Text and Image Plagiarism Detection

Dr. Om Prakash Samantray[1], K. Divya[2], M. Amrutha[3], M. Chandan[4], M. Rajashekar[5]

Associate Professor, Department of Computer Science and Engineering[1]
Students, Department of Computer Science and Engineering[2,3,4,5]
Raghu Institute of Technology, Visakhapatnam, AP, India

Abstract: *Today, much more than in the past are discussed of plagiarism in the research. Conditions of the Web and Possibility of complex and smart searches in a short time, is rated to this, and as a result has arrived significant damages to the research. Tools designed to deal with plagiarism act on the text and ignore images. On the other, an inseparable part of information transfer are images that transfer the large volume of information in an article or scientific research. Because of the images include a very wide range and especially found large amounts of images in the computer's texts, and as respects, flowcharts are carrying a lot of information, could be one of the options of plagiarism. The purpose of this project is examining the plagiarism rate of a paper in terms of images plagiarism using Histogram Model.*

Keywords: KNN, Machine Learning, Plagiarism, Text Plagiarism, Image Plagiarism

I. INTRODUCTION

The issue of plagiarism is often discussed in the educational community across the world. It relates to the act of taking another person's work and passing it as your own. Basically it converts the existing information in a modified format. Plagiarism is said by S. Hannabuss as "is the act of using somebody else's creation or idea without permission and presenting it as one's own . Today with the huge popularity of internet, so many documents are freely available. Now internet is a source for different types of files and data. People can easily get their required information or data from internet and copy instead of writing their own text document with their own mind. In recent times, the detection of plagiarism becomes more important as it is very easy for a plagiarist to find an appropriate text data that can be copied. On the other side it becomes increasingly difficult to correctly identify plagiarized data due to the large amount of possible sources of data over internet. Plagiarism cases are an everyday topic, for example, in academics, journalism, scientific research,politics and even in many other sectors. This approach to plagiarism detection is especially useful when no data collection is available or not all the possible copy sources are present, thus documentto-document comparison algorithms cannot be used. Plagiarism is of various types like literal,image, integral, intrinsic, extrinsic, exact copy, text manipulation etc. Similarly various plagiarism detection methodologies and methods are present to detect plagiarism. Presently systems which are based on the text and image manipulation techniques are not accurate enough for practical applications. Therefore, we have proposed a new easy method which is based on the texta image identification technique through file transfer method which uses a machine learning approach in order to detect plagiarism between text sets and images. It compares two files and identify how many words are similar between two files then we calculate a percentage value according to our threshold value required to detect plagiarism, image hologram percentage helps to get image plagiarism, through which we can get the plagiarised text and image series .

II. RELATED WORK

In some of text-based, citation-based and shape-based plagiarism detection methods have been compared with each other. According to comparison in a copy-paste plagiarism, Text-based plagiarism detection methods have been almost 70 percent whereas citation-based methods inefficient in this regard. About the translated texts plagiarism, text-based methods have been successfully Less than 5 percent, and this value in citation-based method is about 80 percent. Existing system, have not been performed the comparison of images.

A Selamat, IMI Subroto and Choon-Ching Ng(2009)[1] Arabic Script Web Page Language detection Using Hybrid KNN.One of the crucial tasks in the text-based language identification that utilizes the same script is how to produce reliable features and how to deal with the huge number of languages in the world.

Ahmad Gull Liaqat and Aijaz Ahmad(2011)[2]Advanced Supervised Learning in Multi-layer Perceptrons - From

DOI: 10.48175/IJARSCT-9243

Backpropagation to Adaptive Learning Algorithms. Since the presentation of the backpropagation algorithm a vast variety of improvements of the technique for training the weights in a feed-forward neural network have been proposed. UpulBandara and Gamini Wijayrathna(2012)[3]Detection and identification of Source Code Plagiarism Using Machine Learning Approach. Source code plagiarism is currently a severe problem in academia. In academia's programming assignments are used to evaluate students in programming courses.

Imam Much Ibnu Subroto and Ali Selamat, (2014)[4] Plagiarism Detection through Internet using Hybrid Artificial Neural Network and Support Vectors Machine most of the plagiarism detections are using similarity measurement techniques. Basically, a pair of similar sentences describes the same idea.

