International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 4 Issue 4, June 2020 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

# Categorization of Protean Writers by Exploitation of Raspberry Pi

**Pritom Sarker[1], Jannatul Ferdous[2], Nakib Aman Turzo[2], Biplob Kumar[3], Jyotirmoy Ghose[4]**

[1]B.Sc., [2]Lecturer, [3]Student, [4]Lecturer Department of CSE,
[1, 2, 3]Department of Computer Science & Engineering, Varendra University, Rajshahi, Bangladesh, India
[4]NBIU, Department of Computer Science & Engineering, NIT Rourkela, Rourkela, Odisha, India

**ABSTRACT**

Raspberry Pi is a computer though smaller but has versatile functionality. These are useful in assisting variety of educational institutions for teaching and other investigational indagations. In this paper, three prominent Bangladeshi writer's works were catalogued by using Raspberry Pi 3. The significance of this research pivots on low cost computational teaching in different institutions. Mathematica was used for this purpose and it's comprised of two modules which effectively communicate and shape an effective interpreter. It's free on Raspberry. After gathering literature and amassing all text files in BD writer's database, application of variety of algorithms were done for categorization. Then experimental analysis was done. Among different categorizers Markov and Naïve Bayes have high precision and have best training times. This research will help in further distinguishing of languages by using an economical approach and will assist in further investigational studies for better understanding.

***KEYWORDS:*** *Bangladeshi Bangla, Inverse Data Frequency, Linear SVM, Principal Component Analysis, Python, Term Frequency, West Bangla*

***How to cite this paper:*** Pritom Sarker | Jannatul Ferdous | Nakib Aman Turzo | Biplob Kumar | Jyotirmoy Ghose "Categorization of Protean Writers by Exploitation of Raspberry Pi" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-4 | Issue-4, June 2020, pp.1104-1109, URL: www.ijtsrd.com/papers/ijtsrd31332.pdf

**IJTSRD31332**



## I. INTRODUCTION

Raspberry Pi is a small sequencing of single board computers and has wide array of application in research because it's economical and is highly portable. It is one of the bestselling British computers. These are utilized for promotion of fundamental computer science teaching in schools and in developing countries. First generation was released in February 2012, then in 2014 the Foundation released a board with improved design. These boards are of credit-card size and represent standard mainline form-factor. Raspberry Pi 2 features 900 MHz quad core ARM Cortex-A7 processor and 1GiB RAM and was released in February 2015. Raspberry Pi Zero with decreased input/output and general-purpose input/output capabilities was released in 2015. Its model Pi 3 Model B was released in 2016. Hardware of Raspberry Pi had evolved in many versions that the differences in characteristics in the central processing unit type, memory capacity and support of networking. Foundation of Raspberry Pi has done all efforts in putting the power of digital making and computing in the hands of common people by providing these highly economical and high-performance computers. By providing education it assists in helping more people to access it. It provides resources to assist in learning. The indagational objective of this research work is to catalogue a line or paragraph of literature in accordance with its writer. Writings of three well-known Bangladeshi writers were used for training this model. Other part of this investigational work was done using model of Raspberry Pi 3. Another point of this research work is done using Raspberry pi 3 as elaborated above a single board economical computer of 25$. This research would prove a hallmark in providing low cost computing for teaching purposes for educational institutions i.e school and college level machine learning teaching classes. An already installed Mathematica with Raspbian was used for the research. Mathematica is a computational programming tool utilized in maths, engineering and science. This was first out in 1988. It's primarily utilized in coding projects in institutions. The dataset was gathered from different websites. Four novels of each writer named Rabindarnath Tagore, Sharatchandra and Kazi Nazrul Islam were took for the purpose of training and testing purposes.

## II. LITERATURE REVIEW

In a research utilization of functionalities of Mathematica was done which was freely accessible on the Raspberry Pi for the purpose of Multi-label classification algorithm with low cost. Random Kitchen Sink algorithm improved the accuracy of Multi-label classification and brought improvement in terms of memory usage [1].

