

A
PROJECT STAGE -II REPORT
ON
“MACHINE LEARNING APPROACH FOR PLAGIARISM DETECTION OF
IMAGE AND TEXT CONTENT”

Submitted in partial fulfillment of the requirement for the award of the Degree of

BACHELOR OF TECHNOLOGY
IN
COMPUTER SCIENCE AND ENGINEERING



Submitted By

B. SINDHUJA	206B1A0510
G. AKANKSHA	206B1A0530
G. MEGHANA	206B1A0533
K. HARSHITHA	206B1A0553

Under the esteemed guidance of

D.VENU GOPAL,M.Tech,Ph.D



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
KAKATIYA INSTITUTE OF TECHNOLOGY & SCIENCE FOR WOMEN
MANIKBHANDAR, NIZAMABAD, 503003
(Approved by AICTE and Affiliated to JNTUH.)

(2020-2024)



KAKATIYA INSTITUTE OF TECHNOLOGY AND SCIENCE FOR WOMEN

Manikbhandar, Nizamabad - 503003. Ph: 08462-281077.

Approved by AICTE and affiliated to JNTUH - Hyderabad.

Website: www.kitw.ac.in

Dept_email_id: kitw.cse@gmail.com

Date:

CERTIFICATE

This is to certify that the project stage II report entitled “**MACHINE LEARNING APPROACH FOR PLAGIARISM DETECTION OF IMAGE AND TEXT CONTENT**” is a record of bonafied work carried out by **B.SINDHUJA(206B1A0510), G.AKANKSHA(206B1A0530), G.MEGHANA(206B1A0533), K.HARSHITHA(206B1A0553)** students of B.Tech, under our supervision and guidance in partial fulfillment for the award of **Bachelor of Technology** in **Computer Science and Engineering** during the academic year 2023-2024.

PROJECT GUIDE

D.VENU GOPAL

M.Tech,Ph.D

HOD

M.NAGARANI

Assoc.Prof,M.Tech

EXTERNAL EXAMINER

PRINCIPAL

Dr.S. SELVA KUMAR RAJA

B.E, M.E, Ph.D

INDEX

CONTENTS	PAGE NO.
ACKNOWLEDGEMENT	I
DECLARATION	II
ABSTRACT	III
1. INTRODUCTION	1-2
2. LITERATURE SURVEY	3-4
3. SYSTEM ANALYSIS	5-7
3.1 EXISTING SYSTEM	
3.2 PROPOSED SYSTEM	
3.3 SYSTEM REQUIREMENTS	
3.4 SYSTEM STUDY	
4. SYSTEM DESIGN	8-13
4.1 SYSTEM ARCHITECTURE	
4.2 UML DIAGRAMS	
4.3 IMPLEMENTATION	
5. SOFTWARE ENVIRONMENT	14-15
6. MACHINE LEARNING	16-20
7. MODULES USED IN PROJECT	21-22
8. SYSTEM TEST	23-26
	27
9. K-MEANS ALGORITHM	28-36
10. OUTPUT RESULT	37
CONCLUSION	38-39
BIBLIOGRAPHY	

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to our Guide **D.VENU GOPAL,M.Tech,Ph.D** of Computer Science and Engineering, KITW who extended his unconditional support and spared their valuable time with their patience and valuable suggestions regarding our project.

We would like to wish to give a special note of thanks to **HOD, Dr.M.NAGARANI,Assoc.Prof,M.Tech,Ph.D**, CSE Dept. KITW, for her unique way of inspiring students through clarity of thought, enthusiasm and care. Her constant encouragement and assistance are very helpful and made our effort a success.

We are also grateful to the **Dr.S.SELVA KUMAR RAJA,B.E,M.E,Ph.D, Principal**, KITW for providing us with the facilities and resources required for the successful completion of our project. We even thank him for his valuable suggestions at the time of project which encouraged us to give our best in project.

We would also like to thank our faculty and supporting staff of **Computer Science and Engineering Department** and all other departments for their kind co-operation directly or indirectly in making the project a successful one.

Finally, we want to deeply acknowledge all our friends and family members who have encouraged us during the preparation of our project.

By

B. SINDHUJA	206B1A0510
G. AKANKSHA	206B1A0530
G. MEGHANA	206B1A0533
K. HARSHITHA	206B1A0553

DECLARATION

We **B.SINDHUJA, G.AKANKSHA, G.MEGHANA, K.HARSHITHA** hereby declare that this project report has been carried out entirely under the esteemed guidance of **D.VENU GOPAL,M.Tech,Ph.D** for the partial fulfillment of the award of the degree of **Bachelor of Technology in Computer Science and Engineering** at **Kakatiya Institute of Technology & Science for Women**, Manikbhandar, Nizamabad, Affiliated to JNTUH and further it has not been submitted to any other university or institutions for the award of any other degree.

By

B. SINDHUJA	206B1A0510
G. AKANKSHA	206B1A0530
G. MEGHANA	206B1A0533
K. HARSHITHA	206B1A0553

ABSTRACT

In an educational environment, plagiarism is a crucial task that needs to be identified, in recent years all known journals and conferences, as well as universities, request a plagiarism report from students and researchers to prove the originality of published text or scientific paper. Plagiarism detection usually checks the text content via many of the platforms which are available for productive use reliably identifying copied text or near-copies of text and these systems usually fail to detect the images, and files plagiarism since it is originally built for text mainly. In this project, we suggest an adaptive, scalable, and extensible, robust method for image plagiarism which is tested in designs collect from department of architecture University of Technology, this method mainly compare the data (designs images) entered to the system with data sets saved in the database mainly these designs are saved as feature which is one of the artificial intelligence algorithms and match by using k-mean clustering and the similarity check is done with threshold used 40% which can be changed to an accepted levels when needed. Using the k-mean algorithm in clustering, which is a robust artificial intelligence clustering algorithm giving us a strong system that is not discarding any feature extracted from the image. In this project, data sets consist of 45 samples as training images saved and used in the system as the system database and using 48 samples as testing images which consist of original and forgery designs. These testing images were evaluated with 100% matching rate and 81% matching accuracy rating. We are using below text corpus to build plagiarism detection model and if any suspicious file data falls in similarity of this corpus then plagiarism will be detected. This corpus you can see inside 'corpus20090418' folder. 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.

