AUTHOR IDENTIFICATION OF HINDI STORIES

[1]Dr.A.Pandian, [2]Dr. M. Abdul Karim Sadiq, [3]Paritosh Maurya, [4]Nitin Jaiswal

ABSTRACT*—Attribution also called Authorship Identification determines the probability of work that is produced by any author by examining other works from that same author. This process is used in various places like Characterization of work of an author, detecting Plagiarism, Cybercrime analysis etc. In this paper, we are using this process on a corpus of 70 Hindi Stories each from three different authors. Various lexical and structural features are extracted from these works like Word count, Average length of sentence, Frequency of words and characters, Function Words etc. With help of these features we build a dataset and use it as input in J48 decision tree algorithm for determining the best features that help in authorship attribution. We then use these extracted features on different types of algorithm like SMO, Bayes Net, Naïve Bayes, J48 etc. and select the algorithm with the best accuracy for classifying author.*

Keywords*— Author Identification, Feature Selection, Hindi Stories, J48 Decision Tree, Machine Learning, Stylometry, Weka.*

I. INTRODUCTION

In Author identification we determine the most likely author of a given work from a list of given authors. Recently, this task of authorship identification has been steadily increasing in capturing attention of researchers due to its vast applications like Plagiarism detection, email classification, genre detection, threat detection and analysis and various other forensic applications. A lot of information is posted on the internet daily. Some of this information is posted anonymously. The information can be innocuous or detrimental. And hence it's necessary to find the authors of these texts. Each author possesses writing style varying completely from other authors. The task of authorship identification is performed by studying writing styles and characteristics of different authors, extracting features from the authorship known texts and then training a machine learning model to perform the classification. The set of features can include lexical features like the frequency of characters, vowels, function words, etc., semantic features like tense, voices and syntactic features include part of speeches and phrases.

Most of the work performed in this field of authorship identification has been done- in well-known languages like English, Arabic. Researchers have shown their interest in regional languages like Bengali, Punjabi, Tamil, Urdu, etc. However, Hindi still lacks behind in this area.

Hindi just falls behind English and Mandarin as the third most spoken language around the world. Around 366 million people speak Hindi worldwide. The main reason for lack of research in hindi language is the non-availability of tools and corpus, and lack of interest shown by researchers.

[1] Associate Professor. CSE Department. SRM institute of Science and Technology, Chennai, India. pandiana@srmist.edu.in
[2] College of Applied Sciences, Sohar, Oman.
[3] Student, CSE Department, SRM Institute of Science and Technology, Chennai, India, pm5732@srmist.edu.in,
[4] Student, CSE Department, SRM Institute of Science and Technology, Chennai, India, itsnitin@ymail.com.

Received: 22 Feb 2020 | Revised: 13 Mar 2020 | Accepted: 05 Apr 2020

In this paper; we are trying to perform author identification on hindi stories. For this purpose, we are choosing three different authors' namely Sachchidananda Hiranand Vatsyayan "Agnay", Kamleshwar, Vishnu Sharma. The corpus consists of 70 stories from each author collected from the website [16]. We will extract features from their works and give these as input to classifiers for classification.

II. STATE OF THE ART (LITERATURE SURVEY)

In [1] the paper suggests a syntactic approach for achieving authorship identification. Syntactic features along with a categorization method based on dissimilarity were used. Different approaches using polytomy and dichotomy were also tested. These experiments were conducted on four different databases containing literary and journalistic texts in Portuguese and English. On conclusion of the experiments, the proposed approach generated accuracy rates of 77% and 94.5%, for English Language, and 78.3% and 98% for the Portuguese language. in the identification as well as verification, respectively.

In [2], Noise Arabic texts have been used for several authorship attribution experiments. The database used for this experiment consisted of extracted texts of same genre and topics from books of 5 ancient Philosophers. This dataset is called 'AP4'. Different classifiers like SVM, MLP, Linear Regression were used on features, namely character -N grams and words. The Results indicate that classifiers and features used have a great effect on the noise limit ratio for authorship identification. The maximum noise limit for getting a fair authorship identification was determined to be 450 words for each document. [2]

In [3], the paper suggested the use of different models like Multinomial naïve Bayes (MNB), Simple Naïve Bayes(NB), multi-variant Bernoulli naïve Bayes and multi-variant Poisson naïve Bayes (MPNB) for authorship identification of Arabic texts. Works of 10 distinct authors were used for creating a large Arabic database and evaluating this. The results have indicated that MBNB reaches an accuracy of 97.43%. Thus, providing the best result.