Multiclass categorization problem was resolved by the utilization of Support vector machines trained on the timbral textures, rhythmic contents and pitch contents. Different experiments on various musical sounds were done. Catogorization into ten adjective groups of Farnsworth as well as classification on six subgroups was accomplished that were formed by combining these basic groups. For few of the groups accurate performance was achieved [2].

In case of different classic patterns, classes were mutually exclusive by definition. Due to classes overlapping categorization errors. These problems arise in semantic scene and a framework was presented to handle such problems and apply it to the problem of semantic scene classification where a natural scene may contain multiple object such that the scene can be described by multiple class labels [3].

For acceleration of training of kernel machines a map was proposed for input of data for a randomized low dimensional feature space and then applied existing fast linear methods. Two sets of random features were explored and provided convergence bounded on the ability for approximation of various radial basis kernels [4].

Canonical Correlation Analysis is a technique for finding the correlation between two sets of multi-dimensional variables. It is commonly applied for supervised dimensionality reduction in which multi-dimensional variables is derived from class variables. Experiments on multi-label data sets confirmed established equivalence relationship [5].

A class of algorithms for independent component analysis which on the basis of canonical correlation use contrast functions. Illustrations with simulations involved a wide variety of source distribution showed that algorithm outperform many of the presently known algorithm [6].

Introduction of multi-label classification organizes literature into structured presentation and helps in better performance of comparative experimental results of certain multi-label classification methods. It also helps in quantification of multi-label nature of data set [7].

Least-squares probabilistic classifier is an efficient alternative to kernel logistic regression. LSP utilize least square estimation for long linear model which allows to get a global solution analytically in a class wise manner. It's a practically useful probabilistic classifier [8].

A framework of Probabilistic Functional Testing was proposed which guaranteed a certain level of correctness of system under test as a function of two parameters. One was an estimate of reliability and other was estimate of risk and results were based on theory of formula testing [9].

In the real work, number of class labels could be hundreds or thousands and multi-label classification methods may become computationally inefficient. There are many remedies available by choosing a specific class subset and it can be done by randomized sampling which is highly efficient approach and it's proved by theoretical analysis [10].

In another indagation pivot was on subset of these methods that adopt a lazy learning approach and are based on traditional k-nearest algorithm. Firstly implementation of BRkNN was done and then identified two useful adaptations of kNN algorithm for multi-label classification and then it was compared with other multi-label classification methods [11].

MULAN is a Java library for learning of multi-label data. It presents a number of classification ranking threshold and algorithms on dimensionality reduction. It constitutes evaluation framework for measuring performances [12].

The HOMER algorithm constitutes a hierarchy of multi-label classifiers and every algorithm deals with smaller sets of labels. This leads to training with improved predictive performances and complexities of logarithmic testing. Label distribution was obtained via a balanced clustering algorithm called balanced k means [13].

Stratified sampling is a method which takes into account the existence of disjoint groups within a population and produce sample. One research investigated stratification in the context of multi-label data. Two methods were considered and empirically compared with random sampling on a number of datasets and based on number of evaluation criteria [14].

A large number of research in supervised learning deals with analysis of single label data where training examples are associated. Textual data was frequently annotated with more than a single label. It's perhaps the dominant multi-label [15].

An algorithm was presented called Core vector Machines and its performances were illustrated through other SVM solvers. Training time was independent of sample size and gave similar accuracies to other SVM solvers [16].

A paper proposed an ensemble method for classification of multi-label. The random K label sets built each member of ensemble by by taking into account a single subset. Proposed algorithm took into account label correlation using single label classifiers that are applied on subtasks with manageable number of labels [17].

Independent factor analysis methods were being introduced for recovering independent hidden sources. All the probabilistic calculations were performed analytically [18]. Wolfram language and Mathematica are both already available for Raspberry Pi. Programs can be activated by auto command line through a notebook interface on the Pi. For continues task it's wise to use looping scenario and in Mathematica it can be done by using different statements [19].

Different statistical tests were done on more algorithms on multiple data sets and theoretical examinations were done. Recommendation of a set of simple and robust tests for comparison were given i.e the Wilcoxin signed rank tests and Friedman test for comparison of more classifiers [20].