1. INTRODUCTION

1.1 MOTIVATION

There are two main types of plagiarism as Text Based Plagiarism and Image Based Plagiarism. Text Based Plagiarism includes ‘copying textual information available from internet or other resources without proper permission and presenting it as their own’ Image Based plagiarism includes "copying an image or portions of an image from the Internet or from classroom resources without permission or proper acknowledgment.” Hashing techniques are used in the process of plagiarism detection. Here are different algorithms for plagiarism. Here we are using corpus for image and text.

1.2 PROBLEM DEFINITION

The corpus and the measures form the first controlled evaluation environment dedicated to plagiarism detection. Unlike other tasks in natural language processing and information retrieval, it is not possible to publish a collection of real plagiarism cases for evaluation purposes since they cannot be properly anonymized. Therefore, current evaluations found in the literature are incomparable and often not even reproducible. Our contribution in this respect is a newly developed large-scale corpus of artificial plagiarism and new detection performance measures tailored to the evaluation of plagiarism detection algorithms

1.3 OBJECTIVE OF PROJECT

We aimed to create a corpus that could be used for the development and evaluation of plagiarism detection systems that reflects the types of plagiarism practiced by students in an academic setting as far as realistically possible.

In this technological era, social media has a major role in people’s daily life. Most people share text, images, and videos on social media frequently (e.g. Twitter, Snapchat, Facebook, and Instagram). Images are one of the most common types of media share among users on social media. So, there is a need for monitoring of images contained in social media. It has become easy for individuals and small groups to fabricate these images and disseminate them widely in a very short time, which threatens the credibility of the news and public confidence in the means of

social communication. This research attempted to propose an approach to extracting image content, classify it and verify the authenticity of digital images and uncover manipulation. Instagram is one of the most important websites and mobile image sharing applications on social media. This allows users to take The International journal of analytical and experimental modal analysis Volume XIII, Issue IX, September/2021 ISSN NO:0886-9367 Page No: 243 photos, add digital photographic filters and upload pictures. There are many unwanted contents in Instagram's posts such as threats and forged images, which may cause problems to society and national security. This research aims to build a model that can be used to classify Instagram content (images) to detect any threats and forged images. The model was built using deep learning algorithms, specifically Convolutional Neural Networks (CNNs), including the AlexNet network and transfer learning with AlexNet. The results showed that the proposed Alex net network offers more accurate detection of fake images compared to the other techniques with 97%. The results of this research will be helpful in monitoring and tracking in the shared images in social media for unusual content and forged images detection and to protect social media from electronic attacks and threats.

2. LITERATURE SURVEY

The previous work explains the various approaches that were undertaken and researched in this context of work approach

2.1 Convolutional Neural Networks for fake news detection

AUTHORS: L. Zheng, Y. Yang, J. Zhang, Q. Cui, X. Zhang, Z. Li, et al

The identification of fake news and images is very difficult, as fact-finding of news on a pure basis remains an open problem, and few existing models can be used to resolve the problem. It has been proposed to study the problem of “detecting false news.” Through a thorough investigation counterfeit news. Many useful properties are determined from text words and pictures used in counterfeit news. There are some hidden characteristics in words and images used in fake news, which can be identified through a collection of hidden properties derived from this mode 1 through various layers. A pattern called TI -CNN has been proposed. By displaying clear and embedded features in a unified space, TI -CNN is trained with both text and image information at the same time.

2.2 Machine Learning implementation for identifying fake accounts in social network

AUTHORS: R. Raturi

Maturity's 2018 architecture was proposed to identify counterfeit accounts in social networks, especially on Facebook. In this research, a machine learning feature was used to better predict fake accounts, based on their posts and the location mentioned their social networking walls. Support Vector Machine (Sid) and Complement Naive Bayes (CNB) were used in this process, to validate content based on text classification and data analysis. The analysis of the data focused on the collection of offensive words, and the number of times they were repeated. For Facebook, SVM shows a 97% resolution where CNB shows 95% accuracy in rec data is not properly validated before publishing.

2.3 Detection and localization of image forgeries using resampling features and deep learning.

AUTHORS: J.Bunk,J.Bappy,H.Mohammed,T.M.Nataraj,L.,Flenner,A.,Manjunath,B.,etal.

In 2017 study by Bunk eta, two systems were proposed to detect and localize fake images using a mix of re -sampling properties and deep learning. In the initial system, the Radon conversion

of re sampling properties is determined on overlapping pictures corrections. Deep learning classifiers and a Gaussian conditional domain pattern are then used to construct a heat map. A random Walkers e.g., organizing Bag of Words (BOW)-based counterfeit accounts. The result so the study confirmed that the main problem related to the safety of social networks is that authentication method uses total areas. In the next system, for identification and localization, software re sampling properties is passed on overlapping object patches over a long-term memory (LSTM)-based network. In addition, the detection/localization performance of both systems was compared. The results confirmed that both systems are active in detecting and settling digital image fraud.

2.4 Detecting fake news with Machine Learning method.

AUTHORS: S. Aphiwongsophon & P. Chongstitvatana

Phonographic and Longstanding, filamentous automated learning techniques to detect counterfeit news. There common techniques were used in the experiments: Bayesian, Neural Network and Support Vector Machine (SVM). The normalization method is a major step to preprocessing data before using the automatic learning method to sort information. There sluts show Naive Bayes to have a 96.08% accuracy in detecting counterfeit news. There are two other advanced methods, the Neural Network Machine and the Support Network (SVM), which achieve 99.90% accuracy.

