International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

A Hybrid Neural Network Approach for Author Classification in Written Articles

Gagandeep Singh[1]
Assistant Professor, Department of Computer Science, Sri Guru Teg Bahadur Khalsa College, Sri Anandpur Sahib, Punjab
gsgagan01@gmail.com

Tarandeep Kaur[2]
Research Scholar, Guru Nanak Dev University, Amritsar, Punjab
deep.taran4@gmail.com

Abstract:
With the exponential growth of digital content, accurately determining the authorship of textual material has become increasingly critical. Authorship attribution, the task of determining the writer of a given text, has emerged as an important area of research within the domains of Natural Language Processing (NLP) and Machine Learning (ML). This paper presents an approach for classifying the authorship of written articles using NLP techniques and supervised ML algorithms. Initially, a dataset comprising text samples from three distinct authors was prepared. The text data was pre-processed to remove noise, and essential features were extracted using the TF-IDF technique. These features were then utilized to train and evaluate three supervised classifiers: Support Vector Machine (SVM), Naïve Bayes, and Logistic Regression. The performance of each classifier was assessed using accuracy, precision, recall, and F1-score metrics. Among the models tested, the SVM classifier achieved the highest accuracy of 94%. The results demonstrate that the proposed approach is effective for authorship classification and holds promise for applications in digital forensics, content verification, and intellectual property protection.

Keywords: Enhanced MLP, Information gain, Decision Trees, Neural Networks

1. INTRODUCTION

In some real-world situations, the capacity to naturally characterize archives into a settled arrangement of classifications is exceptionally attractive. Like verbal discoursed and compositions, reports additionally have the creator's unmistakable style and distinguishing proof of the creator of an obscure archive is a critical and testing undertaking in PC security. This gets even more difficult when we need to do distinguishing proof of a style of a gathering of assorted people acting in comparable conditions, like writers writing in certain abstract period or individuals writing in a specific social part. The sum of what these have been propelled by the way that we leave attributes of creation in our compositions because of the way that we have novel methods for composing. *Creator Identification*: Authorship distinguishing proof is the undertaking of foreseeing the in all probability creator of a content given a predefined set of competitor creators and various content examples per creator of undisputed initiation. From a machine

learning **perspective,** this assignment can be viewed as a solitary mark multi-class content arrangement issue where the competitor creators assume the part of the classes.

The primary objective is to build up a savvy vigorous classifier that is equipped for dissecting the writer of a bit of composing given a predefined set of competitor writers and various examples for each writer. To grow such a classifier, it is essential to recognize an arrangement of qualities or elements that is sufficiently solid to recognize the diverse creators. Moreover, a clever learning calculation must be worked in a programmed approach to have the capacity to distinguish the initiation of a bit of writing. To build up a wise programmed hearty classifier, it is vital to have the correct list of capabilities and a solid learning calculation. There has been incredible enthusiasm for the use of machine learning techniques with the end goal of creator distinguishing proof, and these strategies have turned out to be helpful and fruitful for this reason.

2. DIFFERENT METHODOLOGIES FOR AUTHOR CLASSIFICATION

There are multiple techniques used for author attribution/identification. These can be broadly categorized in 4 categories.

***Statistical Based methods*:** Statistical based methods are based on the analysis and identification of joint dissemination of some style factors making it conceivable to choose whether two writings demonstrate a huge relationship of style or not. Statistical based methods are the most common approaches used for author identification.

Rule Based Methods: Rule based methods are based on the rules formulated for solving problem. In this, goal is to identify the probability of one event happening under the given contexts and events are defined by rules. It simply works on if and only if the certain clause exhibits attributes already identified or not. If the clause exhibits, then it can be formulated that A =< B, determining relation between two entities we are comparing.

Machine Learning Based methods: Statistical and stylometric based methods are the most used methods but they have limitations when it's come to increase in complexity in author identification. Machine based learning technique with respect to building classifiers that induce the lexical and syntactic characteristics that portray the style of a creator. The techniques which are a piece of this approach rely on two phases: the principal comprises in speaking to the source messages as vectors of marked and multivariate words. The second comprises in utilizing learning strategies to recognize the limits of each class, in the meantime limiting a grouping misfortune work. Examples are SVM, Neural Networks.

Hybrid Approach: Hybrid methods combine the best features of Statistical, stylometric, machine learning, rule-based learning and try to solve the problem. Learning techniques are being used to infer attributes that discriminate the styles of authors.