## III. HARDWARE AND SOFTWARE

Hardware utilized for this purpose called Pi or Raspberry Pi is an economical and cheap desktop PC which can be connected straight to internet and is able to show HD videos. Since Pi works on Linux so there is no dire need to buy OS.

Wide array of programming languages can be selected by using the Raspberry Pi. It is a versatile platform for any kind of fun utility and investigational purpose. Mathematica is a technical computing platform and importantly constitutes two programs which communicate with each other and shape an interactive interpreter: the front end provides GUI and Kernel for calculation purpose. It performs the part of progressive text processor. Formation of notebooks, portable documents are carried out by the front-end. Mathematica which is the software included with the official Raspbian on a Raspberry Pi. This made this investigational research an economical implementation in data mining algorithms.

## IV. IMPLEMENTATION

Following is the workflow of the research:

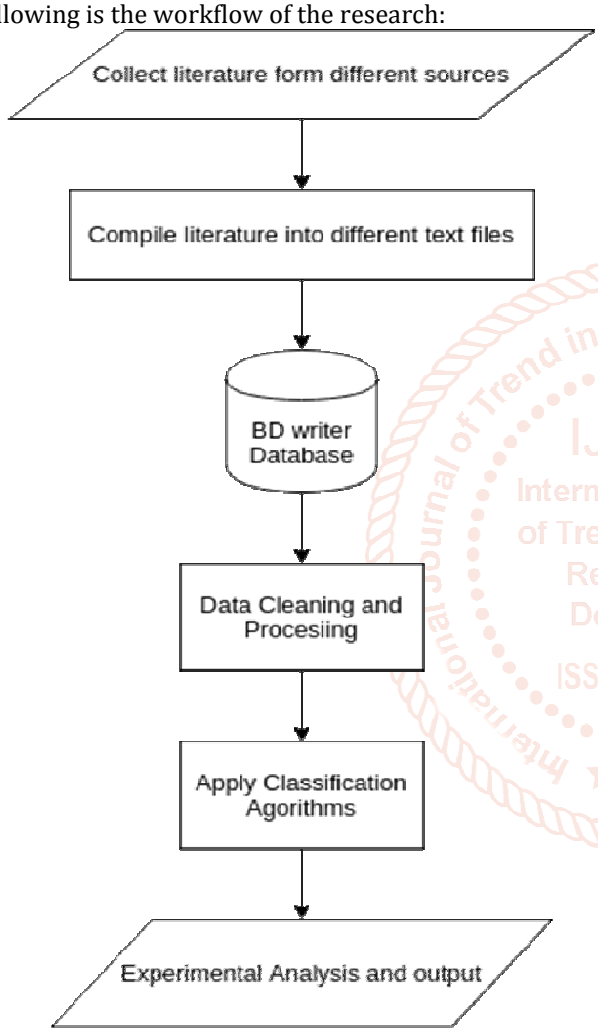


**Fig 1**

Three versatile writings were being gathered for training of the system. Following is the elaboration of these novels.

1. Kazi Nazr-ul-Islam
   Badhonhara, Kuhelika, Mrittukhuda and Sheulimala
2. Rabindranath Tagore.
   Bou thakuranir Hat, Choturongo, Duibon and Gora
3. Sharatchandra
   Borodidi, Devdas, Parineeta and Srikanto

Numbers of lines used of each writer

| S/N | Writer Name | No. Of Lines |
|---|---|---|
| 1 | Rabindranath Tagore | 4467 |
| 2 | Kazi Nazrul Islam | 3212 |
| 3 | Sharatchandra | 6302 |