III. EXISTING SYSTEM

In some of text-based, citation-based and shape-based plagiarism detection methods have been compared with each other. According to comparison in a copy-paste plagiarism, Text-based plagiarism detection methods have been almost 70 percent whereas citation-based methods inefficient in this regard. About the translated texts plagiarism, text-based methods have been successfully Less than 5 percent, and this value in citation-based method is about 80 percent. Existing system, have not been performed the comparison of images.

IV. PROPOSED SYSTEM

The proposed system has two phases: training and testing. They are seen as in train phase used of Histogram in learning stage and in the test phase in the recognition stage taken help from the modelling done by this network. Data analysis method and input image similarity detection rate with images in the database is based on the query image correlation rate with each test images and select images with the highest correlation. Correlation levels obtained at this stage report as the tested image plagiarism and the final interpretation is the responsibility of the expert.

V. SYSTEM IMPLEMENTATION

- **Preprocessing:** The aim of preprocessing is an improvement of the image data that suppresses unwanted distortions or enhances some image features important for further processing.
- **Train the Model:** The functions of keras will start the training of the module..
- **Evaluate the Model:** We have images in our dataset which will be used to evaluate how good our model works.

5.1 Algorithm

K-Nearest Neighbour is one of the Machine Learning algorithms based on Supervised Learning technique.

K-NN algorithm assumes the similarity between the new case and available cases and put the new case into the category that is more similar to the available categories.

K-NN algorithm stores and compares all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified and identified into a well suite category by using K- NN algorithm.

K-NN algorithm or method can be used for Regression as well as for Classification but mostly it is used for the Classification problems.

5.2 System Architecture

A system architecture or systems architecture is the conceptual model that architecture is the conceptual model that defines the structure, behavior and more views of a system. An architecture description is a formal description is a formal description and representation of a system, organized in a way that supports reasoning about the structures and behaviors of the system. System architecture can comprise system components, the externally visible properties of those components, the relationships between them.

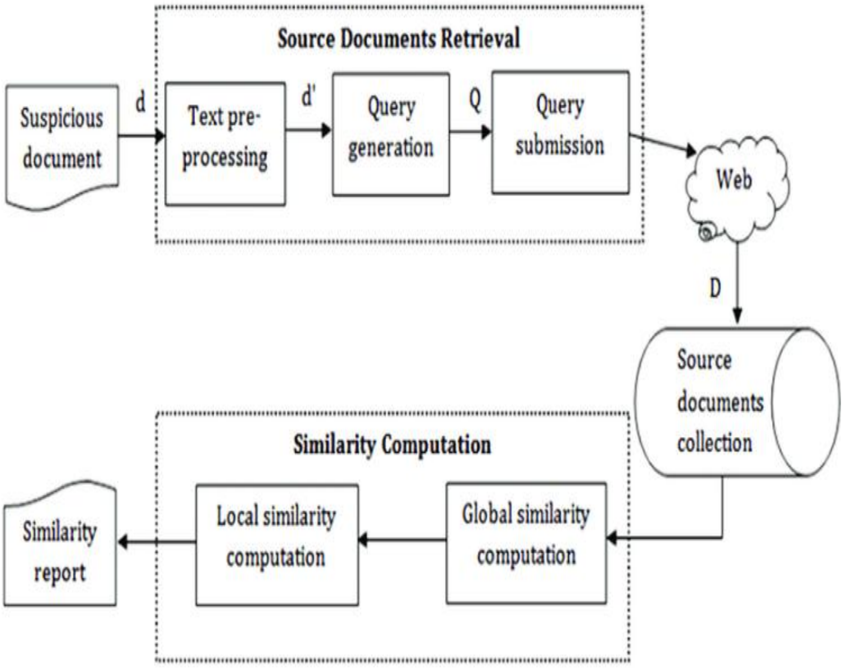
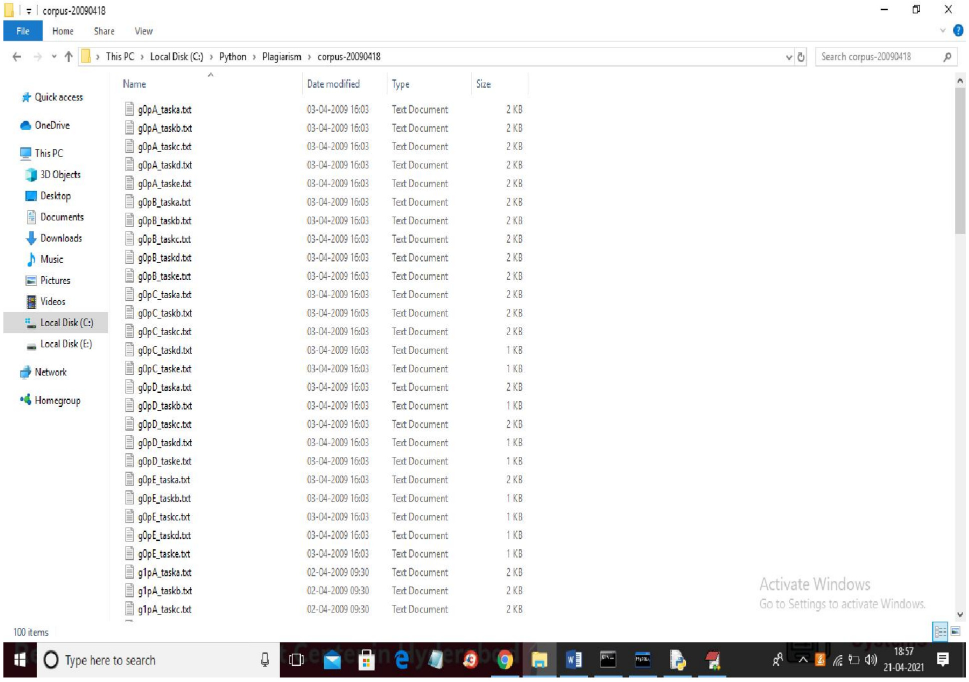