In [4], the paper posed an approach to perform authorship attribution of documents written in Marathi Language. For analysis of text, a set of concrete stylistic and lexical features were adopted. Two distinct models namely Sequential minimal optimization with rule-based decision tree approach (SMORDT) and statistical similarity model were developed. Articles written by 5 different authors were used as dataset. Precision, Recall, f-measure and accuracy form the basis of evaluation of the performance of model. The values calculated were as 80%, 50%, 61.54% and 80% respectively.

In [5], researchers have tried to identify authors gender using authorship attribution from Arabic texts. For this they have suggested that writing styles of similar gender have some common aspects. These aspects can be secured using stylometric features. Second method of approach they focused on was keyword occurrence called bag-of-words(BOW) approach. 1000 Arabic articles written by distinct authors of both gender were used as Dataset. The result show that first approach was more efficient and more accurate than latter one. The best accuracy achieved for the first approach was using the JRip rule-based classifier. It was 80.4%. On the other hand, the best accuracy for the latter approach was 73.9% and achieved using the support vector machine (SVM) classifier.

In [6], three graph based models were introduced by authors for the task of authorship attribution. These graphs consider the interaction of phraseological patterns, character sequences and structure of the sentences in the

Received: 22 Feb 2020 | Revised: 13 Mar 2020 | Accepted: 05 Apr 2020

document for each author. Graphs for each author are combined to generate an aggregated weighted training graph for all of these models. Then the testing graph is compared with training graph with help of simple graph traversal technique. Documents of six authors from Bengali literature were used as corpus for this experiment. Experimental results display that our models significantly outperform four state-of-the art models.

In [7], researchers have analysed the methods of authorship identification on Azerbaijani texts. N- grams algorithm with values n=1 and n=2 was used for this approach. 50 newspaper articles written by 4 different authors were used as dataset in this experiment. For experimental results, they have been able to achieve an accuracy of 75%.

In [8], researchers have used a 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 analysing. Hence it is indicated that lexical n-grams are the best features for author identification.

In [9], the paper has proposed a semantic association model showing writing styles of various authors using word dependency relations, voice, and non-subject stylistic words. Stylistic features were extracted using an unsupervised approach. Two models namely, (PCA) Principal component Analysis and (LDA) Linear Discriminant Analysis were used for identification of the author. Two publically available text corpus were used as dataset for extracting features namely Reuters corpus volume 1(RCV1) and English Books. The above proposed approach has a significant improvement of performance over SVM, genetic algorithms and Neural networks.

In [10], the authors present two modules- one that allows novel stylistic spaces to be recognised and another that acts as a classifier to perform the task of authorship identification from those features derived from first modules. The methodology uses feature selection, anomaly detection, classification, and visualization algorithms. Self-organizing maps were used for anomaly detection and visualization. These maps describe the basis of the classifier. The proposed method generated lowest error under a novel stylistics space which is based on the rate of introduction of new words. Also, similar or better results were generated under bag-of-words-related stylistics spaces approach.

In [11], the authors extracted various features from Bengali poems and made a dataset of it using it to train the system. Different features like count of characters, spaces, vowels, etc. were considered. The training algorithm used was J48 algorithm. The highest accuracy achieved was 87.4% which is considerably higher than other papers according to the authors.

In [12], Features like bigram, unigram and latent semantic features were considered. The resemblance of texts was tested using these models. The algorithms used in this project are SVM, Random forest tree and Logistic Regression. The proposed model gave an accuracy of 80% by using these models. The datasets used in this are essays, novels, reviews, articles of Dutch, English, Greek and Spanish languages.