**Table 1**

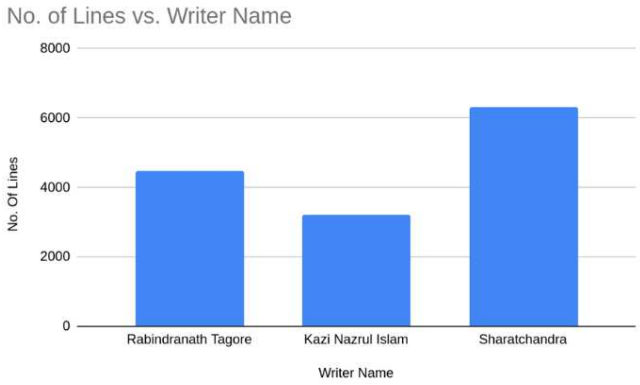


**Fig 2**

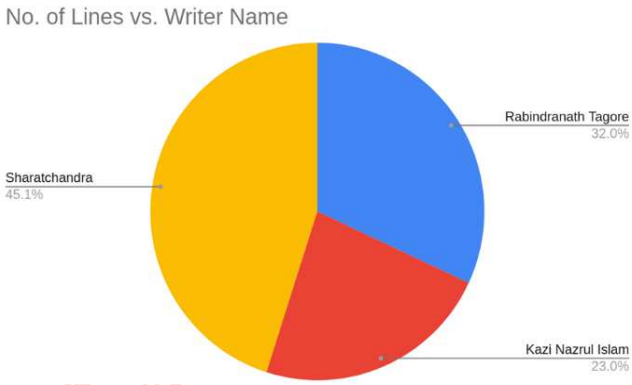


**Fig 3**

Code segment utilized to import the text files in Mathematica are as follows:

```
Badhonhara = Import["/home/pi/Downloads/Writer Classification Dataset/Kazi Nazrul Islam/Badhonhara.txt"];
Kuhelika = Import["/home/pi/Downloads/Writer Classification Dataset/Kazi Nazrul Islam/Kuhelika.txt"];
Mrittukhuda = Import["/home/pi/Downloads/Writer Classification Dataset/Kazi Nazrul Islam/Mrittukhuda.txt"];
Sheulimala = Import["/home/pi/Downloads/Writer Classification Dataset/Kazi Nazrul Islam/Sheulimala.txt"];
BTH = Import["/home/pi/Downloads/Writer Classification Dataset/Rabindronath tagore/Bouthakuranirhat.txt"];
Chaturongo = Import["/home/pi/Downloads/Writer Classification Dataset/Rabindronath tagore/Chaturongo.txt"];
Duibon = Import["/home/pi/Downloads/Writer Classification Dataset/Rabindronath tagore/Duibon.txt"];
Gora = Import["/home/pi/Downloads/Writer Classification Dataset/Rabindronath tagore/Gora.txt"];

Barodidi = Import["/home/pi/Downloads/Writer Classification Dataset/Sharatchandra/Barodidi.txt"];
Devdas = Import["/home/pi/Downloads/Writer Classification Dataset/Sharatchandra/Devdas.txt"];
Parineta = Import["/home/pi/Downloads/Writer Classification Dataset/Sharatchandra/Parineta.txt"];
Srikanto = Import["/home/pi/Downloads/Writer Classification Dataset/Sharatchandra/Srikanto1.txt"];
```

**Fig 4**

Six algorithms were used for classification purpose and their accuracies are given below-

| Classifier Name | Training Time | Training Accuracy | Testing Accuracy |
|---|---|---|---|
| Decision Tree | 59.6 | 40 | 33 |
| Gradient Boosted Trees | 71 | 44 | 12 |
| Logistic Regression | 62 | 11 | 2 |
| Markov | 90 | 78 | 97 |
| Naïve Bayes | 153 | 78 | 98 |
| Support Vector Machine | 65 | 11 | 3 |

**Table 2**

Above values depict that Markov and Naïve Bayes have given the high precision. In English literature it is already confirmed that in Mathematica Markov categorizer works well. Our investigation pivoted that Naïve Bayes gives same results for Bangla literature in the cost of training time. For precision testing Markov and Naïve Bayes are again the best among all.