3. CLASSIFICATION ALGORITHMS

Machine Learning Techniques enables the computer to learn by itself without being programmed. Machine learning algorithms are more efficient in contrast to those of non-machine learning. Machine learning work in a similar way like data mining, both acquire knowledge from data and find relevance in the data. Machine learning algorithms can be categorized into supervised and unsupervised algorithms.

Some of the supervised machine learning algorithms for classification of data is listed below:

Multi-Layer Perceptron Neural Network: Artificial neural networks are the part of artificial intelligence and MLP is a type of neural network. Multi-Layer Perceptron (MLP) is a feed-forward network that maps the group of inputs to their corresponding outputs. Fig 1demonstrates a feed-forward multi-layer perceptron neural network. MLP is made up of simple neurons termed as a perceptron. Neural network generates information by enabling input perceptron's consisting of the values labelled on them. The activation function of neurons is calculated by the formula mentioned below in the output layer:

$$a_i = \sigma(\sum_j W_{ij}\, O_j) \qquad \text{Eq. No.1}$$

Where a_i represent level of activation for ith neurons; j is the set of neurons of the previous layer; W_{ij} is the weight of the connection between neurons i and j; O_j represents the output of jth neuron and $\sigma(x)$ is the transfer function.

$$\sigma(x) = \frac{1}{1+e^x} \qquad \text{Eq. No. 2}$$

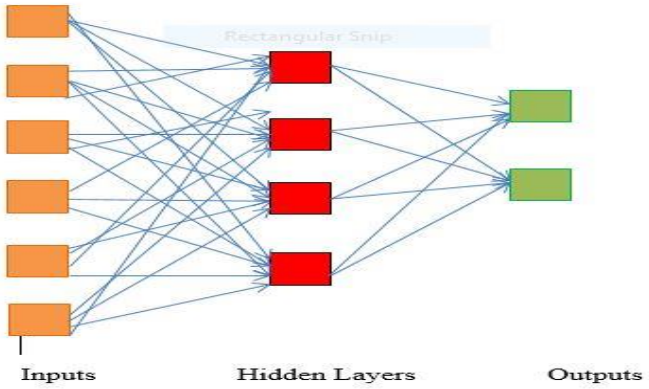

Fig.1 Multi-Layer Neural Network Architecture.

A back-propagation strategy based on delta rule is used to train multilayer perceptron network. MLP consist of various neurons divided into various layers, as follows:

- **Input Layer:** This layer generates the input for the network. The number of neurons depends upon the number of inputs given to the network.
- **Hidden Layer:** The layer that maps the input to the corresponding output is named as a hidden layer. Hidden layers vary in number.
- **Output Layer:** The layer from where the resultant can be seen. The number of neurons in output layer depends upon the learning of the kind of problem.

The type of relationship between the input and output vectors in MLP is a non-linear relationship. This is done by interconnecting the neurons in the antecedent and succeeding layers. Outputs are achieved by multiplying them with weight coefficients. In the training phase, the neural network is given the information regarding training only. Later, the weights of the network are tuned between [-0.5, 0.5] to minimize the error rate between the expected and observed outputs and to enhance the frequency of

training to a predicted level. A sequence untrained inputs are applied to the input to formalize the training. These input set are different from the inputs that are used for the training of the network. Training of the neural network is highly complex due to many variables. MLP holds lots of advantages over other algorithms even if a correct relationship is not induced between the input and output or if the essential and exact information is not achieved [4]. Non-Linear Activation Function of MLP makes MLP different from other networks. Algorithmic steps for MLP- Neural Network can be modeled as:

