International Journal of Engineering & Technology, 7 (4.19) (2018) 26-31

International Journal of Engineering & Technology

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

Research paper

Author Identification for Telugu Classical Poems

A.Pandian[1], V.V.Ramalingam[2], K.Manikandan[3], V.Jeevan Reddy[4], Pavuluri Sai Krishna[5]

[1] Associate Professor, [2,3]Assistant Professor (S.G), [4,5]B.|Tech Student
[1,2,4,5] Dept. of CSE, SRMIST, Chennai, [3]Dept.of IT, SRMIST, Chennai

Abstract

Artist finding is the errand of distinguishing the creator of a given test from an arrangement of suspects. The free worry of this errand is to characterize a fitting portrayal of test that catches the composition styles of creators. In this task, weka based machine learning instruments are utilized for distinguishing proof of creator for highlight extraction of reports spoke to utilizing variable size character n-grams. We wrote our own java program to extract the features like number of words, sentences etc. From, the poem which in turn fed as input to weka tool for the identification of author then after testing the input with all the algorithm all the accuracy rates are noted down to see which algorithm is given us the best accuracy rate. Now to find the author name for an anonymous poem the poem features are extracted using the java code and the output is taken in the java file given to the weka tool and tested with the algorithms and then the author name is given to the anonymous poems.

Keywords: *Author Attribution, Stylometry, Telugu dataset, Natural Language Processing, Word n-gram, Char n- gram.*

1. Introduction

Creator distinguishing proof is the undertaking of recognizing who composed a given bit of content Newcomer clarifies of a given arrangement of applicant creators (suspects). From machine learning point of view, it can be seen as multiclass single-name content grouping errand where creator speaks to a class (name) of a given content. The dismember of stylometry and origin backpedals to the nineteenth century, with Mendenhall leading the pack by portraying the style of various creators through the recurrence conveyance of expressions of different lengths. Through the primary portion of the twentieth century, numerous factual examinations were taken after presenting measures for composing styles including Zipfs conveyance and Yules K measure.

Concurrent initiation ID began by Most teller and Wallace deal with the federalist papers, where they connected Bayesian measurable examination on the frequencies of a little arrangement of capacity words (e.g "and", "to", "the"), as expressive highlights of test. In the dames letters numerous highlights have been proposed to catch expressive highlights including vocabulary wealth measures, linguistic highlights, work words frequencies and character n-gram frequencies. Abysm learning has been effectively connected to different normal dialect handling assignments delivering execution comes about beating beforehand cutting edge procedure. For example, connected profound learning on the space adaption of assessment investigation by utilizing abnormal state highlight portrayal extricated utilizing profound neural systems and outflanked the condition of workmanship techniques on the arrangement assignment. And also, profound with the fast advancement of data, more correspondence and capacity of reports is performed carefully. An impressive extent of business documentation and correspondence, Degree, still takes puts in physical shape and the fax machine stays indispensable apparatus of correspondence around the world. Because, of the way that of this, optical character acknowledgment (OCR) is ending up increasingly imperative. In any case, all the current takes a shot at OCR make a critical certain suspicion that the content and dialect of the archive to be handled is known.

Common mediation in recognizing the content and dialect of report in managing gigantic pictures can't fulfill the prerequisite of speed and computerization. History, content distinguishing proof, by method for the front handling innovation of OCR framework, is basic and noteworthy. Author identification can be seen as a classification problem of texts:" Given a set of documents written by a same author, set can be large or composed of only one element, we have to decide if a new document has been written by the same author as the others". We have to solve a problem of classification having as response a binary value ("yes" or "not") r a probability to belong to the set of known documents. However, one of the specificities of this problem is that only elements belonging to one of the two possible classes are given: the documents having the same author, but the second class are not explicitly described. Moreover, sometimes the number of positive examples is reduced to only one document and, the task becomes much more difficult. To mitigate the absence of negative examples, one can try to produce some of them. This way is explored by different author among which Seidman who builds a class of impostors randomly chosen on the web on the basis of ten more frequent words in the available documents. Other authors, like Zhang et al and Halvani transform this problem of classification with two classes into problem with several classes, either by adding external classes or by dividing the initial classes into several. These same authors increase the size of the class containing the know documents when this last one is reduced to only one. Thus, these approaches allows to transform the problem into a classical from of classification, but during the construction of the set of negative examples there is a risk to take some documents very different from the known documents.