Detailed Classifier Information is given below:

| Table 3: Decision Tree | |
|---|---|
| Data type | Text |
| Classes | Kazi Nazrul Islam. Rabindranath Tagore. Sharatchandra Chattapoddhiya |
| Accuracy | (40 ± 10)% |
| Method | decision tree |
| Single evaluation time | 1.47 s/example |
| Batch evaluation speed | 0.842 examples/s |
| Loss | 1.13 ± 0.097 |
| Model memory | 2.94 MB |
| Training examples used | 9 examples |
| Training time | 59.6 s |

| Table 4: Gradient Boosted Trees | |
|---|---|
| Data type | Text |
| Classes | Kazi Nazrul Islam. Rabindranath Tagore. Sharatchandra Chattapoddhiya |
| Accuracy | (44 ±44)% |
| Method | Gradient Boosted Trees |
| Single evaluation time | 1.74 s/example |
| Batch evaluation speed | 0.768 examples/s |
| Loss | 1.10 ± 0.18 |
| Model memory | 3.03 MB |
| Training examples used | 9 examples |
| Training time | 1 min 11 s |

| Table 5: Logistic Regression | |
|---|---|
| Data type | Text |
| Classes | Kazi Nazrul Islam. Rabindranath Tagore. Sharatchandra Chattapoddhiya |
| Accuracy | (11 ±16)% |
| Method | Logistic Regression |
| Single evaluation time | 1.43 s/example |
| Batch evaluation speed | 0.811 examples/s |
| Loss | 0.998 ± 0.20 |
| Model memory | 2.98 MB |
| Training examples used | 9 examples |
| Training time | 1 min 2 s |

| Table 6: Markov | |
|---|---|
| Data type | Text |
| Classes | Kazi Nazrul Islam. Rabindranath Tagore. Sharatchandra Chattapoddhiya |
| Accuracy | (78 ±47)% |
| Method | Markov |
| Single evaluation time | 4.02 s/example |
| Batch evaluation speed | 16.1 examples/s |
| Loss | 0.659 ± 0.42 |
| Model memory | 1.06 MB |
| Training examples used | 9 examples |
| Training time | 1 min 30 s |

| Table 7: Naive Bayes | |
|---|---|
| Data type | Text |
| Classes | Kazi Nazrul Islam. Rabindranath Tagore. Sharatchandra Chattapoddhiya |
| Accuracy | (78 ±47)% |
| Method | Naive Bayes |
| Single evaluation time | 40 s/example |
| Batch evaluation speed | 1.57 examples/s |
| Loss | 0.659 ± 0.42 |
| Model memory | 13.09 MB |
| Training examples used | 9 examples |
| Training time | 2 min 33 s |

| Table 8: Support Vector Machine | |
|---|---|
| Data type | Text |
| Classes | Kazi Nazrul Islam. Rabindranath Tagore. Sharatchandra Chattapoddhiya |
| Accuracy | (11 ±16)% |
| Method | Support Vector Machine |
| Single evaluation time | 1.72 s/example |
| Batch evaluation speed | 0.757 examples/s |
| Loss | 1.24 ± 0.22 |
| Model memory | 3 MB |
| Training examples used | 9 examples |
| Training time | 1 min 5 s |

The confusion matrixes here includes the output of categorizing each literature of each writer according to the trained data.

