TEXT ANALYSIS FOR AUTHOR IDENTIFICATION USING MACHINE LEARNING

[1]Shubhesh Amidwar, [2]Siddharth Baxi, [3]Keerteesh Rao, [4]Sunil Kale
[1]Student, [2]Student, [3]Student, [4]Assistant Professor
[1]Computer Engineering,
[1]Rajarshi Shahu College of Engineering (Pune University), Pune, India

Abstract*—This India is the home to thousands of languages due to its diversity and culture. Usually while referring to literatures, we cannot make out whether we are referring to the original one or not. Another issue that comes into play is that the work of well-known authors is copied as it is which leads to the spread of plagiarism. In this project, we attempt to find a solution for this issue by implementing a system for author identification using machine learning and text data mining in Devanagari script. We implement this system by using modified versions of SVM (Support vector machine), Naïve Bayes, K-NN (K nearest neighbour) by giving the Devanagari script as input. We carry out comparison and analysis of the result of each algorithm which we will achieve in terms of accuracy of identifying the author and showing the result. We conclude this with best suitable techniques and features.

***Index Terms*— Devanagari, K-NN, Machine learning, Naïve Bayes, SVM.**

I. INTRODUCTION

Author identification is a very important problem from point of view of disputed authorship of some literary works. It is also a very well know problem in the field of digital forensics. To solve this problem various methods have been suggested belonging to either statistic dedicated computation or machine learning. The author identification process usually starts with the training phase. In the training phase, used texts of known authors are selected form which Stop words are removed. The next phase is the verification phase where the unattributed texts are compared with the previously computed text using machine learning algorithms. Then from the available set of authors, the one that matches most closely is chosen. In short, author identification involves comparison of writing style against a corpus of texts (text body used for linguistic analysis) of known authorship.

The existing system makes use of statistical analysis for both test sample and the feed sample. It accepts the sample and the system tokenizes them. It finds out distinct tokens along with their frequencies. For storage purposes the existing system uses linked list as the data structure. This linked list contains information about the token string and the frequency of that word. Parsing of the text is done and whenever a token is encountered while parsing it checks into the linked list that the token is already present in the list or not. If the token is present then the count is incremented by one otherwise a new node is created for that token with count = 1. After this, the system finds out P most common tokens from the feed and the test sample used. As the value of P increments, the accuracy of the result increases.

The proposed system makes use of machine learning in contrast to the statistical approach used in the existing system. This system is adaptable because the accuracy of the machine learning algorithm is trained. Since we are using multiple algorithms, we will conclude with the best features as well the best technique for author identification.

II. PROPOSED SYSTEM

The architecture of the proposed system is given as follows:

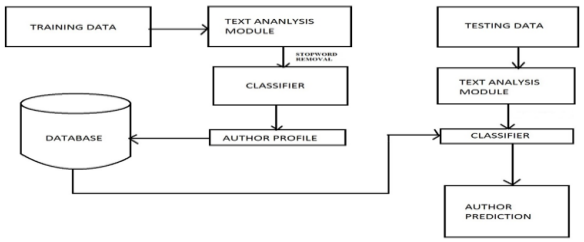
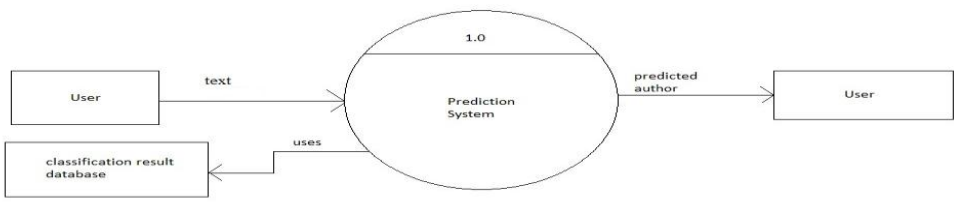


Fig.1

The data flow diagram (level 1) of our system is given as follows:



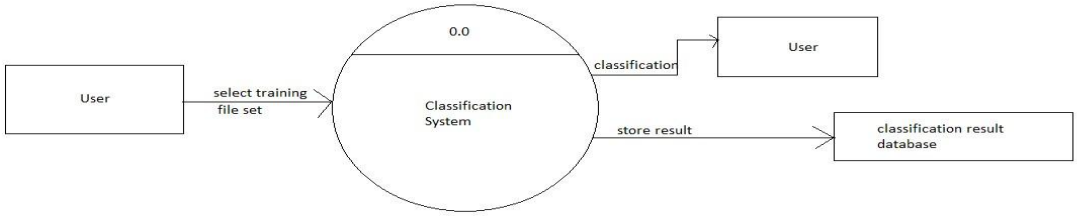


Fig.2

III. IMPLEMENTATION

I. Pre- processing

In this phase, we first remove all the stop-words from the training data set. Stop-words is nothing but sets of commonly used words in any language. The reason why stop words are critical to many applications is that, if we remove the words that are very commonly used in each language, we can focus on the important words instead. In our system, we have already declared over hundred stop-words. The stop-word removal will help us in improving our accuracy.