2.5 Fake image detection using Machine Learning.

AUTHORS: M. Villan, A. Kuruvilla, K.J. Paul, & E.P. Elias

A neural network was successfully trained by analyzing the 4000 fake and 4000 real images error level. The trained neural network has succeeded in identifying the image as fake or real, with a high success rate of 83%. The results showed that using this application on mobile platforms significantly reduces the spread of fake images across social networks. In addition, this can be used as a false image verification method in digital authentication, court evidence assessment, etc. It develops and tests reliable fake image detection program by combining the resultant data analysis (40%) and neural network output (60%).

3. SYSTEM ANALYSIS

3.1 EXISTING SYSTEM

The existing methodology maybe sufficient for detecting plagiarism of images when the source and suspected image have not been rotated by a large margin, but in case of rotational changes the existing methodology will fail. The proposed methodology will ensure that even if the image is rotated plagiarism is detected if it has occurred or if an attack of rotational change has been made. Also, the existing system is not efficient to detect plagiarism properly for different types of images. The proposed system will ensure that by using adaptive threshold values. The algorithm makes sure that the matching time of the images is less by reducing the search field by a significant factor each time the refinement is done.

3.1.1 DISADVANTAGES

Plagiarism detection usually checks the text content via many of the platforms which are available for productive use reliably identifying copied text or near-copies of text and these systems usually fail to detect the images, and Files plagiarism since it is originally built for text mainly.

3.2 PROPOSED SYSTEM:

The Proposed Text and Image of images plagiarism detection will take input from the user which will be suspected plagiarized image according to the user. Then the Phash value of that image would be generated using the corpus algorithm. Now the input image would be checked for plagiarism against the images in local database. In Database, image is stored with their respective Phash values. The plagiarism detection engine will follow a series of steps to find out plagiarism. This would include calculating hamming distance between Phash values of input image and images in database. At the end based on results achieved in detection engine, results will be displayed. In the Same way text file also detected using corpus algorithm.

3.2.1 ADVANTAGES:

We observe in this connection that the evaluation of plagiarism detection algorithms is not

standardized, i.e., most of the time the algorithms are evaluated on homemade corpora using various different performance measures. This situation renders the existing research almost incomparable.

3.3. SYSTEM REQUIREMENTS

3.3.1 HARDWARE REQUIREMENTS:

- | | | |
|--------------|---|-------------------|
| 1. System | : | Pentium IV 2.4GHz |
| 2. Hard Disk | : | 1TB |
| 3. Monitor | : | 15VGA Color |
| 4. Mouse | : | Logitech |
| 5. Ram | : | 4GB |

3.3.2 SOFTWARE REQUIREMENTS:

- | | | |
|---------------------|---|------------|
| 1. Operating System | : | Windows |
| 2. Coding Language | : | Python 3.7 |

3.4. SYSTEM STUDY

FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

1. ECONOMICAL FEASIBILITY
2. TECHNICAL FEASIBILITY
3. SOCIAL FEASIBILITY

ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus, the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

SOCIAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

4. SYSTEM DESIGN

4.1 SYSTEM ARCHITECTURE:

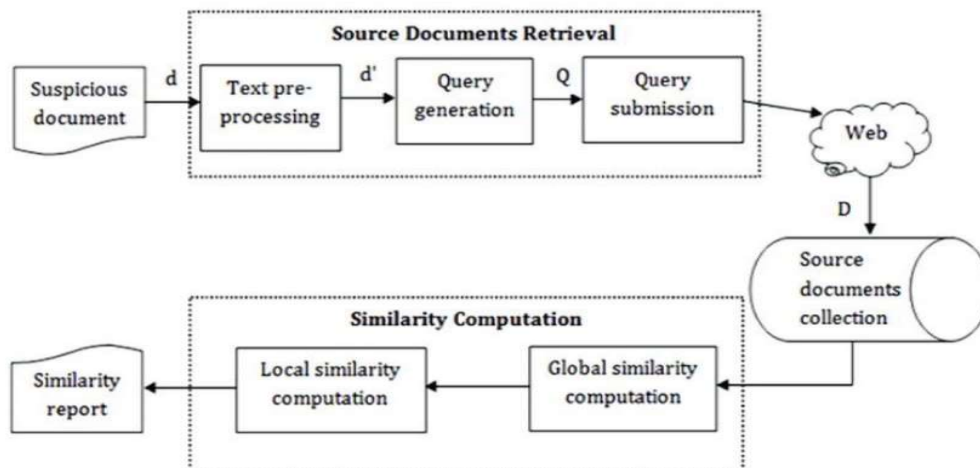


Fig 4.1.1: System Architecture

4.2 UML DIAGRAMS

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.

The goal is for UML to become a common language for creating models of object-oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modeling and other non-software systems.

The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems.

The UML is a very important part of developing objects-oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

GOALS:

1. The Primary goals in the design of the UML are as follows.
2. Provide users a ready-to-use, expressive visual modeling Language so that they can develop and exchange meaningful models.
3. Provide extendibility and specialization mechanisms to extend the core concepts.
4. Be independent of particular programming languages and development process.
5. Provide a formal basis for understanding the modeling language.
6. Encourage the growth of OO tools market.
7. Support higher level development concepts such as collaborations, frameworks, patterns and components.
8. Integrate best practices.