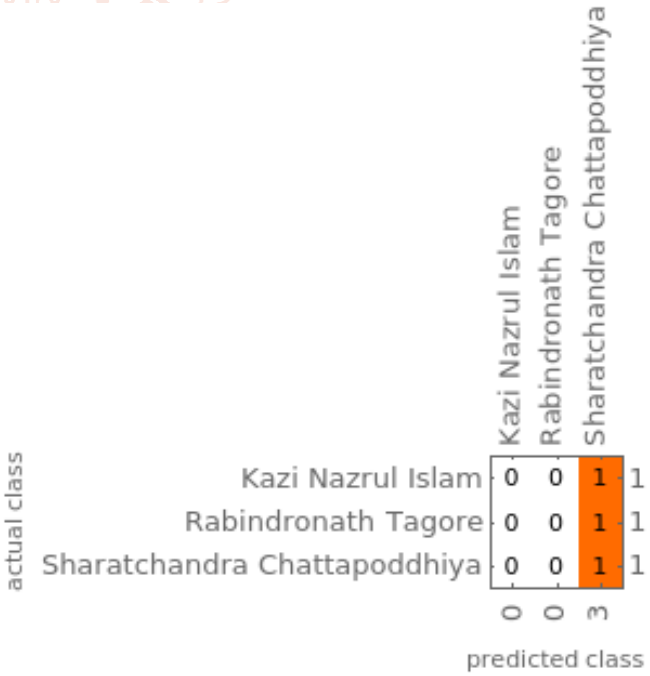


**Fig 5: Decision Tree**

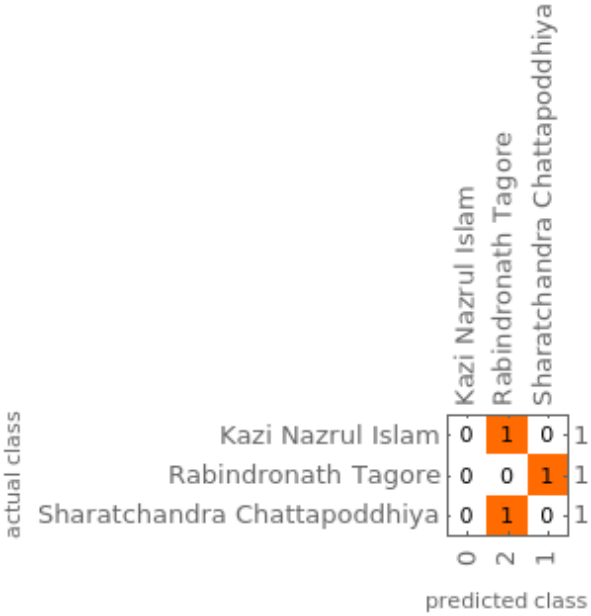


**Fig 6: Gradient Boosted Trees**

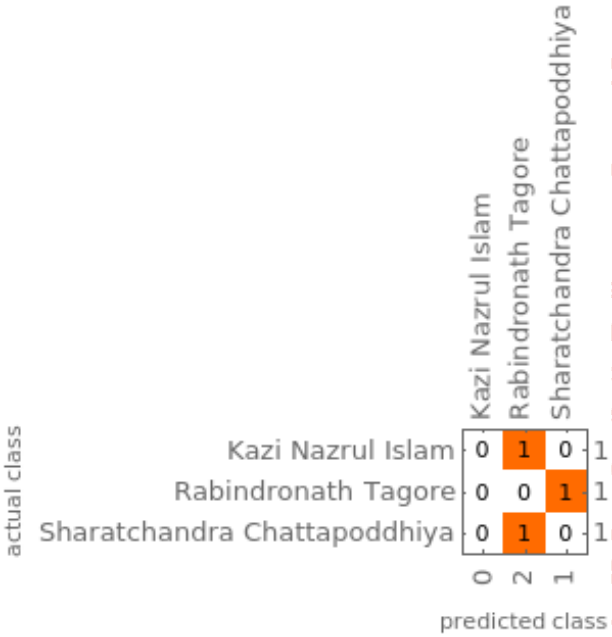


**Fig 7: Logistic Regression**

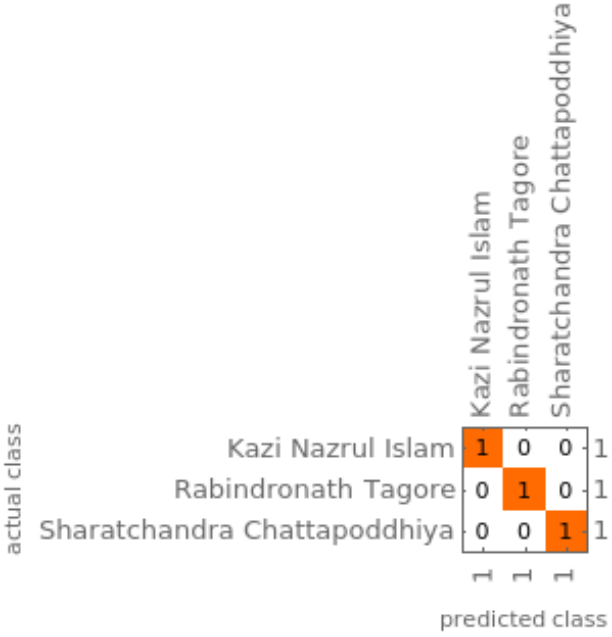


**Fig 8: Markov**

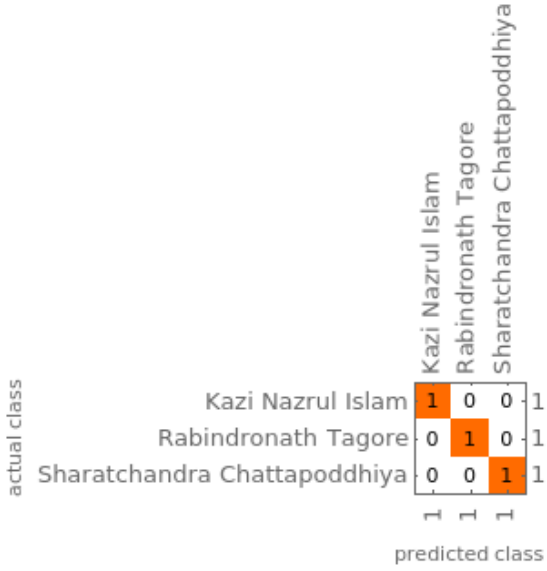


**Fig 9: Naive Bayes**

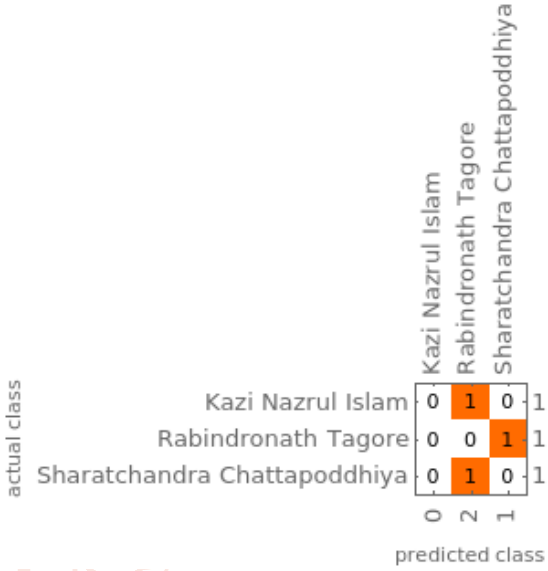


**Fig 10: Support Vector Machine**

**From the confusion matrix we noticed that Naïve Bayes and Markov classifiers identified the 3 literatures accurately.**

```
classicaut = Classify[<|"Kazi Nazrul Islam" → {Badhonhara, Kuhelika, Mrittukhuda}, "Rabindronath Tagore" → {BTH, Chaturongo, Duibon},
   "Sharatchandra Chattapoddhiya" → {Devdas, Parineta, Srikanto}|>]

