International Journal of Engineering & Technology, 7 (4.19) (2018) 407-411

International Journal of Engineering & Technology

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

Research paper

Analysis of Writer Styles in Punjabi

A. Pandian[1], StephenWahi[2], Yash Tokas[3], K. Manikandan[4], V.V.Ramalingam[5]

[1,5]Associate Professor, Department of CSE, SRM University, Kattankulathur.
[2,3]UG Student (B. Tech.), Department of CSE, SRM University, Kattankulathur.
[4]Assitant Professor (S.G), Department of IT, SRM University, Kattankulathur.
**Corresponding author E-mail: pandian.a@ktr.srmuniv.ac.in*

Abstract

Author Identification alludes to the issue of distinguishing the creator of a mysterious content. From the machine learning perspective, this is a solitary mark content arrangement assignment. This errand is done on the supposition that the creator of an obscure content can be separated by looking at a couple of lexical highlights extricated from that obscure content with those of writings having known writers. In this paper, Authorship Identification process is connected on Punjabi verse dataset comprising of Punjabi ballads composed by 5 unique writers. Different highlights extensively ordered as measurable (word-check, roast tally, and so forth.), linguistic (i.e. lexical) and semantically (dialect subordinate) are first chosen utilizing the J48 Decision Tree Algorithm. They chose highlights are thusly, utilized as a contribution to the J48 classifier and the approval of the proposed framework is assessed based on Precision, Recall, F-score and Accuracy.

Keywords: *Authorship Identification, Punjabi poetry corpus, Feature extraction, J48 Decision Tree, J48 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 are recognized. Identifying them would be of more use to the people.

Thus, by using a sensible computational procedure, makers of the unidentified lyrics may have an opportunity to be found for their unaccounted works. Thomas Bayes (1871) used quantifiable speculation for finding issues with recognizable proof of creation in the federalist papers. Auguste de Morgan" (1851)had proposed the word mean length as a factor to choose origin of an article. Seeing those producers of a refrain on the assistance from ensuring complex characters is the writer attribution issue anchored close-by etymological" examination. Finishing trademark extraction may help however that is just a hint of a greater challenge with this creation attribution, which incorporates separating a genuine and just those each sometimes utilized Characteristics in words, period for sentence, earth shattering characters utilized, length about articulations and so on.

In reference [1], numerous segments are investigated "that can be utilized as ascribes to highlight extraction from datasets. Enron E-mail was the dataset utilized and characterization was finished utilizing expectation– boost calculation and bisecting K-implies calculation" giving a 90 % exactness.

In reference [2], order of segments express to the Tamil Language was finished utilizing calculations like help "vector machine, proximal help vector machine and arbitrary kitchen sink calculations". SVM performs order by making two disjoint spaces and arranging each section as one of the two, while Proximal SVM First assigns server farms to the closer of the two parallel lines and characterizes the dataset in like manner. Irregular Kitchen Sink figuring utilizes all the conceivable free factors and produces a quantifiable check. The precisions achieved are 95.7%, 95.8%and 96.82% individually.

In reference [3], a precision of 87.5% is accomplished by us in grandom timberland "calculation on 86052 words and 500788 characters."

In reference [4], a precision of 82% is refined on Arabic ballads, which uses SVM, neural systems and Markov chain as classifiers for information.

In reference [5], particular highlights are removed from a Tamil dataset that contains roughly 5000 words. Classifiers produce "an exactness of 72% to 82%. FLD and RBF calculations are utilized to vanquish the conflicting issue. FLD calculation performs gathering by making a straight mix of parts that limits no under two classes of things. Extended Basis" Function figuring is essentially a vague neural framework structure. It works in setting of neuron parameters.

In reference [6], an Arabic dialect dataset is utilized. Arrangement is performed utilizing the Markov "chain calculation producing an accuracy of 96.96%. The best way to deal with concentrate highlights appropriate to the Arabic tongue is illustrated. Each part that is identified with the dataset and that additionally fulfills the characterized Markov property is a legitimate unit that can be utilized for arrangement. These components are picked consequently used to assemble the classifier.

In reference [8], the issue of creation recognizable proof of old Tamil contents is handled. These contents are first digitalized, and afterward characterization is performed utilizing SVM Classifier and uni-gram, bi-gram highlights which results in a precision of 83%.N-grams are frequently utilized when the information is talk or a substance corpus. Uni-gram is a size one n-gram and bi-gram is a size two n-gram.

In reference [9] , the writer scatters the covering issue utilizing the Fisher's Linear Discriminant and Radial Basis Function

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

calculations on the Enron email dataset, while in 10, segments are gathered with the end goal to translate the inception of a specific article from the Enron email dataset by utilizing winding reason count for gathering in with an exactness of 80% to 90%."