It us widely acknowledged that people around the world are increasingly using the computer technologies and computer-mediated communications to connect with each other. The internet's seamless accessibility and user-friendly platform have revolutionized the sharing of information and communication, facilitating an international web of virtual communities.

2. Literature Review

In reference [1] the poet show the method to extract features from Tamil dataset that contains character count of 28420 and word count 5000 with an precise of percentage (72 to 82). It disperses the over lapping problem by using FLD and RBF algorithms. Work shows the era of Authorized signatures of emails in Tamil utilizing features mainly syntactic, lexical. For purpose to make it easy final activity, converting large dimension of signature into 2D pattern using FLD algorithm. the converted 2D pattern used for train the data for RBF and ESNN network. The enhanced classification of email in tamil is given by changeover of patterns using Fisher's linear discriminant algorithm along with training of RBF and training of ESNN.

This is the advanced method for developing signature database and for perfect Author attribution in tamil mail forensics with a precision of 80 to 90% .

In reference [2] the author made the possibility of identifying authorship on classical tamil poems. The dataset of authors with using a different algorithm, Bayes Net and have got good accuracy of 90%.

In reference [3] the author attribution is based on data compression model. They have taken six different compression models that mainly Zip, BZip, GZip, LZWP, PPM and PPMd combines with three different compression distance measures such as NCD, CDM, CCC. Combining these compression models will be a great help for the authorship attribution rather than classifying the model with many features

In reference [4] in this the author collected 3000 poems written by three Bengali authors. the author used three classifiers from weka i.e Naïve bayes, SVM SMO and J48 decision tree for testing the performance on the development set. By this classification the accuracy rate for the Naïve bayes is highest than the two other algorithms. As Naïve bayes was faster to train the machine than using the SMO. The accuracy rate where always between 95% and 99%.

In reference [5] the author demonstrates that character based features are better than the word based features and they noticed a failure for the classifier using word based features (attribution score about 10% and 70%. According to them character tri-grams are best. By using this method, the performance of the classifier is good even with small texts Bengali authors works are used to perform classification using Naïve Bayes Classification and N-gram namely bi-gram. The author has attained a classifier accuracy of 98%. n-gram which is sized one is a unigram and an n-gram which is sized two is a bi-gram. The n-grams typically are gathered from content corpus.

In reference [6] the author has used the LINGO algorithm to categorize the words from the input document. They used the LINGO algorithm as it was the best for their regional language.

In reference [7] here the author has been using the XPCFG model as it separates the production rule into two sets and it also adds lexical and syntactic features by capturing the non-terminal, terminal and punctuation. It also helps in assigning the scores to rules to quantify the importance of each rule. The scores are calculated by using chi-square score. The average error rate is said to be 0.128.

In reference [8] here the author has used n-gram for identification of the authors for texts. Here they gained 99.1% result for Bi-gram and 99.6% for Tri-gram. These two bi-gram and tri-gram can be used predict the author for a specific text.

In reference [9] here the author unveils that n-gram is the best suited feature for representing stylometry profiles of small size text which indeed is help for authorship attribution.

In reference [11] here the author has proven authorship identification is possible on Tamil Classical Poems and have achieved an accuracy of about 75% using the specified algorithm and again the same authors.

In reference [12] here the author note the similarity of four different author identification methods that is chi square method, delta method, Z-score method, Kullback liebler divergence method and by the results of aforementioned techniques of each macro average, micro average gives greater value as by K.L.D method which is value sufficiently great in parameter and it performs greater than remaining.

