International Journal of Engineering & Technology, 7 (4.19) (2018) 13-16

International Journal of Engineering & Technology

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

Research paper

Authorship Identification of Punjabi Poetry

A. Pandian[1*], StephenWahid[2], Yash Tokas[3], V.V.Ramalingam[4]

[1]Associate Professor, [2, 3]UG Student (B. Tech.), [4]Assistant Professor (S.G),
[1,2,3,4] Department of CSE, SRM University, Kattankulathur
**Corresponding Author E-Mail: [1]pandian.a@ktr.srmuniv.ac.in*

Abstract

The problem of identifying the author of an anonymous text is basically Authorship Identification. It is nothing but a single-label text-categorization task, from the ML point-of-view. An assumption is made that an unknown text's author can be differentiated by comparing a few lexical features extracted from theunknown text with the same of texts having known authors. In this paper, the process of Authorship Identification is executed on Punjabi poetry dataset consisting of Punjabi poems written by 5 different poets. Various features broadly categorised as statistical (word-count, char-count, etc.), syntactical (i.e. lexical) and semantically (language dependent) are first selected using the J48 Decision Tree Algorithm. The selected features are in turn, used as an input to multiple classifiers (like SVM, SMO, Bayes Net & Naive Bayes) and the proposed system's validation is evaluated on the basis of Precision, Recall, F-score and Accuracy.

Keywords: *Authorship Identification, Punjabi poetry corpus, Feature extraction, J48 Decision Tree, Bayes Net Classifier, Naive Bayes Classifier*

1. Introduction

In Indian regional languages, authors of many old poems and texts are not yet known. For instance, in the Punjabi language section, authors of various poems are not alleged. In Punjabi, a vast number of authorless poems is linked with a few poets, whose name and works arerecognized. Identifying them would be of more use to the people.

So, by utilizing a sensible computational technique, creators of the unidentified poems might have a chance to be discovered for their unaccounted works. Thomas Bayes (1871) utilized quantifiable hypothesis for discovering issues with identification of creation in the federalist papers. Auguste de Morgan (1851)had proposed the mean length of a word as a factor to decide the authorship of an article.

Perceiving those creators of a lyric on the support from claiming complex characters is the writer attribution issue clinched alongside etymological examination. Finishing characteristic extraction might help but that's only the tip of the iceberg with this creation attribution, which includes extracting a real and only those each every so often used Characteristics in words, period for sentence, momentous characters used, length about expressions etc.

In [1], 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 usingbisecting K-means algorithm and E-M algorithm giving a 90 % precision.

In [2], classification of components explicit to the Tamil Language was doneusing algorithms like SVM, proximal SVM and random kitchensink computations. SVM performsclassification by creating two disjoint spacesand classifying every entry as one of the two, while Proximal SVMfirst designates data centers to the closer of the two parallel lines and classifies the dataset accordingly. RandomKitchen Sink figuring uses all the possibleindependent factors and generates a measurable count. The precisions accomplished are 95.7%, 95.8%and 96.82% respectively.

In [3], an accuracy of 87.5% is achieved by usingrandom forest algorithm on 86052 words and 500788 characters.

In [4], an accuracy of 82% is accomplished on Arabic poems, which utilizes SVM, neural networks and Markovchain as classifiers for data.

In [5], specific features are extracted from a Tamildataset that contains approximately 5000 words. Classifiers generate an accuracy of 72 to 82 percent. These algorithms (i.e. FLD & RBF) are used to defeat the clashing issue. FLD algorithm performs grouping by making a straight mix of parts that isolates no less thantwo classes of things. Radial Basis Function calculation issimply an indistinguishable neural network framework. It works in perspectiveof neuron parameters.

In [6], 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 extractfeatures pertinentto the Arabic dialect is demonstrated. Each part that is associated with the dataset and that also satisfies thedefined Markov property is a valid unit that can be used for classification. These elements arechosen hence used to build the classifier.

In[8], the problem of authorship identification of oldTamil scripts is tackled. These scripts are first digitalized, and then classification is performed using SVM Classifier and uni-gram, bi-gram features which results in an accuracy of 83%.N-grams are oftenused when the data is discourse or a content corpus. Uni-gram is a size one n-gram and bi-gram is a size two n-gram.

In [9], the covering issue using the Fisher's Linear Discriminant and Radial Basis Function algorithms is dispersed on the Enron email dataset,while in [10], components are concentrated in order todecode the origin of a particular article from the Enron email dataset by using spiral premise calculation forgrouping in with a precision of 80% to 90%.