In reference [11], the author discussed about Authorship Identification for Tamil Classical Poem using Subspace Discriminate Algorithm.

Finding the writers for un-created Punjabi compositions get is an especially troublesome assignment as there is no framework to remember them" expressly. By separating highlights relating to the Punjabi lingo utilized in its lyrics and by utilizing reasonable computation, essayists for these un-composed sonnets can be seen. Fig. 1 exhibits the engineering that is followed in this system of characterization.

2. Materials and Method

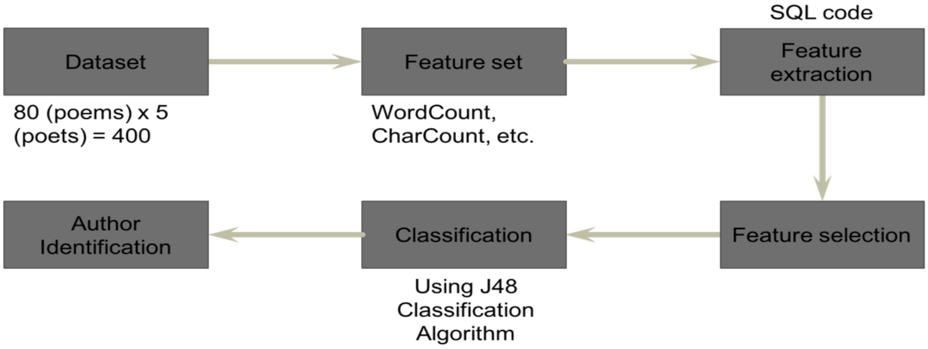

Fig. 1: Architecture Diagram

The datasets used here is 80 poems each of 5 eminent Punjabi poets namely Baba Bulle Shah, Bawa Balwant, Bhai Vir Singh, Fazal Shah, Sultan Bahu and Ustad Daman. The poems for these 5 poets is extracted from different sites like http://www.punjabi-kavita.comand http://www.shivbatalvi.com. By extricating syntactic, lexical and semantic elements as in [8], classification is performed. Main features that are considered are depicted in Table 1.

These highlights are isolated from the dataset and are furthermore used for the gathering methodology. These highlights depict the stylometry of the maker. Stylometry is the basic differentiation in made masterful styles out of different columnists. It contains "lexical, syntactic and semantic components pertinent to the specific tongue. Table - 1 depicts the lexical, syntactic and" factual highlights removed from the dataset. The J48 estimation, gave a precision of 82.66%.

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. Ending Akhar (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 with in each dataset. The 35 features are explained briefly with broad categorizations.

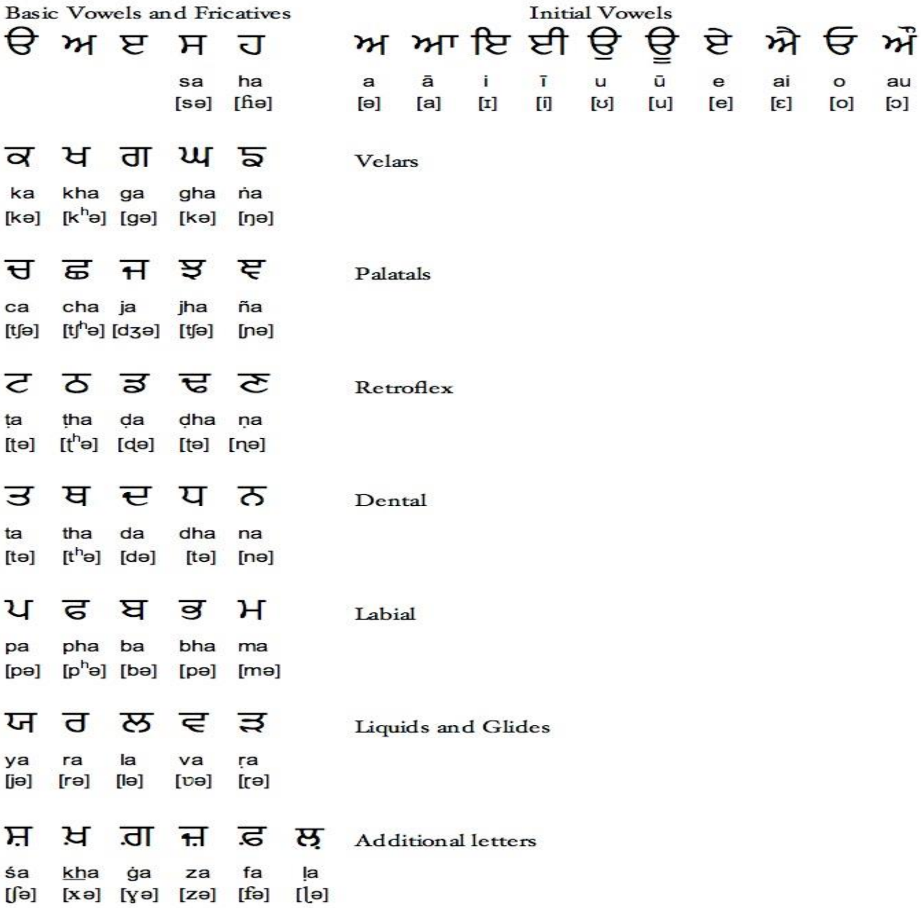
Basic Vowels and Fricatives
ੳ ਅ ੲ ਸ ਹ
sa ha
[sə] [ɦə]
Initial Vowels
ਅ ਆ ਇ ਈ ਉ ਊ ਏ ਐ ਓ ਔ
a ā i ī u ū e ai o au
[ə] [a] [ɪ] [i] [ʊ] [u] [e] [ɛ] [o] [ɔ]
ਕ ਖ ਗ ਘ ਙ
ka kha ga gha ṅa
[kə] [kʰə] [gə] [kə] [ŋə]
Velars
ਚ ਛ ਜ ਝ ਞ
ca cha ja jha ña
[tʃə] [tʃʰə] [dʒə] [tʃə] [ɲə]
Palatals
ਟ ਠ ਡ ਢ ਣ
ṭa ṭha ḍa ḍha ṇa
[ʈə] [ʈʰə] [ɖə] [ʈə] [ɳə]
Retroflex
ਤ ਥ ਦ ਧ ਨ
ta tha da dha na
[tə] [tʰə] [də] [tə] [nə]
Dental
ਪ ਫ ਬ ਭ ਮ
pa pha ba bha ma
[pə] [pʰə] [bə] [pə] [mə]
Labial
ਯ ਰ ਲ ਵ ੜ
ya ra la va ṛa
[jə] [rə] [lə] [ʋə] [ɽə]
Liquids and Glides
ਸ਼ ਖ਼ ਗ਼ ਜ਼ ਫ਼ ਲ਼
śa k͟ha ġa za fa ḷa
[ʃə] [xə] [ɣə] [zə] [fə] [ɭə]
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