In reference [13], the poet deals the covering issue utilizing f and radial basis algorithm by using Enron email dataset, while in [10], the author uncovers how to concentrate components to find the origin of an article by using spiral premise calculation for grouping in Enron email dataset with a precision of 80% to 90%.

In reference [14], the author exhibits the strategy to identify authors from Email dataset (200399 mails) with an accuracy of 90%. It uses Zip, GZiP, NCD, CCC algorithms to beat the covering issue.

In reference [15], the author presents the way to take out attributes and identify the precise of the classifier that arises. Using Tamil scripts and upon application of various algorithms like SVM, SVM+Bi grams has achieved an accuracy of 78%-83%.

3. Material and Methods

Finding the authors for un-authorized poems in Telugu get the chance to be particularly troublesome because there is no machine for recognizing the un-authorized poems interestingly. By taking these attributes important for Telugu writings and usage of suitable calculations, writers for these obscure writings can be perceived. Grouping is done by utilizing content handing procedure. Content handling is the system for getting first class information from substance that consolidates true cases from the substance.

The database contained here is 10 shathakas from 10 great Telugu poets namely Buchana, Pakki Venkata Narasimha, Bharthur Hari, Swami Parmanandha, Gopitham, Sadanandha Yogi, Kancharala Gopana, Dasu Sreeramulu etc. Each shataka consist of 108 poems written by poets that dates back to several years.Out of 836poems 418 for the training the machine and 418 for testing. By taking types of attributes such as statistical type, Syntactic elements as clarified in the grouping process is performed. The rundown of features that are considered is listed in table-A.

The attributes are taken from the data set and used for doing classification. These attributes characterize the statistical analysis of variations in literary style of the creator. Stylometry is for examining composed literary styles from manually written poems that can be used as part of finding the author for un-authorised poem. Statistical analysis of variations in literary style of writer (stylometry) incorporates taking of lexical, syntactic, statistical attributes that are take out from the database table. Using C4.5 algorithm, an precise of 87.4% was achieved.

The J48 set of rules includes two parameters, confidence component and minimal quantity of items. those factors should be various that allows you to reap a few differences inside the accuracy. The confidence aspect need to be various from 0.1 to 1. zero at the same time as the minimum range of items need to numerous from 1 to variety of functions taken into consideration. After performing the tweaks, the final accuracy performed is 87.6% self-belief factor being 0.2 and minimum variety of items being four, the peak accuracy turned into carried out.

3.1. Feature Extraction

Standpoint extraction handle assembles an arrangement of created qualities are from the underlying arrangement of information that is planned to human translation. Dataset cannot be specifically utilized as a part of the tool to perform arrangement. Without equal the features that are extricated from the dataset from the dataset can be utilized to assemble the classifier. This classifier that is built is then used to perform the classification process on the dataset in hand. Duotypes of attributes, Unembellished, syntactic and Unbiased are taken. Lexical attributes Incorporate class such as noun, verb, adjective, and pronoun. Syntactic features include noun phrase, verb phrase and prepositional phrase. Semantic attributes are those that include a set of features that intensifies the meaning of a word. In addition to these attributes, statistical attributes are also taken 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 and mean.

The attributes recorded in table-A are taken from the database table. The database table is initially changed over into unicode format so it can perused in Microsoft excel. Systems(Pc) not comprehend Telugu features. They bargain just with numbers in their memory. Unicode gives the conversion of information or data into code framework that includes all the languages 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. Sqlite browser 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.