In [11],Tamil letters are viewed from their old scripts with the help of the LabVIEW tool and using segmentation, classification on the dataset is achieved. The Enron email wastreated as the dataset used and CALO (Cognitive Assistant that learns andClassifies) was used to accumulate the dataset, which contains e-mailsfrom approximately 150 clients.

In [12], relevant feature extractionis demonstrated and the accuracyof each respective classifier is calculated. Enron email is the dataset used 6 types of features are selected. An accuracy of 90.08% was achieved.Multiple algorithms were further used to calculate respective accuracies: Versatile Metropolis Algorithm gave 68.19%,N-Bayesgave 79.07%, Bayes Net algorithm gave 79.86%, CMAR algorithm gave 88.47%, CBAalgorithm gave 84.18% and finally, 90.08% was achieved by the CMARAA algorithm.

In [13-16], distinctive elements are used to performclassification and their respective precisions are noted. The expectation-maximization algorithm is an iterative classification technique. It performs a cycle between two phases Eand M. The desired step (E) constitutes thespecific occurrence of the likelihood and theaugmentation step (M) amplifies the typical likelihood recorded in each desired step.

2. Materials & Method

Finding the authors for un-authored Punjabi writings get is a particularly troublesome task as there is no system to recognize them explicitly. By extracting features pertaining to the Punjabi dialect used in its poems and by using suitable calculation, writers for these un-authored poems can be perceived. Fig. 1 demonstrates the architecture followed in such a classification.

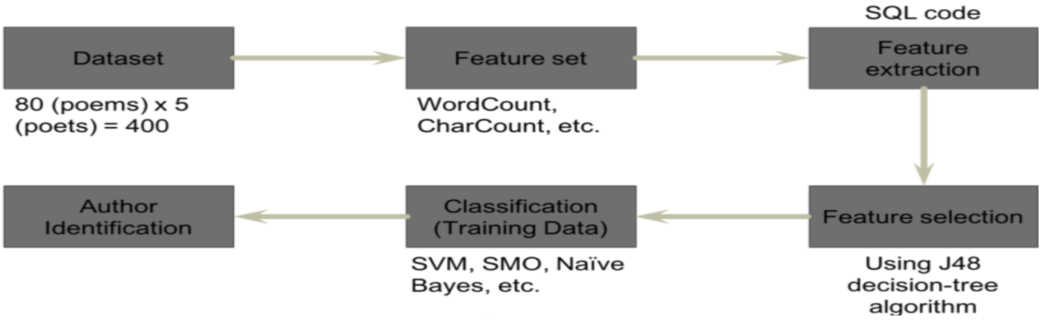
Dataset
80 (poems) x 5 (poets) = 400
Feature set
WordCount, CharCount, etc.
SQL code
Feature extraction
Feature selection
Using J48 decision-tree algorithm
Classification (Training Data)
SVM, SMO, Naïve Bayes, etc.
Author Identification

Fig. 1: Architecture Diagram

The datasets used here is 80 poems each of 5 eminent Punjabi poets namely Baba Bulle Shah, BawaBalwant,BhaiVir Singh, Prof. Mohan Singh and Prof. Puran Singh. The poems for these 5 poets is extracted from different sites like punjabi-kavita.comand shivbatalvi.com. By extricating syntactic, lexical and semantic elements as in [15], classification is performed. Main features that are considered aredepicted in Table 1.

The dataset is used to extract the mentioned features and these features are further used for the classification process. The author’s stylometry is characterized by thesefeatures. Stylometry is defined as the basic difference in composed literary styles of multiple writers. It consists of semantic, lexical and syntactic elementsapplicable to the specific language. Table - 1 depicts all the features extracted from the dataset. An accuracy of 86.66% was given by the J48 algorithm.

Table 1: Features Category

Features type	**Features**
Lexical:	
Character-based	
1.	Akhar(Character) count (N)
2.	Akhar-Space Ratio
3.	Akhar Frequency (35 features)
4.	Vowel count (2 types)
5.	Velar count
6.	Palatel count
7.	Retroflex count
8.	Dental count
9.	Labiel count
10.	LG count
11.	EndingAkhar (A [Aa], N [Na, Ni], L[La, Li])
Lexical:	
Word-based	
12.	Token/Word count(T)
13.	Average token length
14.	Sentence/Line count
15.	Average sentence length (in terms of N, T)
16.	Word Frequency
Syntactic:	
17.	Punctuation frequency (, . ? ! : ; ‘ “) (8 features)
Statistical:	
18.	Mean
19.	Minimum
20.	Maximum
21.	Sum

Figure 2 shows the lexical character features that are concentrated withineach dataset. The 35 features are explained briefly with broad categorisations.