In [13], an Arabic language dataset is used. Classification is performed using the Markov chain algorithm generating a precision of 96.96%. The most ideal approach to extract features pertinent to the Arabic dialect is demonstrated. Each part that is associated with the dataset and that also satisfies the defined Markov property is a valid unit that can be used for classification. These elements are chosen hence used to build the classifier.

In this paper, multiple components are explored that are possible attributes to extraction of features from datasets. Enron E-mail was the dataset used and classification was done using bisecting K-means algorithm and E-M algorithm giving a 90 % precision.

In [15], the paper covered issue using the Fisher's Linear Discriminant and Radial Basis Function algorithms. They are dispersed on the Dataset of Enron email. Components are concentrated in order to decode the origin of a particular article from the Enron email dataset by using spiral premise calculation for grouping in with a precision of 80% to 90%.

III. PRPOSED WORK

Corpus collection-

A set of large amount of texts in any language is called Corpus. The texts present in these corpuses are usually chosen from a distinct set of fields so that they are representative of their language. Our corpus of hindi stories consist of 70 stories from 3 different authors. We have manually collected the poems from the popular website for hindi poetry [16].

Feature Set and Extraction

Feature Extraction accumulates a package of derived information from an initial set of raw data to make processing more manageable. Corpus directly cannot be used as a tool for setting up the classifier (i.e. training the model). But the features that are extracted from the corpus can be used to form a dataset and finally utilized to assemble the classifier. The features were extracted from the corpus using python language as a tool. Code for feature extraction was written in Jupiter notebook and the file containing the corpus was open in Unicode format. Hindi characters cannot be comprehended by computers so the Unicode provides an encoding framework so that computers can understand and comprehend Hindi characters. The set of features that are extracted included count of hindi words and whitespaces and different function words. This was easy to perform. Other features that include simple counting are the number of lines and the number of paragraphs. Once these features were extracted, other features were derived from these using mathematical formula of mean, mode and median. Average word length was calculated by dividing total characters by total words. Similarly, we calculated average character in sentence, average character in paragraph, average words in sentence, ratio of char to whitespace and added them as features. Features related to hindi language were also added like count of Swaron (vowels) and count of Vyannjana (consonants) and ratio of vowels to consonants. Frequencies of function words were also calculated along with occurrences of characters in the text. All the features used in this paper are shown in table [1].

Table 1: Different features used in dataset

1. Feature Type	*2. Features*

3. Lexical	4. Word count
5.	6. Paragraphs count
7.	8. Character count
9.	10. Sentence count
11.	12. Whitespace count
13.	14. Vowel count
15.	16. Consonant count
17.	18. Frequencies of Characters (59 character)
19.	20.
21. Statistical	22. Average character in paragraph
23.	24. Average words in sentence
25.	26. Ratio of char to whitespace
27.	28. Average word length
29.	30. Average character in sentence
31.	32. Ratio of words to whitespace
33.	34. Average words in paragraph
35.	36. Ratio of consonant to vowel
37.	38.
39. Syntactic	40. Occurrence of function words. 41. Conjunction (35 words) 42. Preposition (41 words) 43. Pronoun (75 words)

Creating Database

We use Microsoft Excel to create a database containing the various features that were generated in the previous steps. Firstly, the extracted features were stored in a text file and then imported to an excel file. Each set of feature corresponds each of the three authors. Hence with this dataset, we can train a model.

IV. IMPLEMENTATION

Weka is a tool written in java language and performs machine learning operations. It is a collection of visualization tools and different algorithms for data analysis, along with a graphical user interface for easy access to such functions. Authorship identification is a classification problem. One of the best ways to solve such a problem is with the help of a decision tree. Weka provides one such algorithm called J48, proposed by Ross Quinlan. It is the development of the ID3 algorithm. J48 is optimized to account for missing values, pruning the tree and use of continuous attribute values and derived rules. J48 builds a decision tree. Each internal node of J48 tree denotes a test on an attribute, each leaf holds a class label and each branch denotes the result of the test. This algorithm can be used for extracting the best features which help in the classification process. Weka provides an interface for extracting the best features. These extracted features have a vital part in improving the accuracy of different classifier algorithms. The accuracy of an algorithm varies on each dataset used. Therefore, to find the best algorithm for this dataset, different types of algorithms need to be implemented and the best one needs to be selected. The best features are shown in Table II.