4.2.1 USE CASE DIAGRAM:

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

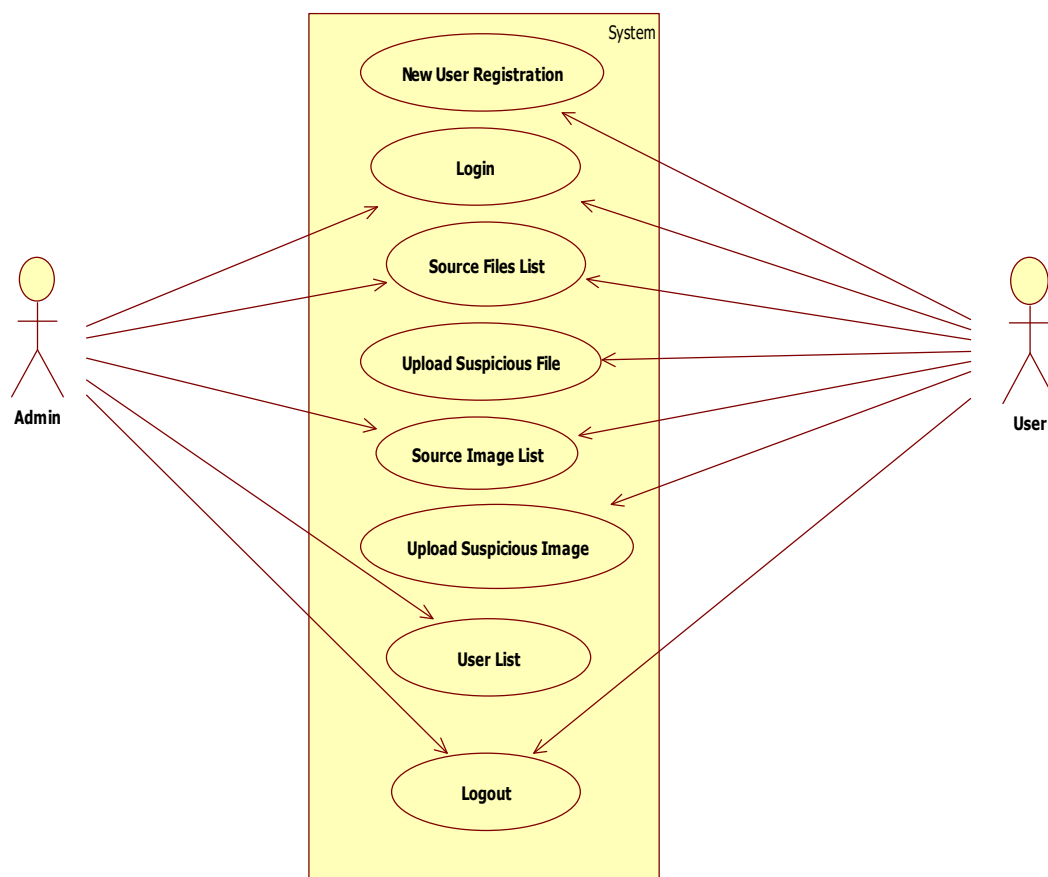


Fig 4.2.1.1: Use Case Diagram

4.2.2 CLASS DIAGRAM:

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

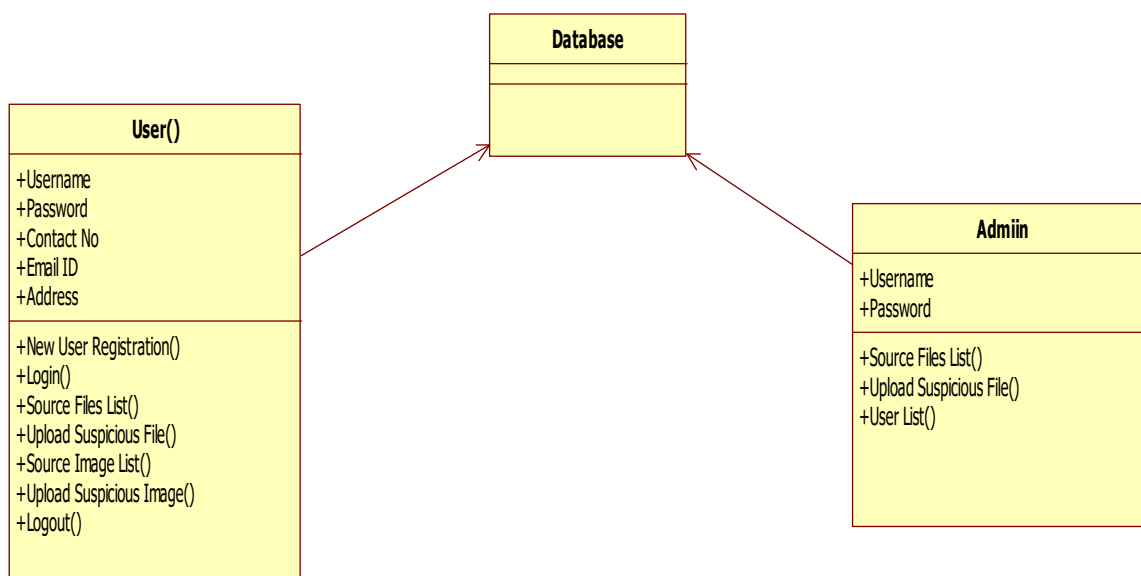


Fig 4.2.2.1: Class Diagram

4.2.3 SEQUENCE DIAGRAM

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

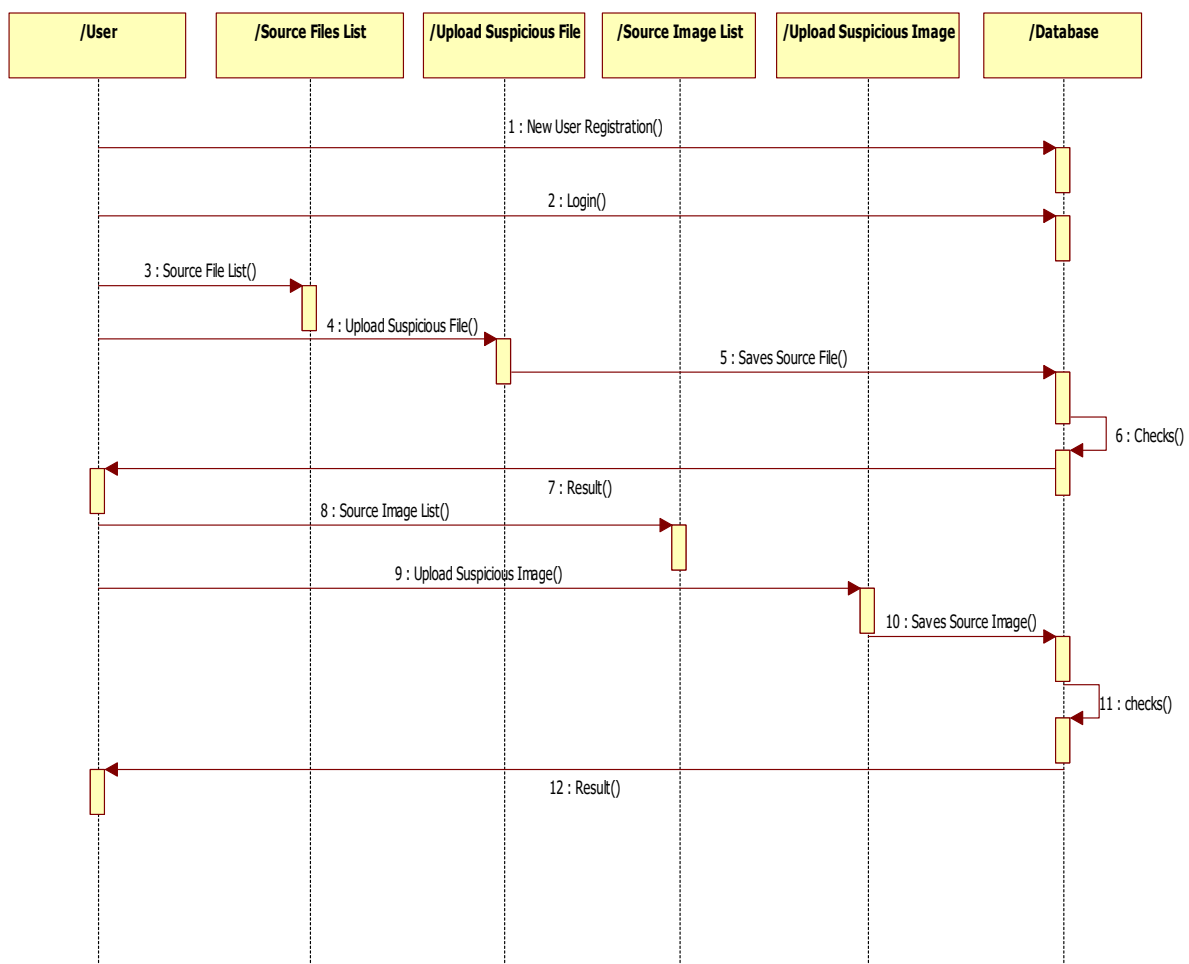


Fig 4.2.3.1: Sequence Diagram

4.3 IMPLEMENTATION:

MODULES:

1. New user Registration

Firstly, user will register in to Application. It helpful to login in to application with username and password.

2. Login

User will login into Application through username and password.

3. Source File List

Folder is created into Upload Source Files link to load all files from corpus folder.

4. Upload Suspicious files

To load suspicious file and get result. User will upload file to Upload Suspicious files the result is execute. LCS score is 1.0 which means 100% matched with corpus file so plagiarism detected and similarly not only this you may enter any text file and get result.

5. Source Image List

In this module from all database images histogram will be calculated and store in array and whenever we upload new test image then both histograms will get matched.

6. Upload Suspicious Image

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. 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. we can both original and uploaded image histogram is matching 100% so plagiarism is detected and now get below result. histogram matching score is 40000 which means all pixels matched so plagiarism is detected in above result.

7. Admin Login

Admin have to login using username and password. After admin successfully login, admin can find the users list containing their information/data.

5. SOFTWARE ENVIRONMENT

What is Python?

1. Below are some facts about Python.
2. Python is currently the most widely used multi-purpose, high-level programming language.
3. Python allows programming in Object-Oriented and Procedural paradigms.
4. Python programs generally are smaller than other programming languages like Java.
5. Programmers have to type relatively less and indentation requirement of the language, makes them readable all the time.
6. Python language is being used by almost all tech-giant companies like – Google, Amazon, Facebook, Instagram, Dropbox, Uber... etc.

The biggest strength of Python is huge collection of standard libraries which can be used for the following –

1. Machine Learning
2. GUI Applications (like Kivy, Tkinter, PyQt etc.)
3. Web frameworks like Django (used by YouTube, Instagram, Dropbox)
4. Image processing (like OpenCV, Pillow)
5. Web scraping (like Scrapy, BeautifulSoup, Selenium)
6. Test frameworks
7. Multimedia