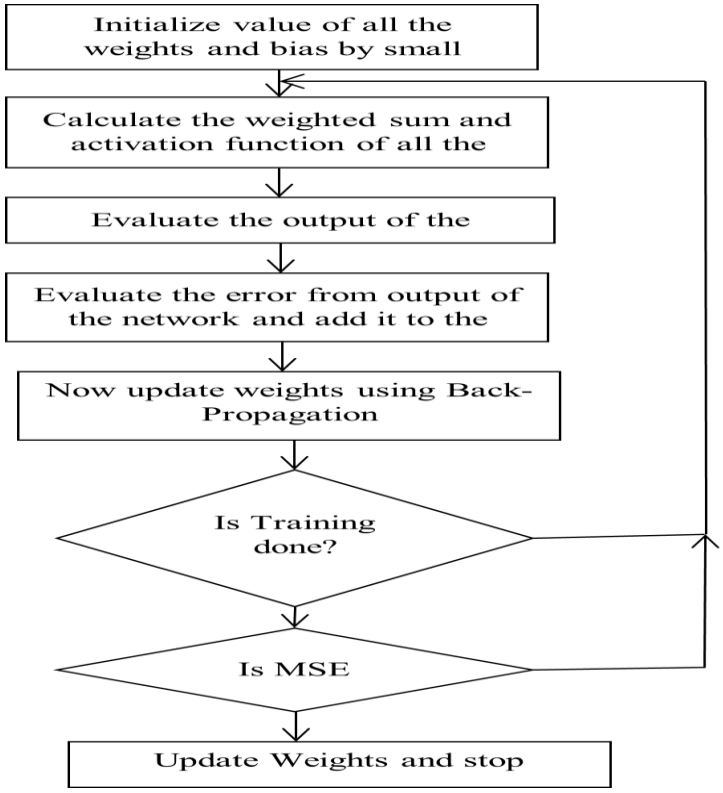

Fig 2 Multi-Layer Neural Network Process diagram

Let a dataset D, consist of training samples and their target values, L be the rate of learning by the network to generate a trained network:

1. Initialize the weights and the biases of the layers using small random values.
2. Compute the weighted sum of the inputs, where

$O_j = I_j$ Eq. No. 3

The output of the inputs is the true input values.

3. Compute the activation functions of hidden layers, where

$I_j = \sum_j W_{ij}\, O_j + \theta_j$ Eq. No. 4

Compute the net input of j with respect to i (previous layer).

4. Compute the output of the layers, where

$O_j = \frac{1}{1+e_j^{-1}}$ Eq. No. 5

5. Compute the error rate by Back-Propagation, Error for output layer:

$Error_j = O_j(1 - O_j)(T_j - O_j)$ Eq. No. 6

6. Error calculation of next hidden layer, h:

$$Error_j = O_j(1 - O_j) \sum_h Error_h w_{jh}$$ Eq. No. 7

7. Weight update:

$$w_{ij}=w_{ij} + \Delta w_{ij}$$ Eq. No. 8

8. Bias update:

$$\theta_j = \theta_j + \Delta\theta_j$$ Eq. No. 9

Where Δw_{ij} and $\Delta\theta_j$ are the changes in weight and bias.

Support Vector Machine: Support vector machine (SVM) is a supervised approach for machine learning. The main idea used in SVM is constructing a hyper-plane that is optimal for the classification of patterns that can be linearly separated. This algorithm work by plotting each information point in a n-dimensional workspace, where n represents the number of features which are equal to the coordinates in the workspace. The optimal hyper plane differentiates the classes at this point

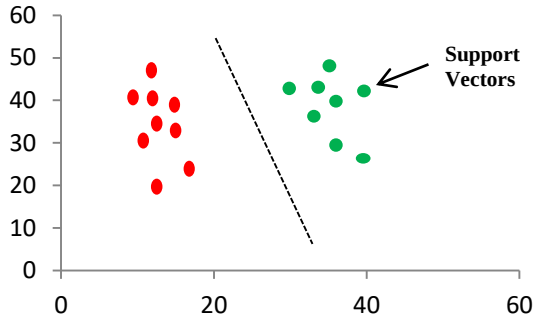

Fig 3: SVM representing the difference between two classes using hyper-plane.

Algorithmic steps of SVM for the classification process are as follows:

1. Train the initial SVM using all the training data to have support vectors decision functions.
2. Eliminate those support vectors generated from the training of initial SVM whose projections have greatest curvatures on the hyper surface by finding the projection of the support vectors along the gradient of decision function used, calculate the notion of curvature for every support vector on the hyper-plane, lastly sort the support vectors in the decreasing order and deduct the top N-percentage of the vectors of support.
3. Retrain the SVM by left over vectors for best decision.
4. Use the group of information point to finally train the SVM, generating support vectors.

SVM classifiers are grouped into linear and non-linear classifiers, as follows:

1. **Linear Classifiers:** Separating the data points in linear order by using a hyper-plane is classified as linear classifiers. There are different hyper-planes but the best way to separate the data using hyper-plane is by maximum margin difference viz. the distance of hyper-plane and the closest information point of any class.
2. **Non-Linear Classifiers:** Sometimes the data is not separated properly or linearly in the high dimensional plane for such separation non-linear classifiers are used that correctly classify the information points and label them to their exact class by using kernel tricks. Some mostly used kernel tricks are as follows:

- **Homogenous kernels:** Polynomial kernels that are used for analysing the similarity of vectors are represented by the expression below:

$k(\vec{ai}, \vec{a_j}) = (\vec{ai}, \vec{a_j})^{d}$ Eq. No. 10

Where k is the kernel function and $(\vec{ai}, \vec{a_j})$ are the vectors of the work space with d as the degree of the polynomial.

