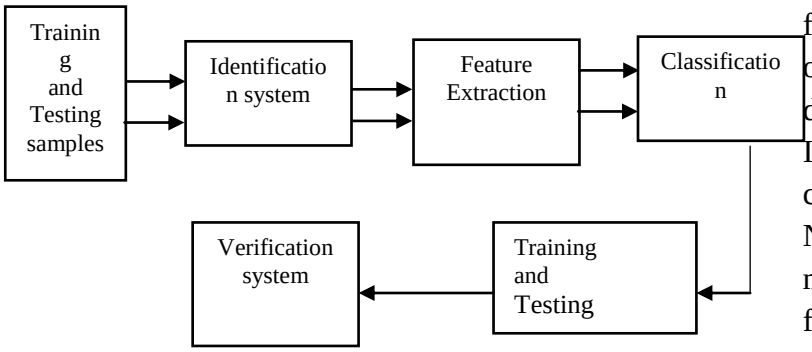

Fig1. Framework of Author Identification and Verification System

II. AUTHOR IDENTIFICATION AND VERIFICATION OF DIFFERENT LANGUAGES

Categorization of research of Authorship Identification and Verification techniques on various Languages.

This Author Identification and Verification is todays need in multiple fields such as in computer forensic application, in copyright issues, in plagiarism etc. here it can be beneficial to identify and verify the authorized person by taking any sample datasets [1].

A. Author Identification on Arabic language

In reference [2] researchers developed a system for identification of an author on a large handwritten Arabic language-based documents. They have introduced a novel system for offline text from documents using Block Wise Local Binary Count (BW-LBC) model for characterizing the hand writing variability. Here this model used for resizing and extracting white pixels from input handwritten documents. The well-known classifier Nearest Neighbor (1-NN) used for classification of those handwritten sample documents and Hamming distance measure used for matching to respective feature vectors. This had achieved best performance on datasets which contains an Arabic and English data.

In future scope they mentioned to use another classifier from machine learning like Artificial Neural Network or the very well support vector machine to study good impact of these classifiers for measuring the performance of this system.

In reference [3] researchers used the different models from machine learning field named as simple Naïve Bayes (NB), Multinomial Naïve Bayes (MNB) algorithm, Multi-variant Bernoulli Naïve Bayes (MBNB) algorithm and Multi-variant Poisson Naïve Bayes (MPNB). Here they have used 10 different authors of Arabic large documents and compared these with the existing methods which are available. The system showed that multi-variant Bernoulli Naïve Bayes (MBNB) gave better result for author attribution of a given written piece of text from Arabic document with accuracy of 97.43% as compared to other algorithms.

They have selected a feature set taken from the table given in reference [4] from the Arabic datasets.

In reference [5] researchers have done Authorship Attribution in Arabic Poetry. Here they have used feature extraction techniques on poetry of 54 different poets. For extraction of the features they have used a Weka library named as RapidMiner for preprocessing text and applied it to the well-known algorithms from machine learning such as Naïve Bayes algorithm, Support Vector Machine algorithm, Sequential Minimal Optimization classification algorithm out of which Sequential Minimal Optimization (SMO) gave a better accuracy of 98.15%. For measuring accuracy of attribution of an author they have used precision and recall measuring parameter.

In [6] researchers presented Textual features for offline text dataset such as Local Binary Patterns (LBP) technique, Local Ternary Patterns (LTP)

May – June 2020
ISSN: 0193-4120 Page No. 25462– 25466

technique, and Local Phase Quantization (LPQ)technique, these kinds of techniques used for extraction of images or sub-images from a handwritten sample document. The texture of the image is given to feature extraction by using LBP, LPQ, or LTP techniques. They have used Arabic databases, and English databases for the experimentation.

B. Identification of Author on Japanese language

In reference [7] researchers worked on text-independent writer identification using JEITA-HP database which contains for each writer, there are 3,306 handwritten samples belonging to 3,214 categories including 2,965 Kanji, 82 Hiragana, 10 numerals, and 157 other categories (English alphabet, Katakana and symbols) are present in the database. For writer's identification they have used the CNN- based method which achieved accuracy rate of 92.80% for identification of writer by training 50 characters for 100 writers and 100 characters for 400 writers from JEITA-HP handwritten offline database.

For applying the method for identification of not known authors or registration of a new writers is a big challenge here.

C. Author Identification on Bengali language

In reference [8] researchers have used a Bengali Literature written by three very well-known authors namely Rabindranath Tagore, Bankim Chandra Chattopadhyay and Sukanta Bhattacharya. Methodology they have used here, they selected 36 documents from which 11 documents are selected for training purpose and around half percentile data from each author used for training and all remaining documents are classified and this used for author identification of Bengali literature.

For larger training datasets accuracy measured is 90% above for the feature like unigram feature and 100% accuracy achieved for bi-gram features of given Bengali Literature written by three different authors.

Challenges are here when smaller sized training set was used the result slightly degraded.