Advantages of Python

1. Extensive Libraries
2. Extensible
3. Embeddable
4. Improved Productivity
5. IOT Opportunities
6. Simple and Easy
7. Readable
8. Object-Oriented
9. Free and Open-Source
10. Portable
11. Interpreted

Advantages of Python over other languages

1. Less Coding

Almost all of the tasks done in Python requires less coding when the same task is done in other languages. Python also has an awesome standard library support, so you don't have to search for any third-party libraries to get your job done. This is the reason that many people suggest learning Python to beginners.

2. Affordable

Python is free therefore individuals, small companies or big organizations can leverage the free available resources to build applications. Python is popular and widely used so it gives you better community support.

3. Python is for everyone

Python code can run on any machine whether it is Linux, Mac or Windows. Programmers need to learn different languages for different jobs but with Python, you can professionally build web apps, perform data analysis and machine learning, automate things, do web scraping and also build games and powerful visualizations. It is an all-rounder programming language.

Disadvantages of Python

1. Speed Limitations
2. Weak in Mobile Computing and Browsers
3. Design Restrictions
4. Underdeveloped Database Access Layers
5. Simple

6. MACHINE LEARNING

6.1 What is Machine Learning

Before we take a look at the details of various machine learning methods, let's start by looking at what machine learning is, and what it isn't. Machine learning is often categorized as a subfield of artificial intelligence, but I find that categorization can often be misleading at first brush. The study of machine learning certainly arose from research in this context, but in the data science application of machine learning methods, it's more helpful to think of machine learning as a means of building models of data.

Fundamentally, machine learning involves building mathematical models to help understand data. "Learning" comes into play when we provide these models with tunable parameters that can be adapted to observed data; in this way, the program can be considered to be "Learning" from the data. Once these models have been fit to previously seen data, they can be used to predict and understand aspects of newly observed data. I'll leave to the reader the more philosophical digression regarding the extent to which this type of mathematical, model-based "learning" is similar to the "learning" exhibited by the human brain. Understanding the problem setting in machine learning is essential to using these tools effectively, and so we will start with some broad categorizations of the types of approaches we'll discuss here.