Figure 1: System Architecture

VI. RESULT



Name	Date modified	Type	Size
g0pA_taska.txt	03-04-2009 16:03	Text Document	2 KB
g0pA_taskb.txt	03-04-2009 16:03	Text Document	2 KB
g0pA_taskc.txt	03-04-2009 16:03	Text Document	2 KB
g0pA_taskd.txt	03-04-2009 16:03	Text Document	2 KB
g0pA_taske.txt	03-04-2009 16:03	Text Document	2 KB
g0pB_taska.txt	03-04-2009 16:03	Text Document	2 KB
g0pB_taskb.txt	03-04-2009 16:03	Text Document	2 KB
g0pB_taskc.txt	03-04-2009 16:03	Text Document	2 KB
g0pB_taskd.txt	03-04-2009 16:03	Text Document	2 KB
g0pB_taske.txt	03-04-2009 16:03	Text Document	2 KB
g0pC_taska.txt	03-04-2009 16:03	Text Document	2 KB
g0pC_taskb.txt	03-04-2009 16:03	Text Document	2 KB
g0pC_taskc.txt	03-04-2009 16:03	Text Document	2 KB
g0pC_taskd.txt	03-04-2009 16:03	Text Document	1 KB
g0pC_taske.txt	03-04-2009 16:03	Text Document	1 KB
g0pD_taska.txt	03-04-2009 16:03	Text Document	2 KB
g0pD_taskb.txt	03-04-2009 16:03	Text Document	1 KB
g0pD_taskc.txt	03-04-2009 16:03	Text Document	2 KB
g0pD_taskd.txt	03-04-2009 16:03	Text Document	1 KB
g0pD_taske.txt	03-04-2009 16:03	Text Document	1 KB
g0pE_taska.txt	03-04-2009 16:03	Text Document	2 KB
g0pE_taskb.txt	03-04-2009 16:03	Text Document	1 KB
g0pE_taskc.txt	03-04-2009 16:03	Text Document	1 KB
g0pE_taskd.txt	03-04-2009 16:03	Text Document	1 KB
g0pE_taske.txt	03-04-2009 16:03	Text Document	1 KB
g1pA_taska.txt	02-04-2009 09:30	Text Document	2 KB
g1pA_taskb.txt	02-04-2009 09:30	Text Document	2 KB
g1pA_taskc.txt	02-04-2009 09:30	Text Document	2 KB

Figure 2: Input Text Files(Data Set)

We are using below images to build histogram model and if any suspicious image similarity finds with this histogram then plagiarism will be detected. See below images used to build histogram model