- **Non-Homogenous kernels:** In Non-homogenous kernels, a free parameter is added that leverage the group of features combined together.

$k(a,b) = (a^T b + c)^d$ Eq. No. 11

K-Means clustering algorithm is the fastest algorithm that work efficiently on a large dataset. The problem of randomization of the MLP classification algorithm degrades its ability to remove vague information from the dataset. To overcome the disadvantage of randomization of the MLP classification algorithm, initial clusters are provided by the K-Means clustering algorithm.

4. INFORMATION GAIN (IG) FEATURE SELECTION

Feature subset selection is the way toward distinguishing and evacuating as much unessential and excess data as could be expected under the circumstances. This lessens the dimensionality of the information and may enable learning algorithm to work speedier and even more adequately. Now and again, accuracy on future classification can be enhanced; in others, the outcome is a more reduced, effectively deciphered portrayal of the objective idea.

Steps of Feature Selection: An element of a subset is great on the off chance that it is highly correlated with the class however very little associated with different elements of the class. [15]

1. Subset generation: We have used four classifiers to rank all the features of the data set. Then we have used top 3, 4, and 5 features for classification.
2. Subset evaluation: Each classifier is applied to generated subset.
3. Stopping criterion: Testing process continues until 5 features of the subset are selected.
4. Result validation: We have used 10-fold cross validation method for testing each classifier's accuracy.

Information gain is depending on Claude Shannon's work on data hypothesis. Info Gain of an attribute A is utilized to choose the best part basis quality. The most noteworthy Info Gain is chosen to assemble the decision tree.

Info Gain(A) = Info(D)−Info(A) Eq. No.12

A) Classification by MLP-NN

Various classification algorithms are present that are used to classify the news articles such as Decision tree, Naive Bayes and SVM. In our research, we use MLP Neural Network, because of its advantages such as generalization and fault-tolerance. The major goal of this proposed work is to upgrade the existing machine learning method in distinguishing article by the name of the creator.

B) Proposed N-K-MLP Algorithm

Let a news article dataset containing *n* news articles to be labelled.

Step 1. Perform dataset pre-processing by lexical analysis, removing stop words and stemming.

Step 2. Optimize the features using information gain feature selection algorithm.

Step 3. Perform K-Means Clustering for selection of initial clusters and for grouping the news in three defined clusters.

Step 4. Provide the K-Means results to the MLP model as initial clusters for avoiding randomization for the detection of vague information and classifies the news articles based on the author.

Step 5. Evaluate the performance.

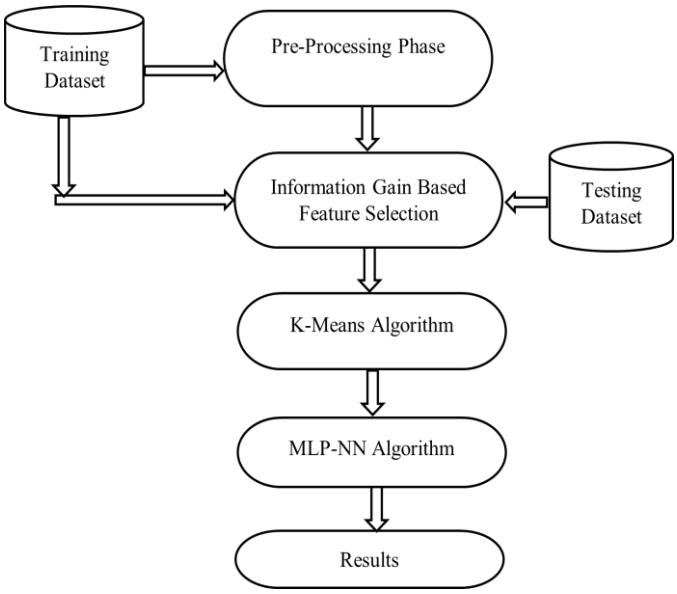

Fig 4 Proposed technique

5. RESULTS AND DISCUSSIONS

This is the first window we come across,here we select the files that our dataset have.that are preloaded when we create database at the backend.We have created newscutting. arff file .Here's it contains our newscutting dataset that we browse through,we upload them as such.here we upload from the main server thing.