6.2 Categories of Machine Learning

At the most fundamental level, machine learning can be categorized into two main types: supervised learning and unsupervised learning.

Supervised learning involves modeling the relationship between measured features of data and some label associated with the data. Once this model is determined, it can be used to apply labels to new, unknown data. This is further subdivided into classification tasks and regression tasks: in classification, the labels are discrete categories, while in regression, the labels are continuous quantities. We will see examples of both types of supervised learning in the following section.

Unsupervised learning involves modeling the features of a dataset without reference to any label, and is often described as "letting the dataset speak for itself." These models include tasks such as clustering and dimensionality reduction. Clustering algorithms identify distinct groups of data, while dimensionality reduction algorithms search for more succinct representations of the data.

6.3 Challenges in Machines Learning

While Machine Learning is rapidly evolving, making significant strides with cybersecurity and autonomous cars, this segment of AI as whole still has a long way to go. The reason behind is that ML has not been able to overcome number of challenges. The challenges that ML is facing currently are –

- 1. Quality of data** – Having good-quality data for ML algorithms is one of the biggest challenges. Use of low-quality data leads to the problems related to data preprocessing and feature extraction.
- 2. Time-Consuming task** – Another challenge faced by ML models is the consumption of time especially for data acquisition, feature extraction and retrieval.
- 3. Lack of specialist persons** – As ML technology is still in its infancy stage, availability of expert resources is a tough job.
- 4. No clear objective for formulating business problems** – Having no clear objective and well-defined goal for business problems is another key challenge for ML because this technology is not that mature yet.
- 5. Issue of overfitting & underfitting** – If the model is overfitting or underfitting, it cannot be represented well for the problem.
- 6. Curse of dimensionality** – Another challenge ML model faces is too many features of data points. This can be a real hindrance.
- 7. Difficulty in deployment** – Complexity of the ML model makes it quite difficult to be deployed in real life.

6.4 Applications of Machines Learning

Machine Learning is the most rapidly growing technology and according to researchers we are in the golden year of AI and ML. It is used to solve many real-world complex problems which cannot be solved with traditional approach. Following are some real-world applications of ML

1. Emotion analysis
2. Sentiment analysis
3. Error detection and prevention
4. Weather forecasting and prediction
5. Stock market analysis and forecasting
6. Speech synthesis

7. Speech recognition
8. Customer segmentation
9. Object recognition
10. Fraud detection

Terminologies of Machine Learning

- 1. Model** – A model is a specific representation learned from data by applying some machine learning algorithm. A model is also called a hypothesis.
- 2. Feature** – A feature is an individual measurable property of the data. A set of numeric features can be conveniently described by a feature vector. Feature vectors are fed as input to the model. For example, in order to predict a fruit, there may be features like color, smell, taste, etc.
- 3. Target (Label)** – A target variable or label is the value to be predicted by our model. For the fruit example discussed in the feature section, the label with each set of input would be the name of the fruit like apple, orange, banana, etc.
- 4. Training** – The idea is to give a set of inputs (features) and its expected outputs (labels), so after training, we will have a model (hypothesis) that will then map new data to one of the categories trained on.
- 5. Prediction** – Once our model is ready, it can be fed a set of inputs to which it will provide a predicted output (label).

Types of Machine Learning

- 1. Supervised Learning** – This involves learning from a training dataset with labeled data using classification and regression models. This learning process continues until the required level of performance is achieved.
- 2. Unsupervised Learning** – This involves using unlabeled data and then finding the underlying structure in the data in order to learn more and more about the data itself using factor and cluster analysis models.
- 3. Semi-supervised Learning** – This involves using unlabeled data like Unsupervised Learning with a small amount of labeled data. Using labeled data vastly increases the learning accuracy and is also more cost-effective than Supervised Learning.
- 4. Reinforcement Learning** – This involves learning optimal actions through trial and error. So the next action is decided by learning behaviors that are based on the current state and that will

maximize the reward in the future.

6.5 Advantages of Machine learning

1. Easily identifies trends and patterns

Machine Learning can review large volumes of data and discover specific trends and patterns that would not be apparent to humans. For instance, an e-commerce platform analyzes the purchase histories of its users to help tailor products, deals, and reminders relevant to them. It utilizes these results to display relevant advertisements to them.

2. No human intervention needed (automation)

With ML, you don't need to babysit your project every step of the way. Since it means giving machines the ability to learn, it lets them make predictions and also improve the algorithms on their own. A common example of this is anti-virus software; they learn to filter new threats as they are recognized. ML is also good at recognizing spam.

3. Continuous Improvement

As ML algorithms gain experience, they keep improving in accuracy and efficiency. This lets them make better decisions. Say you need to make a weather forecast model. As the amount of data you have keeps growing, your algorithms learn to make more accurate predictions faster.

4. Handling multi-dimensional and multi-variety data

Machine Learning algorithms are good at handling data that are multi-dimensional and multivariate, and they can do this in dynamic or uncertain environments.

5. Wide Applications

You could be an e-tailer or a healthcare provider and make ML work for you. Where it does apply, it holds the capability to help deliver a much more personal experience to customers while also targeting the right customers.

6.6 DISADVANTAGES OF MACHINE LEARNING

1. Data Acquisition

Machine Learning requires massive data sets to train on, and these should be inclusive/unbiased, and of good quality. There can also be times where they must wait for new data to be generated.

2. Time and Resources

ML needs enough time to let the algorithms learn and develop enough to fulfill their purpose with a considerable amount of accuracy and relevancy. It also needs massive resources to function. This can mean additional requirements of computer power for you.

3. Interpretation of Results

Another major challenge is the ability to accurately interpret results generated by the algorithms. You must also carefully choose the algorithms for your purpose.