DOI: 10.48175/IJARSCT-9243

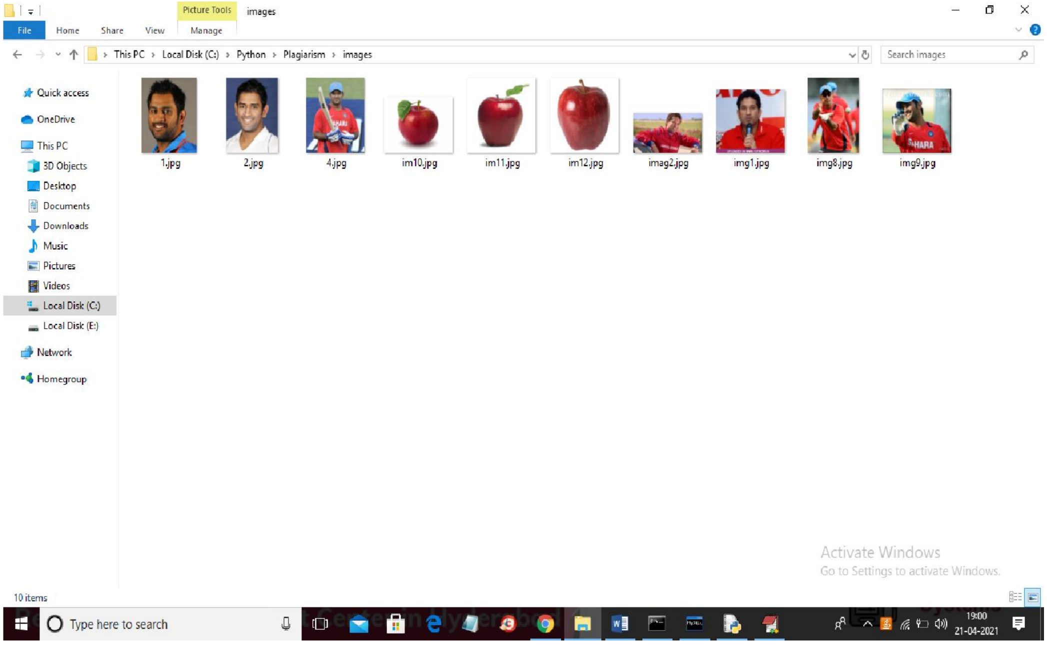


Figure 3: Input Image Files(Data Set

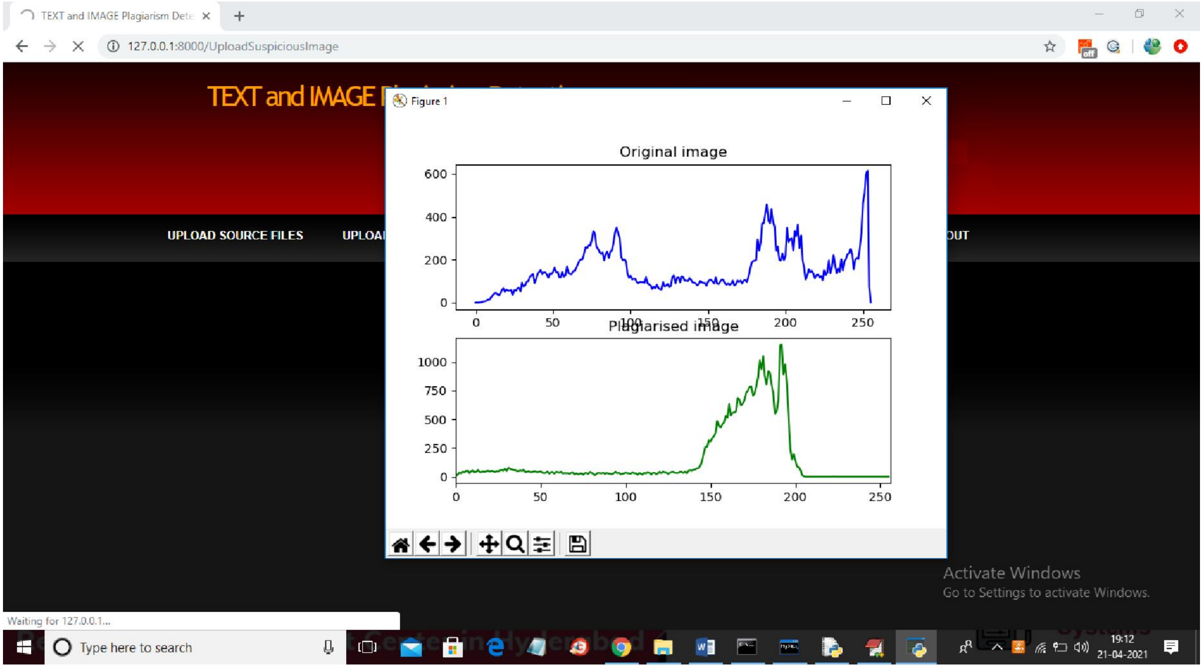
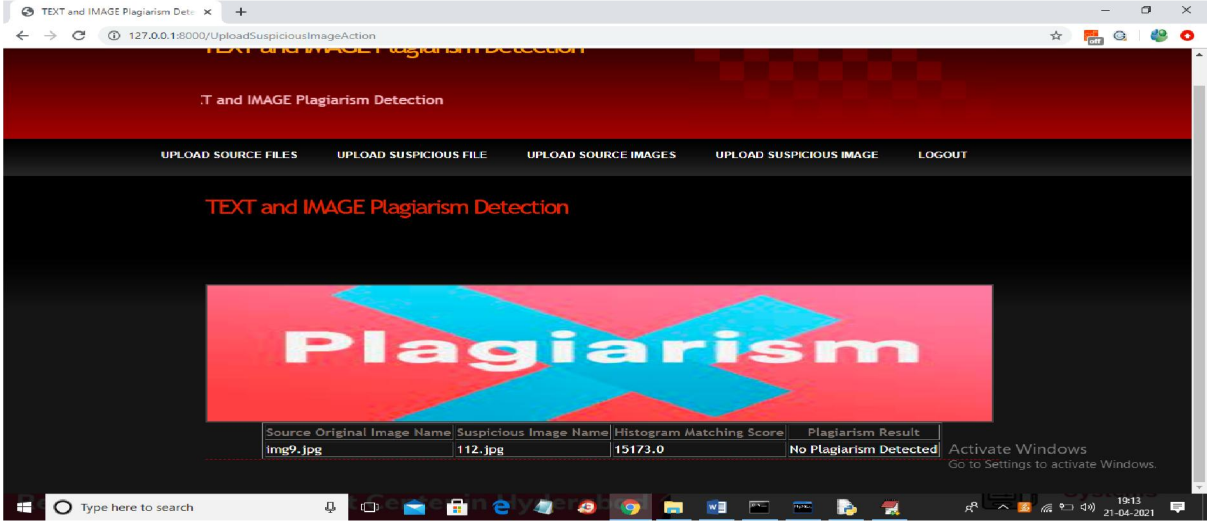


Figure 4: Graph of Output

In above screen we can see for database image and uploaded image we generated histogram and we can see there is no match in histogram so no plagiarism will be detected and now close above graph to get below result.