4. Implementation of Classification Algorithm

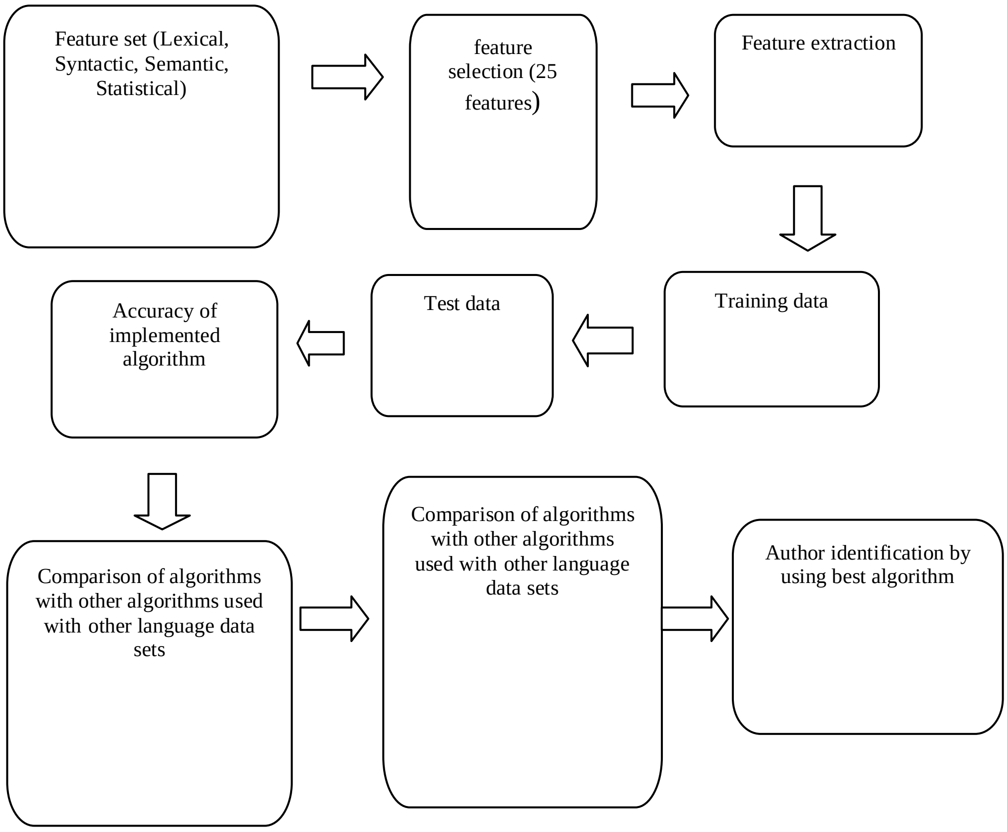
The algorithms chosen and were used for implementing on the dataset in hand. These algorithms are already proven to have given a decent accuracy on various other datasets. The implementation process was performed by the use of two tools weka.

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 has to be selected.

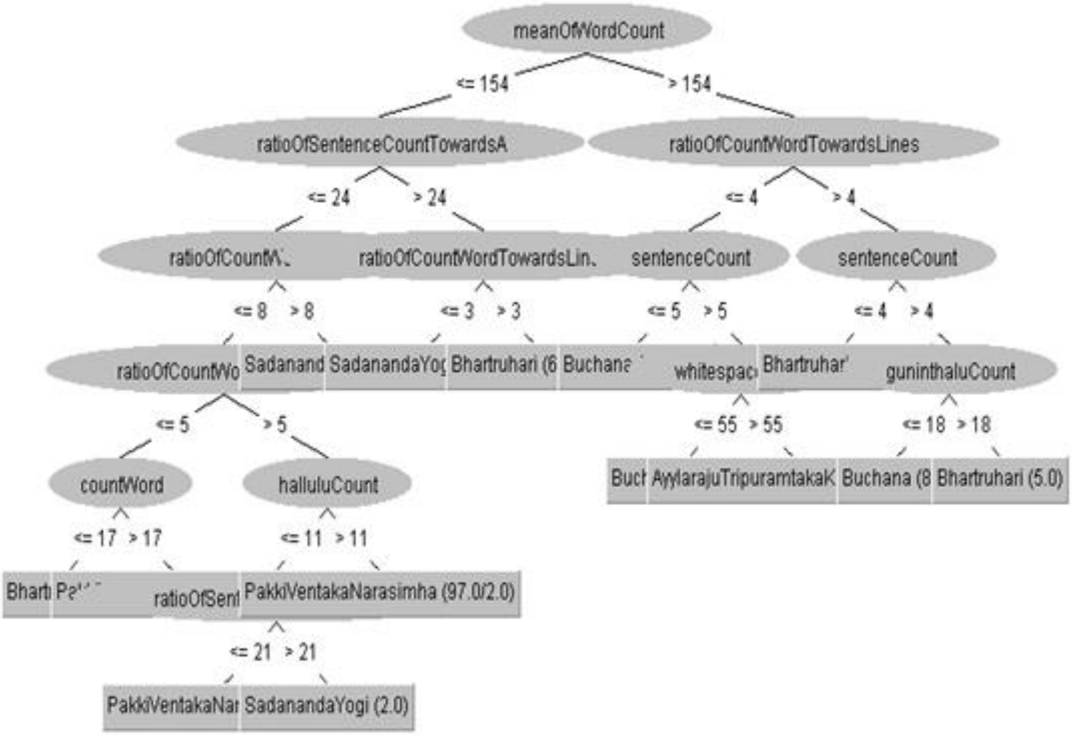
Table A: (list of Attributes)