4. High error-susceptibility

Machine Learning is autonomous but highly susceptible to errors. Suppose you train an algorithm with data sets small enough to not be inclusive. You end up with biased predictions coming from a biased training set. This leads to irrelevant advertisements being displayed to customers. In the case of ML, such blunders can set off a chain of errors that can go undetected for long periods of time. And when they do get noticed, it takes quite some time to recognize the source of the issue, and even longer to correct it.

7. MODULES USED IN PROJECT

TensorFlow

TensorFlow is a free and open-source software library for dataflow and differentiable programming across a range of tasks. It is a symbolic math library, and is also used for machine learning applications such as neural networks. It is used for both research and production at Google.

TensorFlow was developed by the Google Brain team for internal Google use. It was released under the Apache 2.0 open-source license on November 9, 2015.

NumPy

NumPy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays.

It is the fundamental package for scientific computing with Python. It contains various features including these important ones:

1. A powerful N-dimensional array object
2. Sophisticated (broadcasting) functions
3. Tools for integrating C/C++ and Fortran code
4. Useful linear algebra, Fourier transform, and random number capabilities
5. Besides its obvious scientific uses, Numpy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined using Numpy which allows Numpy to seamlessly and speedily integrate with a wide variety of databases.

Pandas

Pandas is an open-source Python library that provides high-performance data manipulation and analysis tool using its powerful data structures. Python was majorly used for data munging and preparation, with minimal contribution to data analysis. Pandas solved this problem. Using Pandas, we can accomplish five typical steps in the processing and analysis of data, regardless of

the origin of data load, prepare, manipulate, model, and analyze. Python with Pandas is used in a wide range of fields including academic and commercial domains including finance, economics, Statistics, analytics, etc.

Matplotlib

Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shells, the Jupyter Notebook, web application servers, and four graphical user interface toolkits. Matplotlib tries to make easy things easy and hard things possible. You can generate plots, histograms, power spectra, bar charts, error chart , scatter plots, etc., with just a few lines of code. For examples, see the sample plots and thumbnail gallery.

For simple plotting the **pyplot** module provides a MATLAB-like interface, particularly when combined with IPython. For the power user, you have full control of line styles, font properties, axes properties, etc., via an object-oriented interface or via a set of functions familiar to MATLAB users.

Scikit – learn

Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python. It is licensed under a permissive simplified BSD license and is distributed under many Linux distributions, encouraging academic and commercial use.

8 .SYSTEM TEST

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub-assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of tests. Each test type addresses a specific testing requirement.

8.1 TYPES OF TESTS

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input: identified classes of valid input must be accepted.

Invalid Input: identified classes of invalid input must be rejected.

Functions: identified functions must be exercised.

Output: identified classes of application outputs must be exercised.

Systems/Procedures: interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration-oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

White Box Testing

White Box Testing is a testing in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is used to test areas that cannot be reached from a black box level.

Black Box Testing

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box. You cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

Unit Testing

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail.

Test objectives

1. All field entries must work properly.
2. Pages must be activated from the identified link.
3. The entry screen, messages and responses must not be delayed.
4. Verify that the entries are of the correct format
5. No duplicate entries should be allowed
6. All links should take the user to the correct page.

Integration Testing

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.

The task of the integration test is to check that components or software applications, e.g., components in a software system or – one step up – software applications at the company level – interact without error.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

TEST CASES

Test cases1:

Test case for Login form:

FUNCTION:	LOGIN
EXPECTED RESULTS:	Should Validate the user and check hisexistence in database
ACTUAL RESULTS:	Validate the user and checking the user against the database
LOW PRIORITY	No
HIGH PRIORITY	Yes

Test case2:

Test case for User Registration form:

FUNCTION:	USER REGISTRATION
EXPECTED RESULTS:	Should check if all the fields are filled by the user and saving the user to database.
ACTUAL RESULTS:	Checking whether all the fields are field by user or not through validations and saving user.
LOW PRIORITY	No
HIGH PRIORITY	Yes

9. K-MEANS ALGORITHM

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if $K=2$, there will be two clusters, and for $K=3$, there will be three clusters, and so on.

It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.

It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

How the K-means algorithm works

To process the learning data, the K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids.

It halts creating and optimizing clusters when either:

1. The centroids have stabilized — there is no change in their values because the clustering has been successful.
2. The defined number of iterations has been achieved.