II. Training Module using Classifier

Separating data into training and testing sets is an important part of evaluating data mining models. Typically, when you separate a data set into a training set and testing set, most of the data is used for training, and a smaller portion of the data is used for testing. Analysis Services randomly samples the data to help ensure that the testing and training sets are similar. By using similar data for training and testing, you can minimize the effects of data discrepancies and better understand the characteristics of the model. After a model has been processed by using the training set, you test the model by making predictions against the test set. Because the data in the testing set already contains known values for the attribute that you want to predict, it is easy to determine whether the model's guesses are correct.

III. Author Prediction

This module consists of 3 algorithms viz. k-Nearest Neighbour, State Vector Machine and Naïve Bayes algorithm. These 3 algorithms will predict the author based on the given test document. Depending on the accuracy of each algorithm, we will select the result.

Snapshot of the User Interface:

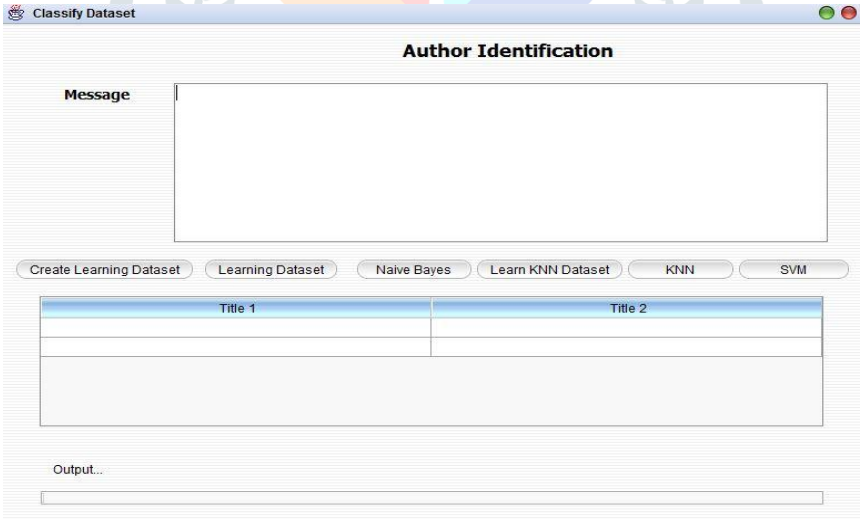


Fig.3

Snapshot of UI of Author Dataset:

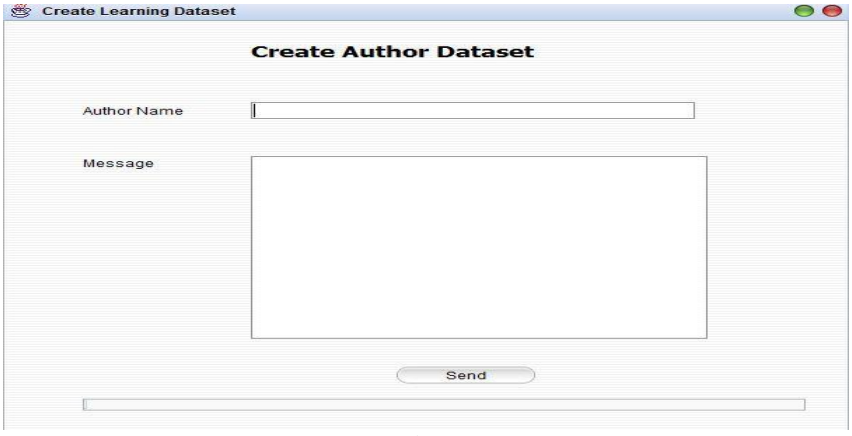


Fig.4

IV. EXPERIMENTAL RESULTS

We now give a text as an input. The following algorithms give the appropriate result.