Attribute type	Attributes
	• Count Word
	• Sentence count
	• Character count
	• Paragraph count
	• White space count
	• Occurrence of achulu, halulu, gunithalu, vothulu
Statistical features	• Mean of Word Count, Median of Word Count, Mode of Word Count
	• Ratio of Count Word TowardsA
	• Ratio of Sentence Count TowardsA
	• Ratio of Character Count TowardsA
	• Ratio of Paragraph Count TowardsA
	• Ratio of White Space Count TowardsA
	• Ratio of Count Towards Lines
	• Ratio of Sentence Count TowardsB
Syntactic features	• Ratio of Character Count TowardsB
	• Ratio of Paragraph Count TowardsB
	• Ratio of White Space Count TowardsB

4.1. Feature Selection


Feature set (Lexical, Syntactic, Semantic, Statistical)
feature selection (25 features)
Feature extraction
Training data
Test data
Accuracy of implemented algorithm
Comparison of algorithms with other algorithms used with other language data sets
Comparison of algorithms with other algorithms used with other language data sets
Author identification by using best algorithm

4.2. Decision Tree

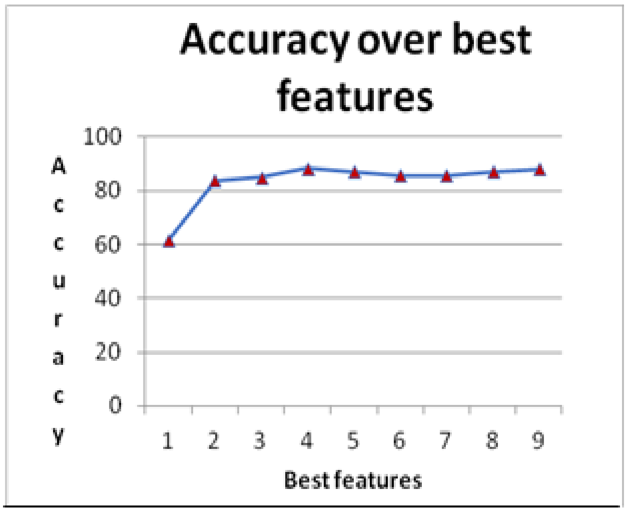


The below graph shows accuracy over 9 best attributes using C4.5 algorithm above one. Best attributes are selected from decision tree algorithm in the Dataset.