Highlight extraction is worried about gathering a plan of got characteristics from the basic course of action of data relating to human interpretation. Datasets can't particularly be used as a contribution to classifiers, i.e. preparing the information. Highlights are removed from the information to shape a Feature Set, and that thusly, must be used to gather the classifier. This classifier that is constructed is then used to play out the characterization procedure on the Feature Set close by.

Three kinds of highlights are removed, i.e. lexical, syntactic and measurable. Lexical highlights incorporate classifications, for example, thing, verb, descriptive word, and pronoun. Syntactic highlights incorporate thing phrase, verb express and prepositional expression.

Despite these features, quantifiable features are in like manner expelled "from the dataset. Quantifiable features record to a significant bit of the classifier accuracy. The classifier precision has extended from 86% to 90% by including quantifiable features to the features set and" playing out a couple of changes in the figuring used. Authentic features fuse Minimum, Maximum, Sum, and Mean.

The "features recorded in table-1 are removed from the dataset. The dataset is at first changed over to Unicode so it might be controlled easily using SQL. PCs can't get a handle on Punjabi characters. They manage just numbers in memory. A Unicode requesting change over each character of the regional tongue and gives an approach to manage PCs to comprehend them". The extraction strategy is finished by using SQL directions, which can remove the foreordained highlights subsequently. Continuation Pro is used to make a database with each one of the sonnets and segments. The separated highlights are in numeric configuration.

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.