Fig.1 represents the Architecture Diagram for the proposed model.

Table 2: Best 9 Features

Best Features
Average Characters in Sentences
Average Characters in Paragraph
Ratio of Character to Whitespace
Frequencies of Aur
Frequencies of Anyatha
Frequencies of Kintu
Frequencies of Mano
Frequencies of Kuch
Frequencies of Visarga

Received: 22 Feb 2020 | Revised: 13 Mar 2020 | Accepted: 05 Apr 2020

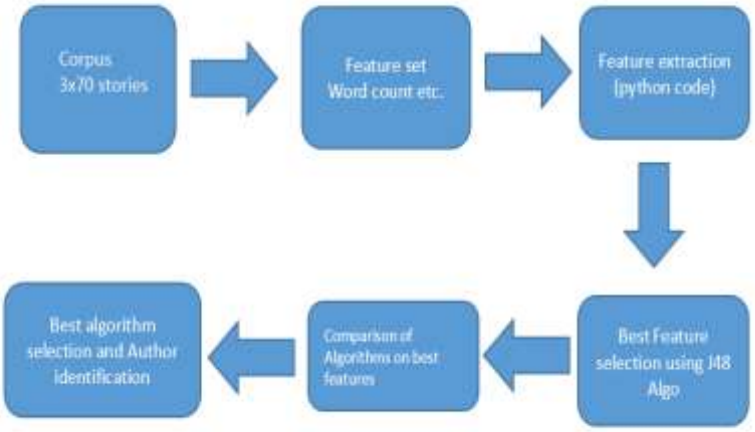

Fig1. Architecture Diagram

Figure 1: Architecture

V. RESULTS AND DISCUSSION

The outcome of the comparison of algorithms to their corresponding accuracies is listed in Table III. These accuracies were found using Weka explorer by training the dataset.

The Logitbost algorithm and the Random Forrest algorithm have given a peak accuracy of 95.83% on the dataset of best attributes. Algorithms like KStar have given an accuracy of 93.75%. Algorithms like Multi-layered Perceptron, Multiclass Classifier and J48 have given same accuracy of 91.67%. Classification via Regression also have performed well giving an accuracy of 89.58%. JRip and Bayes Net have produced an accuracy of 87.5% and 85.41% respectively. Random Tree algorithm has produced an accuracy of 83.33% while Hoeffding Tree has performed to produce an accuracy of 81.25% and Naïve Bayes producing 77.08%.

Algorithms like SMO and OneR have produced an accuracy of 75% and 72.91% respectively. Both LML and Decision Stump algorithms have produced an accuracy of 70.83%. On the other hand, Naive Bayes Multinomial has produced accuracy of 54.16%. Algorithms like ZeroR, MultiScheme and Naive Bayes Multinomial text have produced the least accuracy of 41.667%.

Fig. 2 shows the algorithms with their derived accuracy.

Table 3: Different algorithms and their Accuracy

ALGORITHMS	ACCURACY
SMO	75
Randomizable filtered classifier	83.3333
Naïve Bayes Multinomial Text	41.6667
JRip	87.5

Logitbost	95.8333
KStar	93.75
Attribute selected classifier	64.8649
Random Forrest	95.8333
J48	91.6667
Multilayer Perceptron	91.6667
Random Tree	83.3333
Multi Class Classifier	91.6667
Classification via Regression	89.5883
LML	70.8333
OneR	72.9167
Bayes net	85.4167
Naïve Bayes	77.0833
Hoeffding Tree	81.25
Naïve Bayes Multinomial	54.1667
Decision Stump	70.8333