ClassifierFunction
   Input type: Text
   Classes: Kazi Nazrul Islam, Rabindronath Tagore, Sharatchandra Chattapoddhiya
   Method: Markov
   Number of training examples: 9

classicaut[Barodidi]

Sharatchandra Chattapoddhiya

classicaut[Sheulimala]

Kazi Nazrul Islam

classicaut[Gora]

Rabindronath Tagore
```

**Fig 11**

Following are the output and literature used for testing

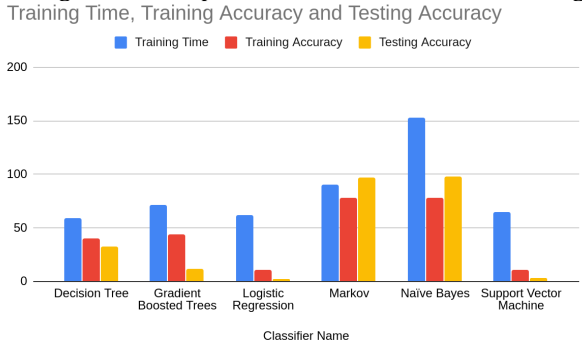


**Fig 12**

From the above dataset it can be said that similar results can be found using Markov and Naïve Bayes. As training time required is much low for Markov model so English language categorization and Bangla language classification can be easily done for writer's anticipation using Markova classifier.