I. Results using Naïve Bayes Algorithm:

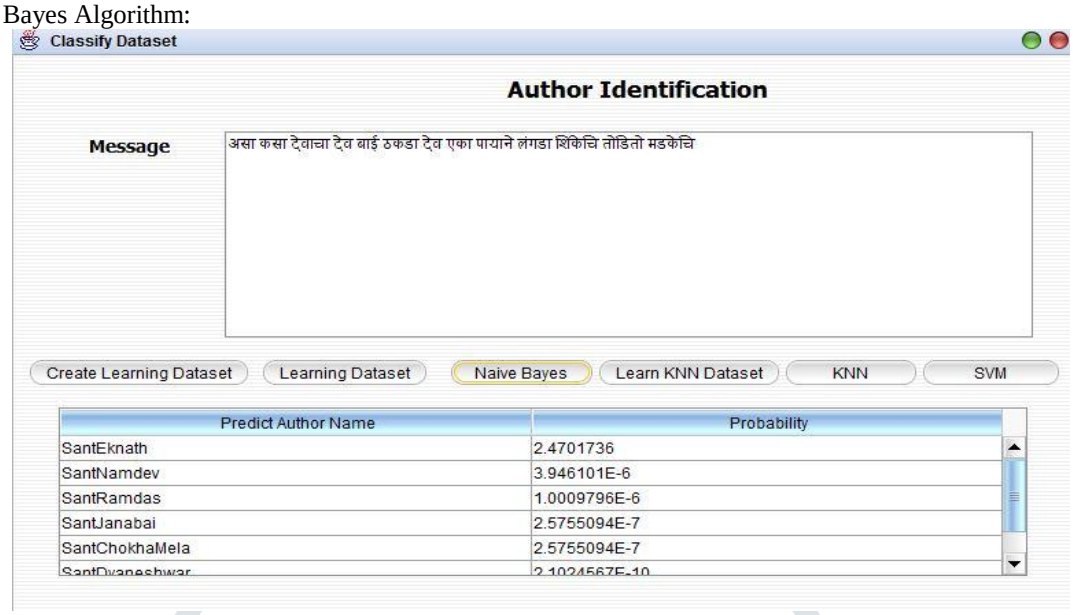


Fig.5

II. Results using K-Nearest Neighbour Algorithm:

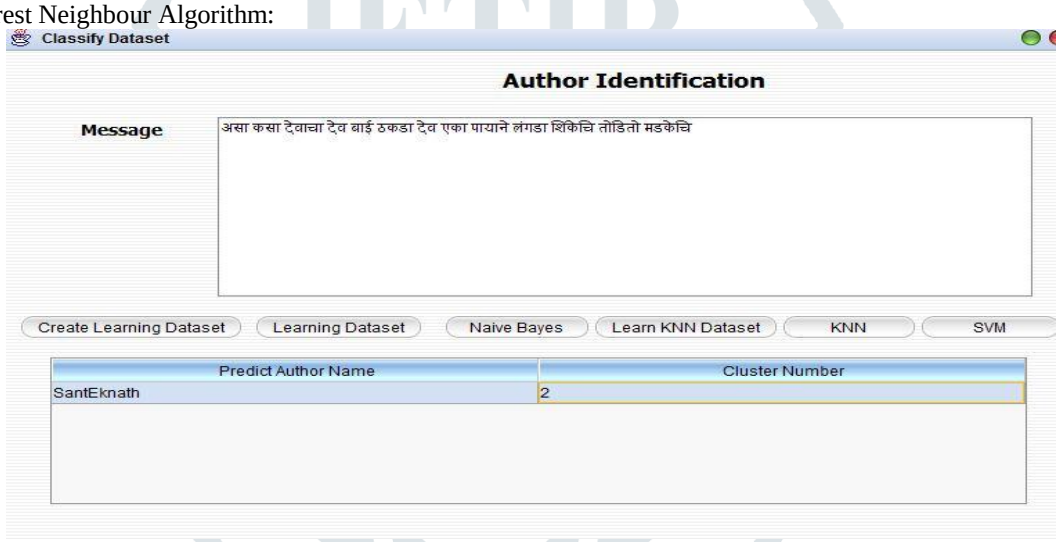


Fig.6

III. Results using State Vector Machine Algorithm:

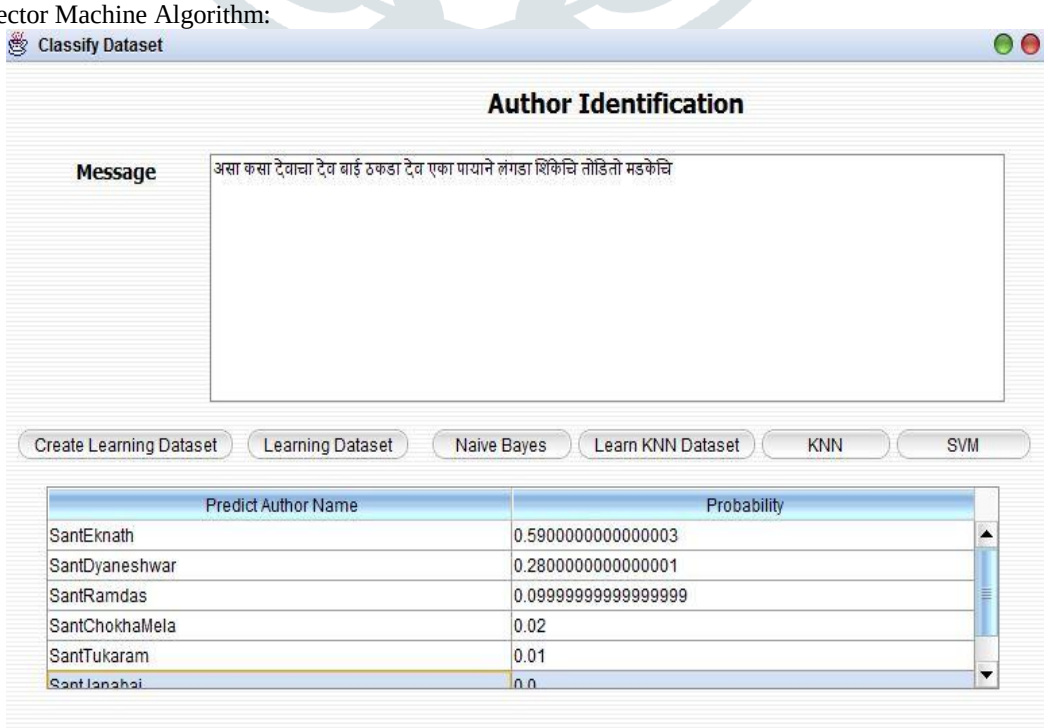


Fig.7

V. CONCLUSION

The project 'Text Analysis for Author Identification using Machine learning' is to build a system which identifies the correct author of the test sample by checking if the pattern of the test sample matches with the style of training data sample for Marathi literature which uses Devanagari script.

The biggest challenge here is reading of PDF files. Hence, text is entered manually. The accuracy of the three algorithms viz. K- nearest neighbour, State Vector Machine and Naïve Bayes is less due to insufficient feature extraction.

The next phase of work will be feature extraction like noun count, verb count, adjective count, etc. The more the features, the more will be the accuracy.

REFERENCES

[1] Satiago Segarra, Mark Eisen & Alejandro Ribeiro, "Authorship Attribution Through Function Word Adjacency Networks", IEEE transactions on signal processing. Vol. 63, No. 20, October 15, 2015.

[2] Rahul Reddy Nadikattu, 2014. Content analysis of American & Indian Comics on Instagram using Machine learning", International Journal of Creative Research Thoughts (IJCRT), ISSN:2320-2882, Volume.2, Issue 3, pp.86-103.