In reference [9] researchers have presented the analysis on behavior and writing style of an author of a Bengali Literature for author identification. They developed two different models one is statistical similarity model which consist of three measures Cosine similarity (COS), Chi-square measure (CS), Euclidean distance (ED), Majority voting (COM) and their combination and other is machine learning model using Decision Tree, Neural Networks and Support Vector Machine algorithms from machine learning.

In this system they have used 150 stories written by Indian Nobel laureate Rabindranath Tagore. Here SVM gave better experimental result.

D. Author Identification on Guajarati language

Very less work has been done on Guajarati Language for author identification, the reference [10] researchers presented recognition of character from offline handwritten Guajarati Language. The Guajarati characters are recognized by Pattern Matching using Neural Network with result of recognition efficiency of 71.66% for the offline Guajarati dataset. Also, in reference [11] researchers presented a Template matching method for identification of character by analyzing the shape of a character and by making comparison the features for distinguishing each character from the offline handwritten dataset.

E. Author Identification on Hindi language

Further, reference [12] shows identification of writer of Devanagari Hindi Language which also used in Marathi, Konkani and Sanskrit languages.

May – June 2020
ISSN: 0193-4120 Page No. 25462– 25466

They have collected data mostly written by students from age 12 to 25 years old and others in between 25 to 60 years old. They have used Devanagari alphabets, numerals and vowels from around 250 documents which are written by around 50 different writers. For identification of writer of these documents they have used 5-fold cross validation scheme. The digitization is done on data after it the 64-dimensional extraction of features done and passed to LIBLINEAR and LIBSVM which are a suitable linear classifiers of WEKA tool for classification purpose. From LIBLINEAR they have achieved accuracy rate of 99.12% for identification of writer of the 5 documents per user from total number of 50 users from a given data.

F. Identification of an Author on Telugu language

In reference [13] researchers have used different unique features for Authorship Attribution (AA) on a Telugu text document. They have used the 300 Telegu news articles which are written by 12 different authors from which for training purpose 20 documents per author used while for testing 5 documents per writer has been used. From a preprocessed text the features have been extracted and then grouped. Support Vector Machine classifier used for building a model while performance measurement F1 metric and accuracy used for Authorship identification. They achieved the result where character ngram features are giving more successful result for Authorship Attribution for a Telegu text.

In reference [14] another work on Telugu text has been carried out on 120 text articles as a training dataset and 5 articles for test dataset has been used. They have used different unigram, bigram, trigram and tetra-gram character and word features among which “Unigram Feature” and “Trigram Feature” gave a best performance when applied to SVM classifier. For accuracy measurement they have used F1 Measure and Exactness measuring parameters.

G. Author Identification on Marathi language

In reference [15] researchers presented a system for Author Identification of a Marathi article. Here, they have used feature extraction technique on a Marathi literature written by 5 authors. They proposed a novel classifier which is a combination of different algorithms such as Sequential Minimal Optimization and Rule-based Decision Tree (SMORDT) which has been tested for Marathi article. They have evaluated the result of accuracy by using different measuring parameters and got the good result. For future work they have presented that accuracy can be enhanced by increasing the feature set like ngram feature for word and character for extraction from Marathi articles and using a topic modelling for enhancing performance of this proposed system.

Reference [16] shown a system for recognition of Marathi handwritten character. Here, the researchers have used 200 set of 40 handwritten Marathi characters for experimentation. For the extraction of a feature and detection they have used Rectangle Histogram Oriented Gradient(R-HOG) from a Marathi handwritten documents. They have used feed-forward neural network and support vector machine for classification and for accuracy measurement they have used False Accept Rate (FAR), False Reject Rate (FRR) and True Accept Rate (TAR) which shown result that performance of feed-forward neural network is outperformed than Support Vector Machine.

This technique gave poor result for identification of some character which looks quite similar to each other.

In [17] researchers proposed technique for author identification on Marathi literature. They have used Katha, Lekh, Natak, Lalit of 20 well authors in

May – June 2020
ISSN: 0193-4120 Page No. 25462– 25466

Marathi. Here the feature extraction is performed on the training and testing dataset from the given Marathi corpus. Word ngram method has been used and applied to Support Vector Machine, Naïve Bayes and K-nearest Neighbor. The experimentation result shown that according to size of training dataset the accuracy gets vary which does not give a satisfactory result.

In [18] researchers did survey on Author Identification on different languages spoken in worldwide. They have shown here that feature extraction and data size of training data has an important role while identification of an author of a given documents. They have presented there is requirement of developing datasets for different languages worldwide is necessary. Also, have shown there is wide scope of research to be done in different languages except English language as the most of the work has been done in it for Author Identification but very few on any other language. They have also shown the style of writing is important parameter for Author Identification and verification. So, selection of the feature is important to identify the stylometry of an author.