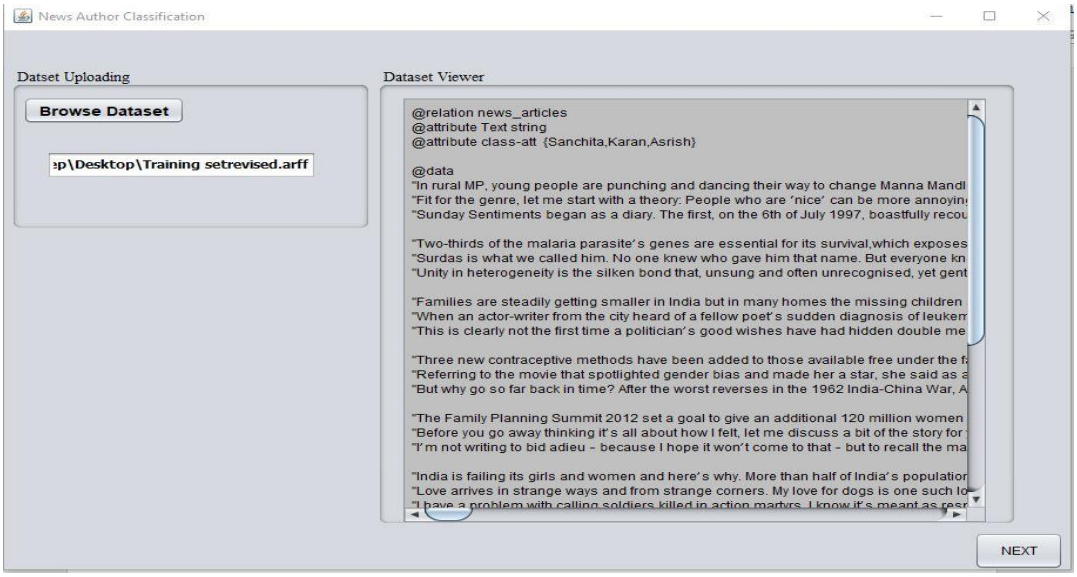

Fig 6 Choosing the dataset and showing the content.

Dataset Filtering

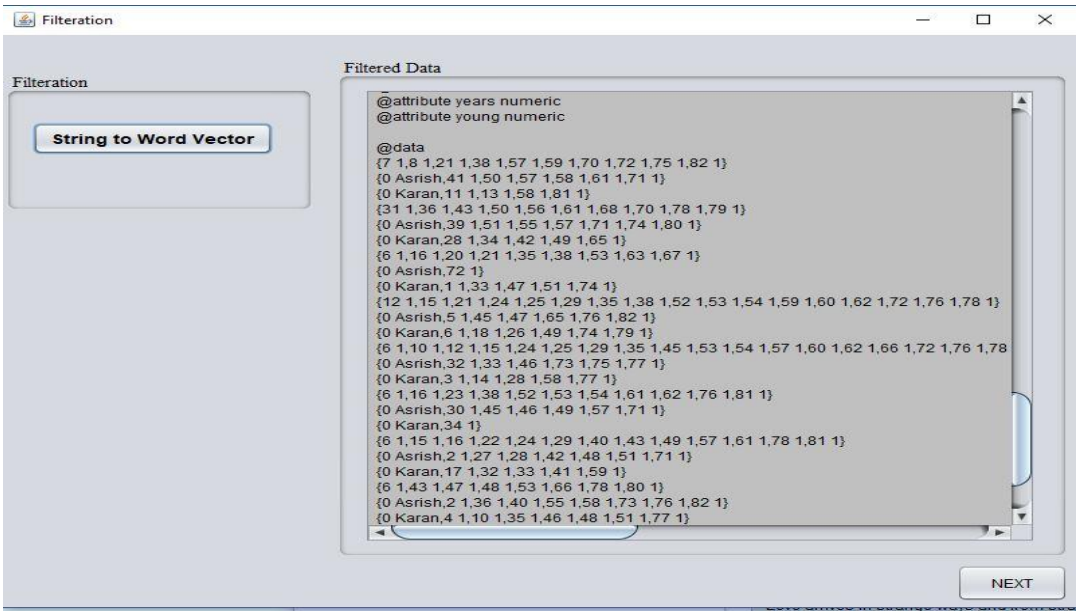

Fig.7 Showing the results of String to word vector filter

A) ***Classification Results:*** The classifier model built is evaluated and tested using k-fold Cross Validation approach (k- fold CV). In this, the training set is divided into 10 smaller sets and results are analyzed.