[3] Sikender Mohsienuddin Mohammad, "AN EXPLORATORY STUDY OF DEVOPS AND IT'S FUTURE IN THE UNITED STATES", International Journal of Creative Research Thoughts (IJCRT), ISSN:2320-2882, Volume.4, Issue 4, pp.114-117, November-2016, Available at :http://www.ijcrt.org/papers/IJCRT1133462.pdf

[4] Anderson Rocha, Walter J. Scheirer, Christopher W. Forstall, Thiago Cavalcante, Antonio Theophilo, Binyu hen, Ariadne R. B. Carvalho & Efstathios Stamatatos, "Authorship Attribution for Social Media Forensics", IEEE transactions on information forensics and security, 2016.

[5] Sikender Mohsienuddin Mohammad, "CONTINUOUS INTEGRATION AND AUTOMATION", International Journal of Creative Research Thoughts (IJCRT), ISSN:2320-2882, Volume.4, Issue 3, pp.938-945, July 2016, Available at :http://www.ijcrt.org/papers/IJCRT1133440.pdf

[6] RR Nadikattu, 2016 THE EMERGING ROLE OF ARTIFICIAL INTELLIGENCE IN MODERN SOCIETY. International Journal of Creative Research Thoughts. 4, 4 ,906-911.

[7] Bharati Ganesh H B, Reshma U & Anand Kumar, "Author Identification based on Word Distribution in Word Space", Centre for Excellence in Computational Engineering and Networking Amrita Vishwa Vidyapeetham, Coimbatore, India, 2015.