In Reference [19] researchers for Author Identification they have used news-paper columns from New York Times, The Indian Express and the correspondence of well-known writers. For experimentation they have used features such as character n-gram, word n-gram and POS n-gram. Here they have used SVM algorithm for classification. For accuracy measurement parameter for author identification used with TFCT (transform feature to current time) function. The limitation here is the negative impact is seen on accuracy when to a to a certain limit the dimension of the feature vector is increased.

III. CONCLUSIONS

From existing system, it has been observed that there are variety of features, which have been extracted from larger and smaller datasets for Author Identification and verification but less work is done so far on Author Verification on Marathi literature. Various parameters have been used for measuring the accuracy for the same and also different model has been build using various machine learning algorithms. In many cases SVM outperformed. If the larger dataset is used, we will get a greater number of features and words for extracting it from a dataset and this will be more useful for identifying the style of writing by an author. Marathi is one of the most spoken language in India. The systems are operated in Marathi languages across the state of Maharashtra and other parts of India. Such systems will help researchers and professionals to work and make a robust frame work for Indian regional languages.

IV. REFERENCES

[1] [1] C. Qian, T. He, and R. Zhang, “Deep Learning based Authorship Identification”, Stanford, CA, 2017.

[2] [2] Abderrazak Chahia, Issam El-Khadiria, Youssef El merabeta, Yassine Ruichekb, Raja Touahnia, “Block Wise Local Binary Count for Off-Line Text-Independent Writer Identification”, published in Expert Systems with Applications ,2017.

[3] Altheneyan A S and Menai M E., "Naïve Bayes classifiers for authorship attribution of Arabic texts." J King Saud Univ - Computer Inf Sci. in 2014.

[4] Abbasi, A., Chen, H., “Applying authorship analysis to Arabic web content”, In: Kantor, P., Muresan, G., Roberts, F., Zeng, D.D., Wang, F.-Y., Chen, H., Merkle, R.C. (Eds.), Intelligence and Security Informatics, vol. 3495. Springer-Verlag, Berlin, Heidelberg, 2015.

[5] Alfalahi Ahmed, Ramdani Mohamed and Bellafkih Mostafa,

May – June 2020
ISSN: 0193-4120 Page No. 25462– 25466

"Authorship Attribution in Arabic Poetry Using NB, SVM, SMO", at IEEE in 2016.

[6] Abderrazak Chahi, Youssef El merabet, Yassine Ruichek, Raja Touahni, "An effective and conceptually simple feature representation for off-line text-independent writer identification", Elsevier January 15, 2019.

[7] Cuong Tuan Nguyen, Hung Tuan Nguyen, Takeya Ino, Bipin Indurkhya, Masaki Nakagawa, "Text independent writer identification using Convolution Neural Network ", Pattern Recognition Letters, Elsevier, 2012.

[8] Das S and Mitra P., "Author Identification in Bengali Literary Works.", published in International Conference on Pattern Recognition and Machine Intelligence ,2011.

[9] Tanmoy Chakraborty, "Authorship identification in Bengali literature: a comparative analysis",2011.

[10] Prasad J. R., Kulkarni U. V. and Prasad R. S, "Offline handwritten character recognition of Gujrati script using pattern matching", In Anticounterfeiting, Security, and Identification in Communication in2009.

[11] Prasad J. R., Kulkarni U. V. and Prasad R. S., "Template matching algorithm for Gujrati character recognition", In Emerging Trends in Engineering and Technology ©IEEE,2009.

[12] Chayan Halder, Kishore Thakur, Santanu Phadikar and Kaushik Roy, "Writer Identification from Handwritten Devanagari Script", Information Systems Design and Intelligent Applications for Springer India in 2015.

[13] Prasad S N, Narsimha V B, Reddy P V and Babu A. V., "Influence of Lexical, Syntactic and Structural Features and their Combination on Authorship Attribution for Telugu Text", Procedia Computer Sci. in 2015.

[14] S. Nagaprasad, N. Krishnaveni, J. K. R. Sastry and A. Vinayababu, "On Authorship Attribution of Telugu Text", Indian Journal of Science and Technology, September 2016.

[15] Kale Sunil Digamberrao, Rajesh S. Prasad, "Author Identification using Sequential Minimal Optimization with rule-based Decision Tree on Indian Literature in Marathi", International Conference on Computational Intelligence and Data Science, 2018.

[16] Kamble P M and Hegadi R S., "Handwritten Marathi Character Recognition Using R-HOG Feature." ProcediaComput Sci. 45: 266– 274,2015 .

[17] Sunil D. Kale1, Rajesh S. Prasad, "Author Identification on Imbalanced Class Dataset of Indian Literature in Marathi" , International Journal of Computer Sciences and Engineering, 2018.

[18] Kale Sunil Digamberrao, Rajesh S. Prasad, "Author Identification on Literature in Different Languages: A Systematic Survey", at International Conference on Advances in Communication and Computing Technology in 2018.

[19] Mubin Shoukat Tamboli, Rajesh Prasad, "Author identification with feature transformation method", in Digital Scholarship in the Humanities, 2019.