For example

1.Taking author as X ,each selected features as x1,x2,x3,x4,x5,x6,x7,x8,x9 .
2. X+x1 is Classified with J48 gives A1.
3.X+x1+x2 is (J48) =A2.
4.X+x1+x2+x3 (J48)= A3.
5.X+x1+x2+x3+x4 = A4.
6.X+x1+x2+x3+x4+x5 = A5.
7.X+x1+x2+x3+x4+x5+x6 = A6.
8.X+x1+x2+x3+x4+x5+x6+x7 = A7.
9.X+x1+x2+x3+x4+x5+x6+x7+x8=A8.
10.X+x1+x2+x3+x4+x5+x6+x7+x8+x9=A9

A1,A2,A3,A4,A5,A6,A7,A8,A9 are the achieved accuracies classified over selected features from J48 Decision tree and graph is made using Excel.



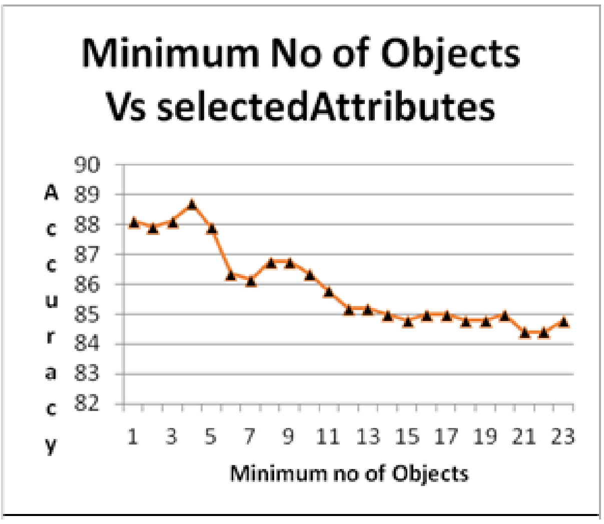
The selected attributes is classified over minimum number of objects that means total number of attributes used in our Dataset.

Here 9 attributes as shown Graph-1 are selected for classifying and for each classification minimum number of objects is chosen from 1 to 23 that gives accuracy for each change of MNO (minimum number of objects) and confidence factor is kept constant i.e 0.25.

For example:

X(Author),selected attributes(x1+x2+x3+x4+x4+x6+x7+x8+x9, minimum no of objects (M1+M2+M3+M4+M5+M6----------+M23) and accuracy as (A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W)

1.X+x1+x2+x4+x5+x6+x7+x8+x9+M1=A
2.X+x1+x2+x4+x5+x6+x7+x8+x9+M2=B
3.X+x1+x2+x4+x5+x6+x7+x8+x9+M3=C
4.X+x1+x2+x4+x5+x6+x7+x8+x9+M4=D
5.X+x1+x2+x4+x5+x6+x7+x8+x9+M5=E
6.X+x1+x2+x4+x5+x6+x7+x8+x9+M6=F
7.X+x1+x2+x4+x5+x6+x7+x8+x9+M7=G
8.X+x1+x2+x4+x5+x6+x7+x8+x9+M8=H
9.X+x1+x2+x4+x5+x6+x7+x8+x9+M9=I
10.X+x1+x2+x4+x5+x6+x7+x8+x9+M10=J
11.X+x1+x2+x4+x5+x6+x7+x8+x9+M11=K
12.X+x1+x2+x4+x5+x6+x7+x8+x9+M12=L
13.X+x1+x2+x4+x5+x6+x7+x8+x9+M13=M
14.X+x1+x2+x4+x5+x6+x7+x8+x9+M14=N
15.X+x1+x2+x4+x5+x6+x7+x8+x9+M15=O
16.X+x1+x2+x4+x5+x6+x7+x8+x9+M16=P
17.X+x1+x2+x4+x5+x6+x7+x8+x9+M17=Q
18.X+x1+x2+x4+x5+x6+x7+x8+x9+M18=R
19.X+x1+x2+x4+x5+x6+x7+x8+x9+M19=S
20.X+x1+x2+x4+x5+x6+x7+x8+x9+M20=T
21.X+x1+x2+x4+x5+x6+x7+x8+x9+M21=U
22.X+x1+x2+x4+x5+x6+x7+x8+x9+M22=V
23.X+x1+x2+x4+x5+x6+x7+x8+x9+M23=W