i) MLP Classification algorithm

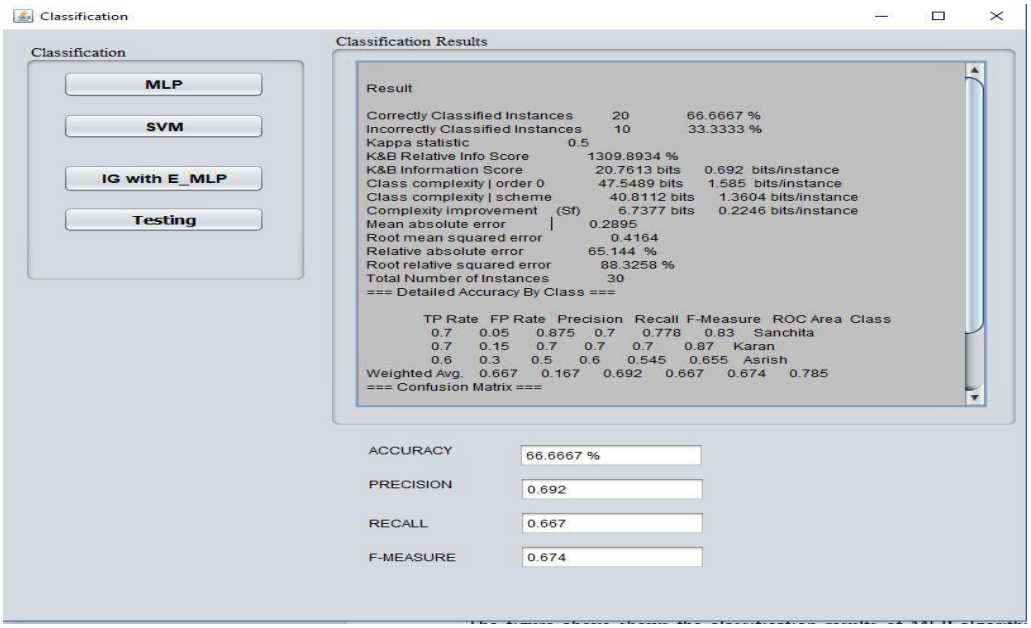

Fig 8: Showing the results of MLP Classification algorithm

The figure above shows the classification results of MLP algorithm. The results show the accuracy of 66.66% i.e., 20 instances are correctly classified out of 30 instances. The kappa statistics is 0.5 Class details parameters are also shown like precision which is 69.2%, recall 66.7 %, F Measure 67.4%, TP Rate 66.7% and FP rate 16.7%.

ii) SVM Classification

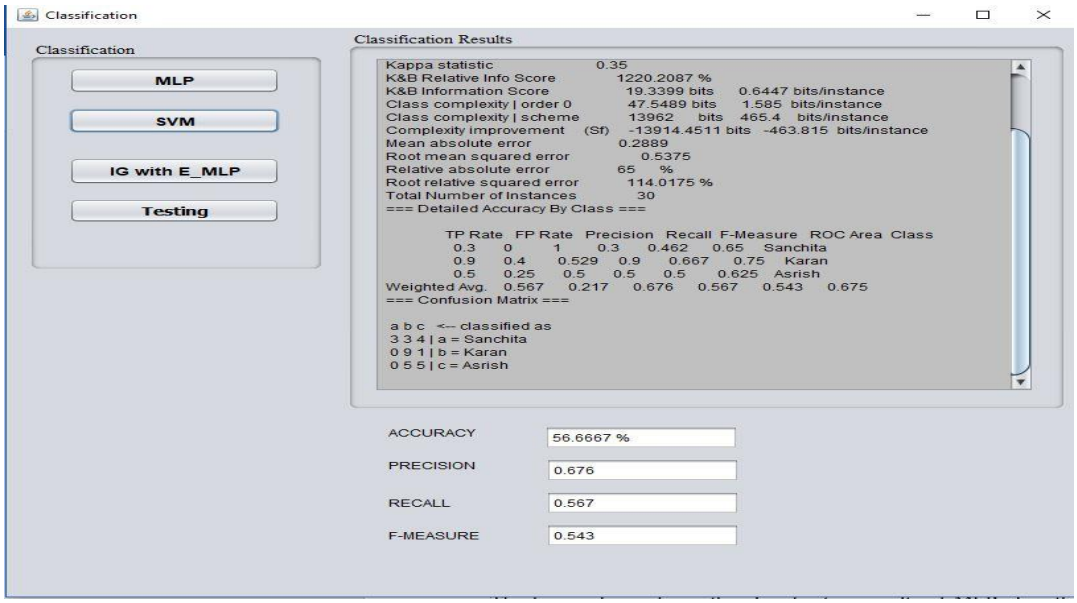

Fig 9: Showing the results of SVM Classification algorithm.