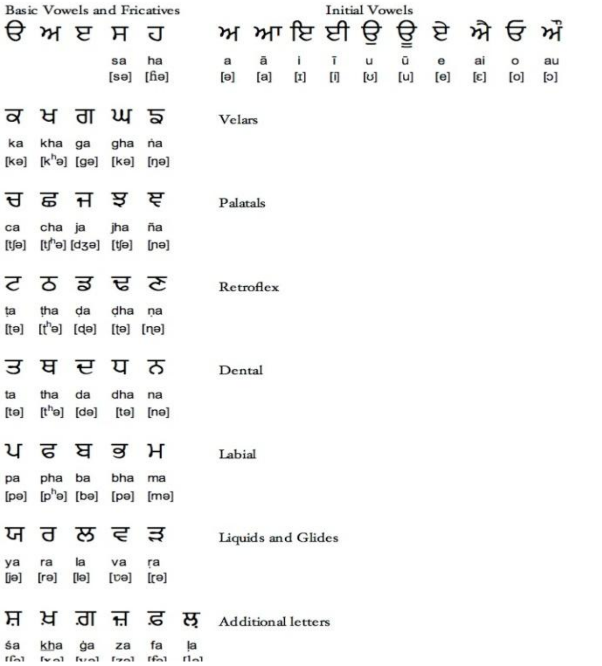
Basic Vowels and Fricatives
ੳ ਅ ੲ ਸ ਹ
sa ha
Initial Vowels
ਅ ਆ ਇ ਈ ਉ ਊ ਏ ਐ ਓ ਔ
a ā i ī u ū e ai o au
ਕ ਖ ਗ ਘ ਙ
ka kha ga gha ṅa
Velars
ਚ ਛ ਜ ਝ ਞ
ca cha ja jha ña
Palatals
ਟ ਠ ਡ ਢ ਣ
ṭa ṭha ḍa ḍha ṇa
Retroflex
ਤ ਥ ਦ ਧ ਨ
ta tha da dha na
Dental
ਪ ਫ ਬ ਭ ਮ
pa pha ba bha ma
Labial
ਯ ਰ ਲ ਵ ੜ
ya ra la va ṛa
Liquids and Glides
ਸ਼ ਖ਼ ਗ਼ ਜ਼ ਫ਼ ਲ਼
śa kha ġa za fa ḷa
Additional letters

Fig. 2: Character Features List

Table 2: Accuracy Percentage of the best features considered

Features	Accuracy Percentage
Minimum	41.33
Palatel Count	59.67
Avg Sentence length	63.33
Char Frequency	69
Mean	69.67
Line count	76.67
Vowel count	81
Word count	81.67
Labiels	80
Dentals	80.33
Avg token length	82.67
Ending akhar	83.33

2.1 Feature Extraction and Selection

Feature extraction is concerned with assembling an arrangement of derived qualities from the underlying arrangement of information pertaining to human translation. Datasets can't specifically be utilized as an input to classifiers, i.e. training the data. Features are extricated from the datato form a Feature Set, and that in turn, can only be utilized to assemble the classifier. This classifier that is built is then used to perform the classification process on the Feature Set in hand.

Three types of features are extracted, i.e. lexical, syntactic and statistical. Example of lexical features are adjective, verb, noun and pronoun. Few examples of syntactic features include verb phrase, noun phrase and prepositional phrase.

In addition to these features, statistical features are also extracted from the dataset. Statistical features account to a major part of the classifier accuracy. The classifier accuracy has increased from 86% to 90% by including statistical features to the features set and performing some tweaks in the algorithm used. Statistical features include Minimum, Maximum, Sum, Mean.

All features mentioned in table-1 are extracted from the dataset. The poem dataset is manipulated into Unicode indexes so that features can beextracted easily using smart SQL queries. Computers can't comprehend Punjabi characters. They bargain with just numbers in memory. Unicode indexing helps to converteach character of the regional language and gives an approach to computers to comprehend them.

The extraction procedure is done by utilizing SQL commands, which can extricate the predetermined features consequently. Sequel Pro is utilized to make a database with every one of the poems and components. The extracted features are in numeric format.

These numeric features that are extracted are all used in the classification process as all of these features play a vital role in improving the classifier accuracy to a great extent.

In order to choose the accuracy contributing features, and neglecting the unwanted ones, feature selection process is done. J48 algorithm is used to perform the feature selection process which is a decision tree algorithm. The authors have used J48 algorithm to perform the feature selection process, which implements the decision tree algorithm. The tree obtained by using the algorithm is shown in figure-2. The table-3 consists of a brief description of the best features.