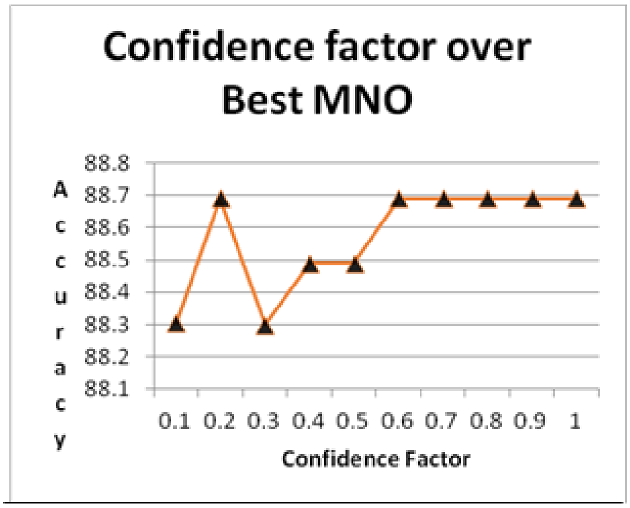
Here in graph-3 the accuracy is obtained by changing the confidence factor with fixed minimum no of objects that is the greatest accuracy that occur in graph 2 and confidence factor varies from (0.1-1.0) classified with J48 algorithm.

the minimum no of objects over selected attributes in graph 2 take the highest accuracy with which minimum no of object had occurred and keep that minimum no of object as fixed value as some (Y) and vary confidence factor from 0.1 to 1.0

x1+x2+x3+x4+x5+x6+x7+x8+x9 is with varying CF 0.1 to 1 and keeping MNO(Y) as constant and classified using J48 and accuracy is achieved that plotted in below.

4.3. The Best Attributes are as Follows

1.Mean of Word count
2.ratio of sentence count towards A
3.ratio of count word towards lines
4.count word
5.hallulu count
6.sentence count
7.white space count
8.guninthalu count



5. Results and Discussions

In the above three graphs the y-axis is taken as accuracy and x-axis differs in each graph. In graph -1 the accuracy rate increases as the best features increases it is. In graph-2 the accuracy is at the highest point i.e (88.7) when the confidence factor value is 0.2,0.6, 0.7, 0.8, 0.9 and 1. The accuracy rate is least (88.3) when the confidence factor is 0.1 and 0.3. And it is in the middle i.e (88.5) when confidence factor is at 0.4 and 0.5. In the last graph the x-axis is MNO(minimum number of objects) here the accuracy is completely independent to the MNO here the highest accuracy rate is obtained when the minimum number of objects is 4 and the accuracy rate has begun to decrease when the MNO value is 10 the accuracy rate has begun to fall down continuously till the end. These are all the results obtained on testing.

The j48 calculation takes after a straightforward calculation. To characterize another thing, it first needs to make a choice tree in light of the trait estimations of the accessible preparing information. In this way, at whatever point it experiences an arrangement of things that segregates the different occasions generally unmistakably. This element that can reveal to us most about the information occasions with the goal that we can arrange them the best is said to have the most elevated data pick up. Presently, among the conceivable estimations of this component, if there is any an incentive for which there is no uncertainty, that is, for which the information occasions falling inside its classification have a similar incentive for the objective variable, at that point we end that branch and dole out to it the objective esteem that we have gotten.

For alternate cases, we at that point search for another characteristic that gives us the most noteworthy data pick up. Subsequently we proceed in this way until the point when we either get an unmistakable choice of what mix of qualities gives us a specific target esteem, or we come up short on traits. If we come up short on qualities, or in the event that we can't get an unambiguous outcome from the accessible data, we dole out this branch an objective esteem that most of the things under this branch pos-sess.

Since we have the choice tree, we take after the request of characteristic determination as we have acquired for the tree. By checking all the separate qualities and their esteems with those found in the choice tree display, we can appoint or pre-dict the objective estimation of this new occasion..