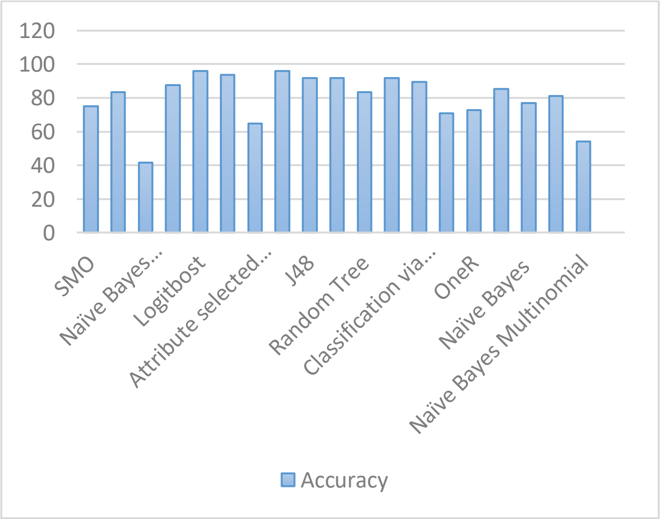

Figure 2: Accuracy of different algorithms

VI. CONCLUSION

In our work, we have examined twenty algorithms for classification problem. Logitboost and Random Forrest have achieved the highest accuracy of 95.8333% on the dataset. Algorithms like J48, Multiclass Classifier and Multilayered Perceptron have also achieved accuracy of 91.67%. Other algorithms like Classification Via Regression, Bayes Net, OneR, LWL etc. have achieved an accuracy ranging from 70%-89%. Algorithms like

Received: 22 Feb 2020 | Revised: 13 Mar 2020 | Accepted: 05 Apr 2020

ZeroR and multi Scheme have achieved the lowest accuracy on the dataset. For future work, the research should include more authors and use different dataset or increase features. Thus, increasing the accuracy of the model.

REFERENCES

1. Paulo Varela et al, “A computational approach for authorship attribution of literary texts using sintactic features”, 2016 International Joint Conference on Neural Networks (IJCNN)
2. S. Bourib et al, “Author Identification on Noise Arabic Documents”, 2018 5th International Conference on Control, Decision and Information Technologies (CoDIT’18), pp 216-221.
3. Alaa Saleh Altheneyan et al, “Naïve Bayes classifiers for authorship attribution of Arabic texts”, Journal of King Saud University – Computer and Information Sciences (2014) 26, 473–484.
4. Kale Sunil Digamberrao et al, “Author Identification using Sequential Minimal Optimization with rule-based Decision Tree on Indian Literature in Marathi” Procedia Computer Science (2018) volume 132, pp 1086-1101.
5. Kholoud Alsmearat et al, “Author gender identification from Arabic text” Journal of Information Security and Applications 35(2017),85-95
6. Tanmoy Chakraborty et al, “Authorship Identification in Bengali Language: A Graph Based Approach” 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp 443-446.
7. Aida-zade K.R. et al, “Authorship Identification of the Azerbaijani Texts Using n-grams”, 2016 IEEE 10th International Conference on Application of Information and Communication Technologies (AICT)
8. Shanta Phani et al, “Authorship Attribution in Bengali Language” Proceedings of the 12th International MConference on Natural Language Processing (2015) pp:100-105.
9. Chunxia Zhang et al, “Authorship identification from unstructured texts” Knowledge-Based Systems (2014), 99-111
10. Antonio Nemeab et al, “Stylistics analysis and authorship attribution algorithms based on self-organizing maps” Neuro Computing (2015) Volume147 pp:147-159.
11. A. Pandian et al, “Author Identification of Bengali Poems” 2018 International journal of Engineering and Technology, Vol 7, No 4.19 pp 17-21.
12. Barathi Ganesh H B et al, “Author Identification based on Word Distribution in Word Space, 2015 IEEE
13. Al-Falahi Ahmed et al, “Authorship Attribution in Arabic Poetry”,78-1-4799-7560-0/15, 2015, IEEE
14. Michael R. Schmid et al, “E-mail authorship attribution using customized associative classification, Digital Investigation, Volume 14(2015) pp 116-126.
15. Pandian, A et al, "Authorship categorization in email investigations using Fisher's linear discriminant method with radial basis function" (2014) Journal of Computer Science, 10 (6), pp. 1003-1014
http://www.hindi-kavita.com

Received: 22 Feb 2020 | Revised: 13 Mar 2020 | Accepted: 05 Apr 2020