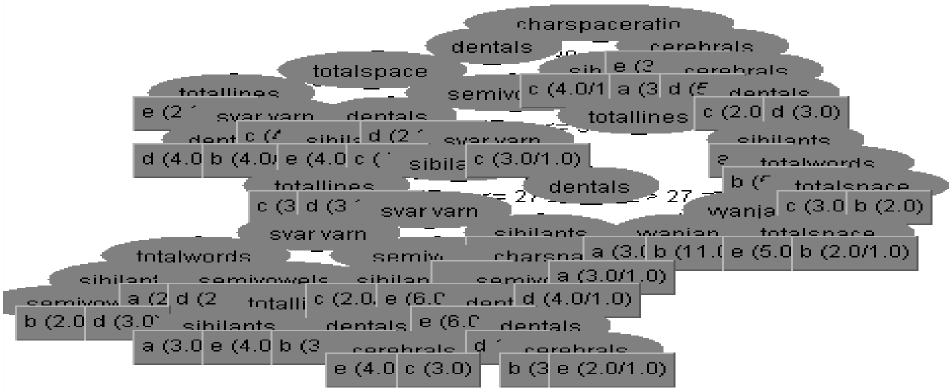

Fig. 2: Decision Tree Construction Using J48 Algorithm

Table 4: Best features description

Features	Description
Ending Akhar	This feature consists of the frequency of the frequently used end-characters of a line in the poem
Avg Token length	This feature is the total number of characters in a poem divided by the number of words.
Word Count	The Word Count feature consists of the overall count of the words present in a particular poem.
Vowel Count	The number of main 3 vowels present in a particular poem

2.2. J48 Classification Algorithm

J48 algorithm is developed by Ross Quinlon. This algorithm will be a development of the ID3 algorithm that might have been being used sooner times. C4. 5 algorithm constructs a choice tree. Following are the steps of the algorithm:

1. Check for the base cases.
2. For every attribute x, split on x and find the information gain.
3. Let the highest information gain attribute be x1.
4. Create a node that splits on x1.
5. Use the subsets of x1 to iterate the same process and add all the nodes as children of x1.

Table 5: Confusion Matrix

	A	B	C	D	E
A	15	1	1	1	0
B	0	24	2	0	0
C	2	1	16	7	0
D	1	0	6	12	0
E	0	0	0	1	0

2.3. Implementation of Classification Algorithm

The algorithms listed in Table – 5 are chosen and were used for implementing on the dataset in hand. These algorithms are already proven to have given a decent accuracy on various other datasets. The implementation process was performed by the use of the Weka tool.

Algorithms are not always guaranteed to provide the same maximum accuracy on all datasets. The accuracy of each algorithm varies on each dataset. So, to find the best suited algorithm for our dataset, all the related algorithms have to be implemented and the best algorithm has to be selected.

3. Results and Discussions

The outcome of the comparison of twenty related algorithms to their corresponding accuracies is listed in Table – 2.

The Random Forest algorithm which has given its best accuracy on certain datasets has given an accuracy of 73.33% on the dataset at hand. The Naïve Bayes algorithm has also performed well on various other datasets while on the dataset at hand it has given a accuracy of 66.6%. The K-Star algorithm has produced an accuracy of about 63.33% while the OneR algorithm has performed to produce an accuracy of 10% and SMO algorithm producing 76.66%. J48 algorithm has produced an outstanding 86% on the dataset at hand. The Multilayer Perceptron algorithm which is considered to perform well on almost all datasets has given an accuracy of 80%.

The LWL and Logit Boost algorithms have given a similar accuracy of 70% respectively, while the Random Tree algorithmhas given 63.66% accuracy on the dataset. The Randomizable Filter Classification algorithm and Random Committeealgorithm have all produced almost the similar accuracy of 60% and 63.33% respectively. The IBK algorithm has produced an accuracy of 83.33% whereas the JRip algorithm has produced an accuracy of 40%. TheOneRand AdaBoost M1 have all produced the least accuracy of 10% and 6.66% respectively.