The figure above shows the classification results of SVM algorithm. The results show the accuracy of 56.66% i.e. 17 instances are correctly classified out of 30 instances. Class details parameters are also shown like precision which is 67.6%, recall 56.7 %, F Measure 54.3%, TP Rate 56.7% and FP rate 21.7%.

Proposed Classification Algorithm

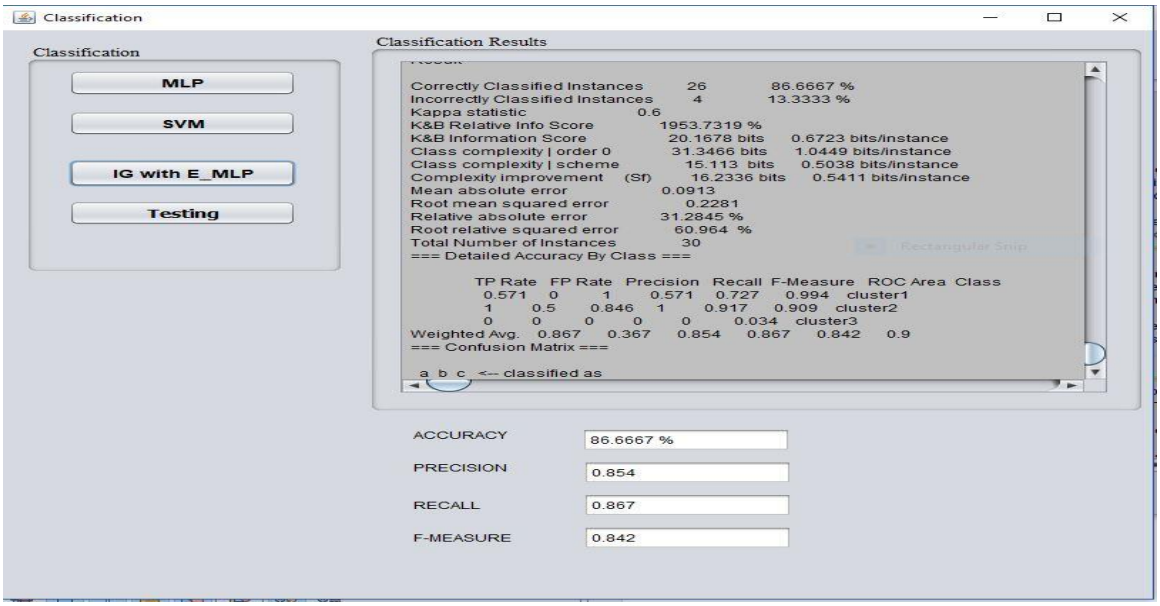

Fig 10: Showing the results of Proposed Classification algorithm.

The figure above shows the classification results of Proposed algorithm. The results show the accuracy of 86.66% i.e., 26 instances are correctly classified out of 30 instances. Class details parameters are also shown like precision which is 85.4%, recall 86.7 %, F Measure 84.2%, TP Rate 86.7% and FP rate 36.7%.

Table: SVM, MLP and E-MLP comparison on dataset.

Parameters /Algorithms	**SVM**	**MLP**	**E-MLP**
Correctly Classified Instances (Accuracy)	56.66%	66.66%	86.667%
Precision	0.676	0.692	0.854
Recall	0.567	0.667	0.867
F-Measure	0.543	0.674	0.842

In table 1 result comparisons are performed for several parameters for E-MLP, MLP and SVM classification algorithms. After the pre-processing of the dataset by performing, lexical analysis and tokenization, stop words removal, and stemming of the dataset, E-MLP classification algorithm performed better over the SVM classification algorithm and MLP classification algorithm. The major problem with the SVM classification algorithm is that it fails to accommodate all the data objects lying far from the density function due to the failure of correctly placing the hyperplane.

6. APPLICATIONS

Author classification has different applications. Some of them are recorded beneath Authorship distinguishing proof. The primary point is to recognize the writer of a bit of composing given various existing compositions by a similar writer. Writer Characterization Author portrayal expects to create the writer profile in view of his/her written work style. The profile gives information about the attributes of the creator, for example, gender and instructive and social foundation. Closeness Detection Similarity recognition which relates to the end goal of distinguishing if there is any sort of written falsification in each bit of composing, including an entire or halfway replication of a bit of work without the consent of the first writer. It is completed by looking at various bits of works and deciding if they were delivered by a similar creator or not. Content order Main point is to sort an arrangement of content archives in view of its substance. Uses of content classification incorporate record separating and report recovery. Other application incorporates criminal law (identifying writers of ransom notes and harassing letters), civil law (copyright and estate disputes), and computer security (mining email content).