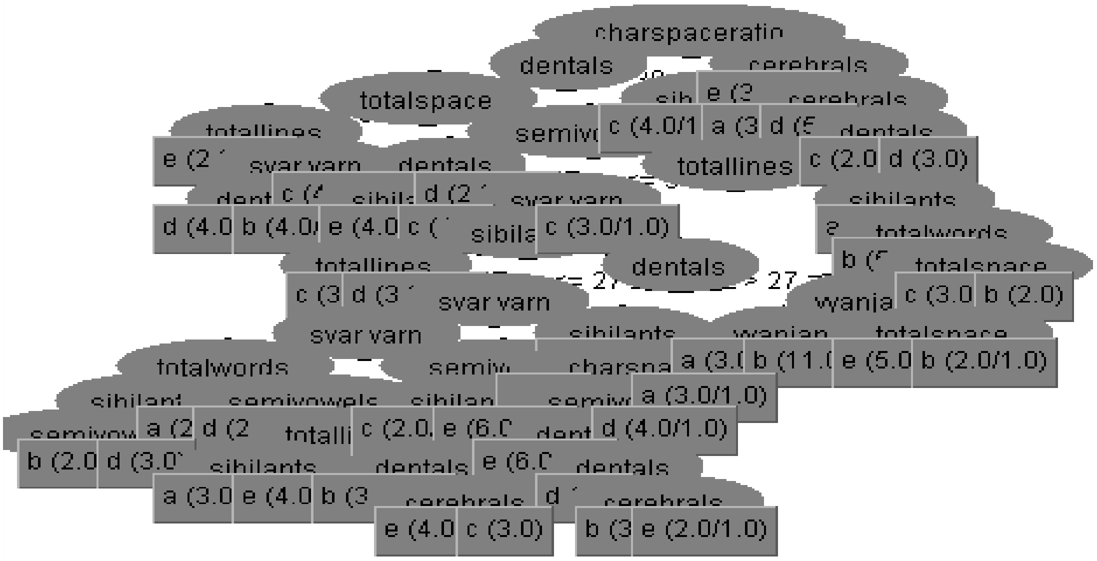

Fig. 2: Decision Tree Construction Using J48 Algorithm

Table 3: 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 calculation is created by Ross Quinlon. This calculation will be an advancement of the ID3 calculation that may have been being utilized sooner times. C4. 5 calculation builds a decision tree. Following are the means of the calculation:

"1. Check for the base cases.
2. For each" characteristic x, split on x and discover the data gain.
3. Give the most elevated data a chance to pick up characteristic be x1.
4. Make a hub that parts on x1.
5. Utilize the subsets of x1 to emphasize a similar procedure and include every one of the hubs as offspring of x1.

3. Results and Discussions

The outcome for the proposed classifiers demonstrate is delineated as a perplexity lattice. The perplexity network acquired by performing characterization utilizing J48 calculation is appeared in Table - 4. It very well may be seen that 17 occasions of An are effectively named A, while 4 cases have been wrongly classified. 28 examples of the creator B are accurately characterized, while 2 occurrences are wrongly arranged as author C's work. The author C seems to have a total of 26 instances out of which 17 are correctly classified as C and 9 instances are wrongly classified. The author D has a total of 17 instances out of which all 11 instances are correctly classified. The instance of author E is wrongly classified.

Table 4: Confusion Matrix

	A	B	C	D	E
A	17	1	1	2	0
B	0	28	2	0	0
C	2	1	17	6	0
D	1	0	5	11	0
E	0	0	0	1	0

The classifier built using J48 algorithm provides an accuracy of 76.84%. To improve the classifier accuracy, two parameters: Confidence factor and Minimum number of objects are considered. After performing some tweaks, the classifier accuracy is improved to 82.6%. The classifier accuracy reaches 82.6% when the minimum number of objects is between 2 and 7, the accuracy drops beyond 7. Similarly, the confidence factor is varied and the corresponding accuracy is recorded. The confidence factor remains the same for all values.

4. Conclusion

The highlights recorded in table-1 were considered and the highlights were chosen by developing a choice tree utilizing J48 calculation. Certain highlights from the rundown of considered highlights list were chosen with the end goal to defeat over fitting of the classifier and furthermore to maintain a strategic distance from the use of non-contributing highlights. The J48 calculation created an exactness of 82.6% on the dataset.

The origin ID prompts a precision of 82.6% on the dataset. Therefore, by removing general highlights that are regular for every provincial dialect, a general initiation ID framework can be created for every single local dialect".

References

[1] Farkhund Iqbal, Hamad Binsalleeh, Benjamin C.M. Fung, Mourad Debbabi, 2015, "E-mail authorship attribution usingcustomized associative classification", Digital Investigation (Elsevier), Vol.7, pp.56-64

[2] Sanjanasri J.P and Anand Kumar M, "A Computational Framework for Tamil Document Classification using Random Kitchen Sink", IEEE 2015, International Conference on Advances 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 and Communication Technology Research (ICTRC2015)

[4] Ahmed Fawziotoom, Emad E Abdullah, Shifaa Jaafar, Aseer Hamdellh, Dana Amer, "Towards Author Identification of Arabic Text Articles", IEEE 2014, 5th International Conference on Information and Communication Systems(ICICS)

[5] Pandian, A., and Md. Abdul Karim Sadiq, 2014, "Authorship Categorization In Email Investigations Using Fisher's Linear Discriminate Method With Radial Basis Function", International Journal 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 Fawzi Otoom, Emad E. Abdullah, Shifaa Jaafer, Aseel Hamdallh, Dana Amer “Towards Author Identification of Arabic Text Articles”, 2014,IEEE, 5th International Conference on Information and Communication Systems (ICICS)

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

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

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

[11] A Pandian, V V Ramalingam, K Manikandan, R P Vishnu Preet. "Authorship Identification for Tamil Classical Poem using Subspace Discriminant Algorithm", Journal of Physics: Conference Series, 2018.