Table 6: List of Algorithms

S.no	Algorithm Used	Accuracy Achieved
1.	J48	86.66%
2.	Random Forest	73.3%
3.	Bayes Net	46.6%
4.	Naïve Bayes	66.66%
5.	KStar	63.33%
6.	OneR	10%
7.	Attribute Selected Classifier	73.33%
8.	Randomizable Filter Classifier	60%
9.	Sequential Minimal Optimization (SMO)	76.66%
10.	Locally Weighted Learning (LWL)	70%
11.	IBK	83.33%
12.	JRip	40%
13.	Random Tree	63.33%
14.	Multilayer Perceptron	80%
15.	Logit Boost	70%
16.	Decision Table	43.33%
17.	Naïve Bayes Multinomial	66.66%
18.	Bagging	53.33%
19.	Random Committee	63.33%
20.	AdaBoost M1	6.66%

4. Conclusion

Out of the twenty algorithms considered for classification, the J48 algorithm has performed well and has given an maximum peak accuracy of 86.66% on the dataset. Other algorithms like IBK and Multilayer Perceptron have also provided a decent accuracy ranging from 80% - 83.33%. Algorithms like OneR and AdaBoost M1 have given the least accuracy of 10% and 6.66% respectively. Out of the 20 algorithms used for comparison, the J48 algorithm has performed well with an accuracy of 86.66%.

References

[1] FarkhundIqbal, HamadBinsalleeh, Benjamin C.M. Fung,MouradDebbabi, 2015, “E-mail authorship attribution usingcustomized associative classification”,DigitalInvestigation(Elsevier),Vol.7,pp.56-64

[2] Sanjanasri J.P andAnand Kumar M, “A Computational Framework for Tamil DocumentClassification using Random Kitchen Sink”, IEEE 2015, International Conference onAdvances in Computing, Communications and Informatics(ICACCI)

[3] Mahmoud Khonji, Youssef Iraqi, Andrew Jones,“An Evaluation of Authorship Attribution Using Random Forests”, IEEE 2015, International Conference on Information andCommunication Technology Research (ICTRC2015)

[4] Ahmed Fawziotoom, Emad E Abdullah, ShifaaJaafar, AseerHamdellh, Dana Amer, “Towards Author Identification of Arabic Text Articles”, IEEE 2014, 5th InternationalConference on Information and Communication Systems(ICICS)

[5] Pandian, A., and Md. Abdul KarimSadiq, 2014, “AuthorshipCategorization In Email Investigations Using Fisher’s LinearDiscriminate Method With Radial Basis Function”, InternationalJournal of Computer Science, Vol.10,No.6,pp.1003-1014 (SNIP: 0.874)

[6] Al-Falahi Ahmed, Ramdani Mohammad, Bellahfkimustafa, Al-Sarem Mohammad, “Authorship Attribution in Arabic Poetry”,78-1- 4799-7560- 0/15, 2015, IEEE

[7] Ahmed FawziOtoom, Emad E. Abdullah, ShifaaJaafer, AseelHamdallh, Dana Amer“Towards Author Identification of Arabic Text Articles”, 2014,IEEE, 5th International Conference on Information andCommunication Systems (ICICS)

[8] BhargavaUrala k, A.G.Ramakrishnan and Sahil Mohammad, “Recognition of Open Vocabulary, Online Tamil HandwrittenPages in Tamil Script”, 2014 IEEE, Vol.42, No.3, pp.6-9.

[9] Pandian A. and Md. Abdul KarimSadiq, 2012, “Detection ofFraudulent Emails by Authorship Extraction”, InternationalJournal of Computer Application Vol.41, No.7, pp.7 – 12.

[10] Pandian A. and Md. Abdul KarimSadiq, 2013, “AuthorshipAttribution in Tamil Language Email For Forensic Analysis”,International Review on Computers and Software, Vol. 8, No. 12, pp.2882-2888, (SNIP: 1.178).

[11] M.Mahalakshmi, MalathiSharavanan, “Ancient Tamil ScriptRecognition and Translation Using LabVIEW”, IEEE, 2013,International conference on Communication and SignalProcessing, April 3-5.

[12] FarkhundIqbal, HamadBinsalleeh, Benjamin C.M. Fung,MouradDebbabi, 2010, “Mining writeprints from anonymous e-mails for forensic investigation”,Digital Investigation(Elsevier),Vol.7,pp.56-64

[13] Bagavandas, M., Hameed, A., Manimannan G, 2009, “NeuralComputation in Authorship Attribution: The Case of SelectedTamil Articles”, Journal Quantitative Linguistics, Vol.16, No.2, pp.115-131.

[14] R Chandrasekaran and G Manimannan, 2013, “Use ofGeneralized Regression Neural Network in AuthorshipAttribution”, International Journal of Computer Applications, Vol.62, No.4, pp.7-10.

[15] Pandian A. and Md. Abdul KarimSadiq, 2014, “A study ofAuthorship Identification Techniques in Tamil Articles”,International Journal of Software and Web Sciences, Vol. 7 No.1, pp.105-108.