7. ISSUES

Problems faced in the author classification includes: 1. Selecting feature set – this includes how the feature will be selected, either by word count (frequency) or by-word length. Further how much word length will be capable of classifying documents, what will be the value of frequency of word occurrence will be sufficient to classify. 2. Lack of text samples of undisputed authorship- In this there is lack of enough test data or some authors and ample amount of data for other candidate authors. 3. Optimization feature set to improve accuracy. 4. Short length texts of certain authors (less than 1000 words) 5. Ambiguity problem: If multiple authors have written articles on same context, it will be little difficult to identify author.

8. CONCLUSION

In the process of Author Identification, the important aspects of a document remain unknown, and it is difficult to analyse the document completely. This has been one of the major issues faced in identifying the authors. The SVM methods provides a solution to overcome this problem. But accuracy became an issue in SVM. To overcome this problem MLP with Info gain is introduced. All these methods are based on Machine Learning. Author Identification can be useful for detection of plagiarism in the context of e-learning.

REFERENCES

1. Bora, Shital P. "Data mining and ware housing." Electronics Computer Technology (ICECT), 2011 3rd International Conference on, E-ISBN :978-1-4244-8679-3, Vol. 1. IEEE, 2011.
2. Singh, Vandana, and Sanjay Kumar Dubey. "Opinion mining and analysis: A literature review." Confluence The Next Generation Information Technology Summit (Confluence), 2014 5th International Conference,978-1-4799-4236-7/14/$31.00_c 2014 IEEE.
3. Srivastava, Shweta. "Weka: A Tool for Data preprocessing, Classification, Ensemble, Clustering and Association Rule Mining." International Journal of Computer Applications 88.10 (2014): 26-29.
4. C. Zhang and Z. Zhang, “A survey of recent advances in face detection,” Technical report -Microsoft Research, 2010.

5. R. Xu. and D. Wunsch, “Survey of Clustering Algorithms,” IEEE Transaction on Neural Networks, vol. 16, no. 3, pp. 645–678, 2005.
6. M. S. Chen, J. Han, and P. S. Yu, “Data mining: An overview from a database perspective,” IEEE Transaction on Knowledge and Data Engineering, vol. 8, no. 6, pp. 866–883, 1996.
7. Kamal Nigam, John Lafferty, Andrew McCallum “Using Maximum Entropy for Text Classification” in International Joint Conference on Artificial Intelligence (IJCAI), Sweden, 1999, pp. 369-376
8. Robert E. Schapire, Yoram Singer “A Boosting based system for text categorization” in Machine Learning - Special issue on information retrieval, Massachusetts, 2000, pp. 135-168 3. Moshe Koppel, Shlomo Argamon, Anat Rachel Shimoni “Automatically Categorizing Written Texts by Author Gender” in Oxford Journals: Literary and Linguistic Community, 2002, pp 401-412.
9. Thorsten Joachims “Text Categorization with Support Vector Machines” Machine Learning: European Conference on Machine Learning Chemnitz, Germany, 2005, pp 137-142.
10. Daniel Pavelec, Edson Justino, Leonardo V. Batista, Luiz S. Oliveira “Author Identification using Writer Dependent and Writer Independent Strategies” in Symposium on Applied Computing, Ceara, 2008, pp. 414-418
11. Vineet Chaoji, Apirak Hoonlor and Boleslaw K. Szymanski “Recursive Data Mining for Author and Role Identification” in 3rd Annual Information Workshop ASIA 08, Albany, 2008, pp. 53-62
12. Dennis Ramdass & Shreyes Seshasai “Document Classification for Newspaper Articles” in 30th annual international ACM SIGIR conference on Research and development in information retrieval, Amsterdam, 2009, pp. 733-734
13. Na Cheng, R. Chandramouli, K.P. Subbalakshmi “Author gender identification from text” in Digital Investigation: The International Journal of Digital Forensics & Incident Response, Amsterdam, 2011, pp 78-88
14. Abdur Rehman, Haroon A. Babri, Mehreen Saeed “Feature Extraction Algorithms for Classification of Text Documents” in International Conference on Communications and Information Technology, Aqabajordan, 2012, pp. 1148-1154