## V. CONCLUSIONS

Markov categorizer is the most suited and economical one because its training time is lower and has greater precision in training and testing. So it has the power to categorize English and Bangla language for writer's prediction. So that languages become more understandable.

## REFERENCES:


[1] A. R. S. K. S. K. S. Surya Ra, "A Low Cost Implementation of Multi-label Classification Algorithm using Mathematica on Raspberry Pi," in Science Direct, India, 2014.

[2] T. M. O. Li, "Detecting emotion in music." in ISMIR, 2003.

[3] A. l. o. o. p. R. M. Browna, "Learning multi-label scene classification," Science Direct, vol. 37, no. 9, pp. 1757-1771, 2004.

[4] A. B. R. Rahimi, "Random features for large-scale kernel machines," Advances in neural information processing systems, 2007.

[5] S. l. Liang Sun, "A least squares formulation for canonical correlation analysis," in Proceedings of the 25th international conference on Machine learning, 2008.

[6] F. R. I. J. Bach, "Kernel Independent Component Analysis," Journal of Machine Learning Research, vol. 3, pp. 1-48, 2002.

[7] G. a. K. I. Tsoumakas, "Multi-Label Classification: An Overview," International Journal of Data Warehousing and Mining, vol. 3, pp. 1-13, 2019.

[8] H. H. M. Y. J. S. a. H. N. Masashi Sugiyama1, "LEAST-SQUARES PROBABILISTIC CLASSIFIER:A COMPUTATIONALLY EFFICIENT ALTERNATIVE TO KERNEL LOGISTIC REGRESSION," in Proceedings of International Workshop on Statistical Machine Learning for Speech Processing, kyoto, Japan, 2012.

[9] G. B. L. Bauzez, "A theory of probabilistic functional testing," in Proceedings of the 19th international conference on Software engineering, 1997.

[10] J. K. Wei Bi, "Efficient Multi-label Classification with Many Labels," in Proceedings of the 30th International Conference on Machine Learning, 2013.

[11] G. T. I. V. Eleftherios Spyromitros, "An empirical study of lazy multilabel classification algorithms," in Springer, Berlin, 2008.

[12] E. S.-X. J. V. I. V. Grigorios Tsoumakas, "Mulan: A java library for multi-label learning," Journal of Machine Learning Research, pp. 2411-2414, 2011.

[13] I. K. I. V. Grigorios Tsoumakas, "Effective and efficient multilabel classification in domains with large number of labels," Proc. ECML/PKDD 2008 Workshop on Mining Multidimensional Data (MMD'08), vol. 21, pp. 53-59, 2008.

[14] G. T. I. V. Konstantinos Sechidis, "On the stratification of multi-label data," in Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Berlin, 2011.

[15] I. K. I. V. Grigorios Tsoumakas, "Mining multi-label data," in Springer, Boston, 2009.

[16] G. L. GAELLE, "Comments on the "Core Vector Machines: Fast SVM Training on Very Large Data Sets"," Journal Of Machine Learning Research, vol. 8, pp. 291-301, 2007.

[17] I. V. Grigorios Tsoumakas, "Random k-Labelsets: An Ensemble Method for Multilabel Classification," in European conference on machine learning, Berlin, 2007.

[18] H. Attias, "Independent factor analysis." vol. 4, no. 11, pp. 803-851, 1999.

[19] A. Kurniawan, "Computational Mathematics with the Wolfram Language and Mathematica: IoT Projects with Wolfram, Mathematica, and Scratch," pp. 97-140, 2019.

[20] J. Demˇsar, "Statistical Comparisons of Classifiers over Multiple Data Sets," Journal of Machine Learning Research, vol. 7, pp. 1-30, 2006.