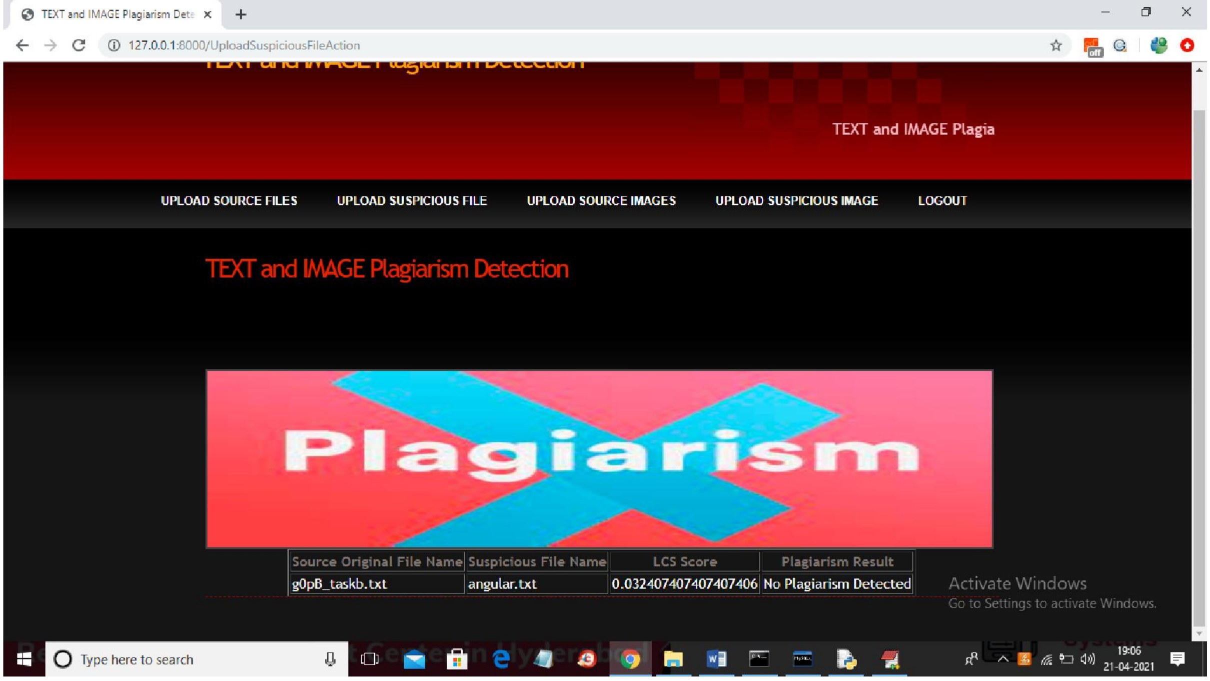
DOI: 10.48175/IJARSCT-9243


127.0.0.1:8000/UploadSuspiciousImageAction
T and IMAGE Plagiarism Detection
UPLOAD SOURCE FILES
UPLOAD SUSPICIOUS FILE
UPLOAD SOURCE IMAGES
UPLOAD SUSPICIOUS IMAGE
LOGOUT
TEXT and IMAGE Plagiarism Detection
Plagiarism

Source Original Image Name	Suspicious Image Name	Histogram Matching Score	Plagiarism Result
img9.jpg	112.jpg	15173.0	No Plagiarism Detected

Figure 5: Output of Image plagiarism checking

In above screen histogram pixel matching score is 15173 out of 40000 pixels so image is not plagiarised and now upload image from "images" folder and see result.



Source Original File Name	Suspicious File Name	LCS Score	Plagiarism Result
g0pB_taskb.txt	angular.txt	0.032407407407407406	No Plagiarism Detected

Figure 6: Output of text Plagiarism checking

In above screen angular.txt file matched very little with g)pB_taskb.txt corpus file and we got similarity score as 0.03 so no plagiarism detected and now upload any file from corpus and see result.

Similarly u can upload any text file and image and test the application.

VII. CONCLUSION

Plagiarism involves converting the existing information in modified format. Today it is found in almost all fields of human activities because use of internet is high, so a lot of attention is given to identify and detect plagiarism. Some experimental results show that in general there is improvement performance in the use of hybrid machine learning methods and techniques in the case of plagiarism. However, the hybrid method does not always produce better and accurate performance. So we have designed a process using machine learning method i.e k-NN which will improve the performance and accuracy. Comparing all methods in this area, we can conclude that the k-nearest neighbour method is

DOI: 10.48175/IJARSCT-9243

much useful in pattern recognition as well as to find copied dataset(text,image) to detect plagiarism. Our method provide more accuracy and efficiency to detect plagiarism. For this, we have implemented the technique which shows how a text set and image set is parsed and checks a particular file with related existing files for plagiarism detection

REFERENCES

[1]. A Selamat, IMI Subroto and Choon-Ching Ng, "Arabic Script Web Page Language Identification Using HybridKNN Method," International Journal of Computational Intelligence and Applications, 2009, pp. 315-343.

[2]. Ahmad Gull Liaqat and Aijaz Ahmad, "Plagiarism Detection in Java Code," Degree Project, Linnaeus University, June 2011, pp. 1-7.

[3]. Upul Bandara and Gamini Wijayrathna ,"Detection of Source Code Plagiarism Using Machine Learning Approach," International Journal of Computer Theory and Engineering, Vol. 4, No. 5, October 2012, pp.674-678.

[4]. Imam Much IbnuSubroto and Ali Selamat, "Plagiarism Detection through Internet using Hybrid Artificial Neural Network and Support Vectors Machine," TELKOMNIKA, Vol.12, No.1, March 2014, pp. 209-218.

[5]. BarrónCedeño, A., & Rosso, "On automatic plagiarism detection based on n-grams comparison," In Advances in Information Retrieval, Vol. 5478. Lecture Notes in Computer Science, pp. 696–700, Springer.

DOI: 10.48175/IJARSCT-9243