6. Conclusion

After doing the necessary literature survey, we have come across the research papers mentioned. We can conclude that J48 algorithm gives the best accuracy rate of 88.69% than other algorithms. The results shown are more accurate as compared to other algorithms. The quality and quantity of features generated directly affect the success of the classifying algorithm.

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.

The addition of the frequency of words to feature set has brought a great difference to the accuracy rate. The frequency of words is a stand out feature.

References

[1] Dr.A. Pandian and M.A.K. Sadiq (2013). “Authorship attribution in email investigations using Fisher’s linear discriminate method with radical basis function,” International Journal of Computer Science.

[2] Dr.A. Pandian, V.V. Ramalingam R. Preet and Dr.R. Varadharajan (2016). “Authorship identification for tamil classical(mukkoodar pallu) using bayes net algorithm.” INDJST..

[3] S. Nagaprasad, P. Vijayapal Reddy and A. Vinaya Babu (2015) “Authorship Attribution based on Data Compression for Telugu text.” International Journal of Computer Applications (0975-8887) volume 110-No.1,january 2015.

[4] Shanta Phani, Shibamouli Lahiri and Arindam Biswas (2015).“Authorship Attribution based on n-grams, feauture selection for bengali language,”.

[5] Itrc.iiit.ac.in/icon2015_proceedings/PDF/37_rp.pdf.

[6] Navinder Kaur and Amandeep Verma (2015), “Authorship Attribution of Punjabi Poetry using SVM Classifier.” kaur et al.,International Journal of advance Research in Computer Science and Software Enginnering 5(5),May-2015,pp.1055-1061.

[7] Aishwarya Sahini, Kaustubh Sarang, Susmitha Umredkar, and Mihir Patil, “Automtic Text Categorization of Marathi Language Documents.” Aishwarya Sahani et al, / (IJCSIT) International Journal of Computer Science and Information Technologies,Vol. 7(5) ,2016,2297-2301.

[8] Ibrahim S.I. Abuhaiba and Mohammed F.Eltibi (2016). “Author Attribution of Arabic Texts using Extended Probabilistic Context Free Grammar Language Model.” I.J.Intelligent Systems and Applications,2016,6, 27-39 Published Online June 2016 in MECS(http://www.mecs-press.org/)DOI: 10.5815/ijisa.2016.06.04

[9] Feryal I. Haj Hassan and Mousmi A.Chaurasia. “Author Verfication of Arabic Language using n-gram analysis method for Classifying text.” 2012 International Conference on Innovation and Information Management (ICIIM 2012) IPCSIT vol.36 (2012) © (2012) IACSIT Press, Singapore.

[10] Shabeeb PK (2017). “Authorship Attribution Technique for Malayalam transcripts based on n-gram model.” International Journal of Innovative Research in Science,Engineering and Technology, Website: www.ijrset.com Vol. 6, Issue 2,February 2017.

[11] Ahmed M. Mohsen, Nagwa M. El-Makky and Nagia Ghanem (2016). “author indentification using deep learning.” Machine learning and applications(ICMLA)15th IEEE international conference.

[12] Pandian, V.V. Ramalingam and R. Preet (2016b). “Authorship identification for tamil classical poem(mukkoodar pallu) using c4.5 algorithm,”INDJST

[13] Parth Mehta, Prasenjit Majumder (2013). “Authorship Attribution based on Optimum parameter selection for K.L.D for Gujarati.” International Joint Conference on Natural Language Processing, pages 1102-1106,Nagoya,Japan,14-18 October 2013.

[14] Panidian. A, V.V.Ramalingam and R.P.Vishnu Preet,2016, “Authorship Identification for Tamil Classical Poem (Mukkkoodar Pallu) using C4.5 Algorithm”,Indian Journal of science and Technology,Vol 9(47),DOI:10.17485/ijst/2016/v9i47/107944,December 2016

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

[16] 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).