10. OUTPUT RESULTS

HOME PAGE



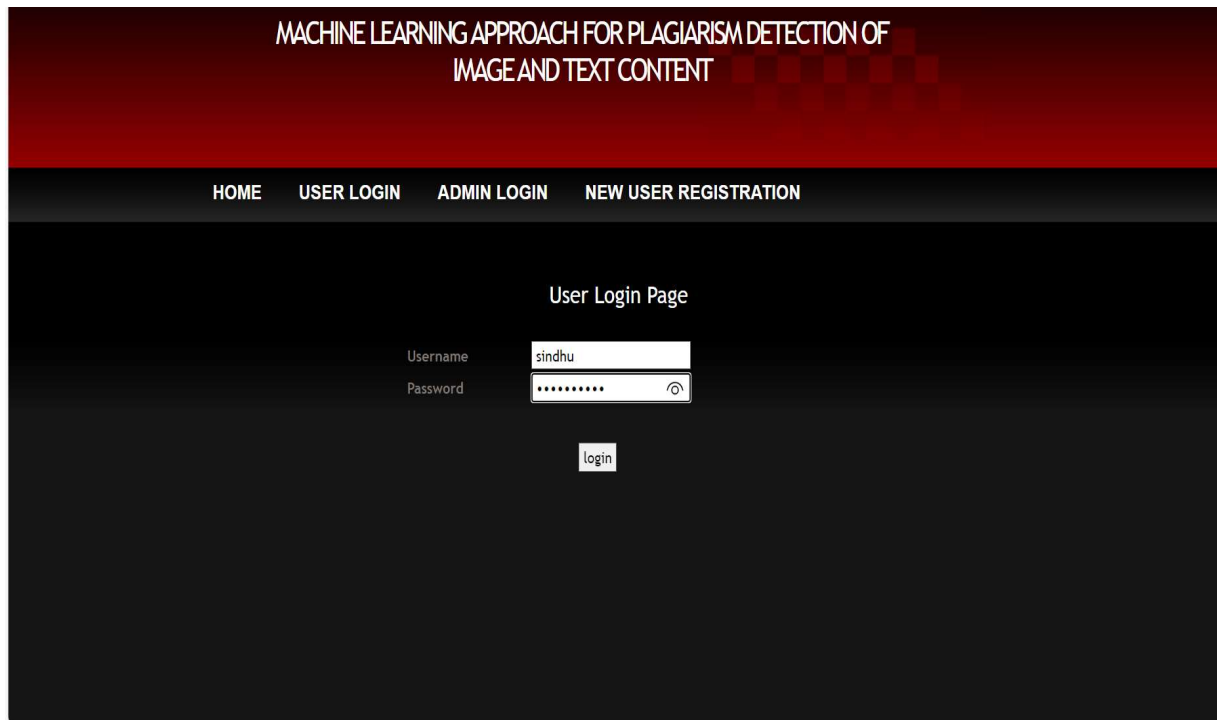
Fig 10.1: Home page

NEW USER REGISTRATION

The screenshot shows the "New User Signup" page. It has the same header and navigation bar as the home page. The main content area is black and contains a "New User Signup" form. The form has five input fields: "Username", "Password", "Contact No", "Email ID", and "Address". Below the input fields is a "Register" button.

Fig 10.2: New User Registration

USER LOGIN PAGE



The screenshot shows the 'User Login Page' of a web application. The header is dark red with the title 'MACHINE LEARNING APPROACH FOR PLAGIARISM DETECTION OF IMAGE AND TEXT CONTENT'. Below the header is a navigation bar with links: HOME, USER LOGIN, ADMIN LOGIN, and NEW USER REGISTRATION. The main content area is dark grey and contains a login form. The form has two input fields: 'Username' with the value 'sindhu' and 'Password' with masked characters. A 'login' button is positioned below the password field.

Fig 10.3: User Login Page

USER HOME PAGE

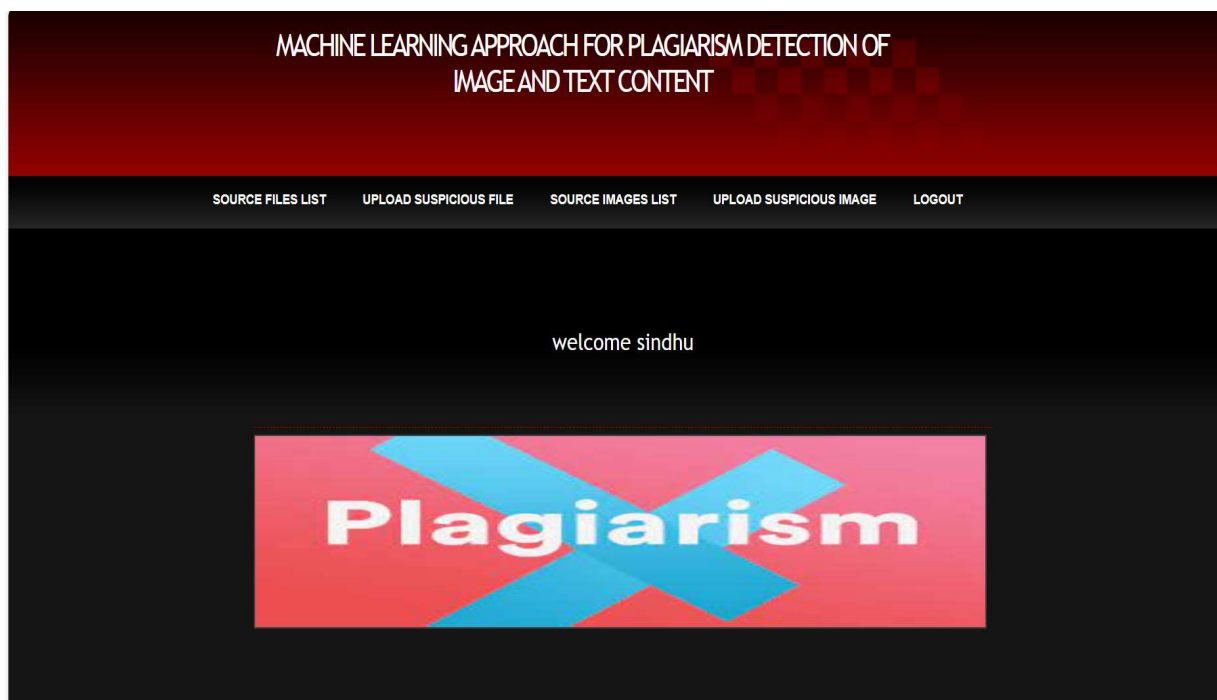


Fig 10.4: User Home Page

SOURCE FILE LIST

MACHINE LEARNING APPROACH FOR PLAGIARISM DETECTION OF
IMAGE AND TEXT CONTENT

SOURCE FILES LIST

UPLOAD SUSPICIOUS FILE

SOURCE IMAGES LIST

UPLOAD SUSPICIOUS IMAGE

LOGOUT

Source File Name	Words in File
g0pA_taska.txt	1
g0pA_taskb.txt	113
g0pA_taskc.txt	129
g0pA_taskd.txt	104
g0pA_taske.txt	112
g0pB_taska.txt	155
g0pB_taskb.txt	122
g0pB_taskc.txt	176
g0pB_taskd.txt	107
g0pB_taske.txt	130
g0pC_taska.txt	104
g0pC_taskb.txt	97
g0pC_taskc.txt	86
g0pC_taskd.txt	88
g0pC_taske.txt	83
g0pD_taska.txt	101
g0pD_taskb.txt	45
g0pD_taskc.txt	93

127.0.0.1:8000/UploadSource

Fig 10.5: Source File List

UPLOAD SUSPICIOUS FILES

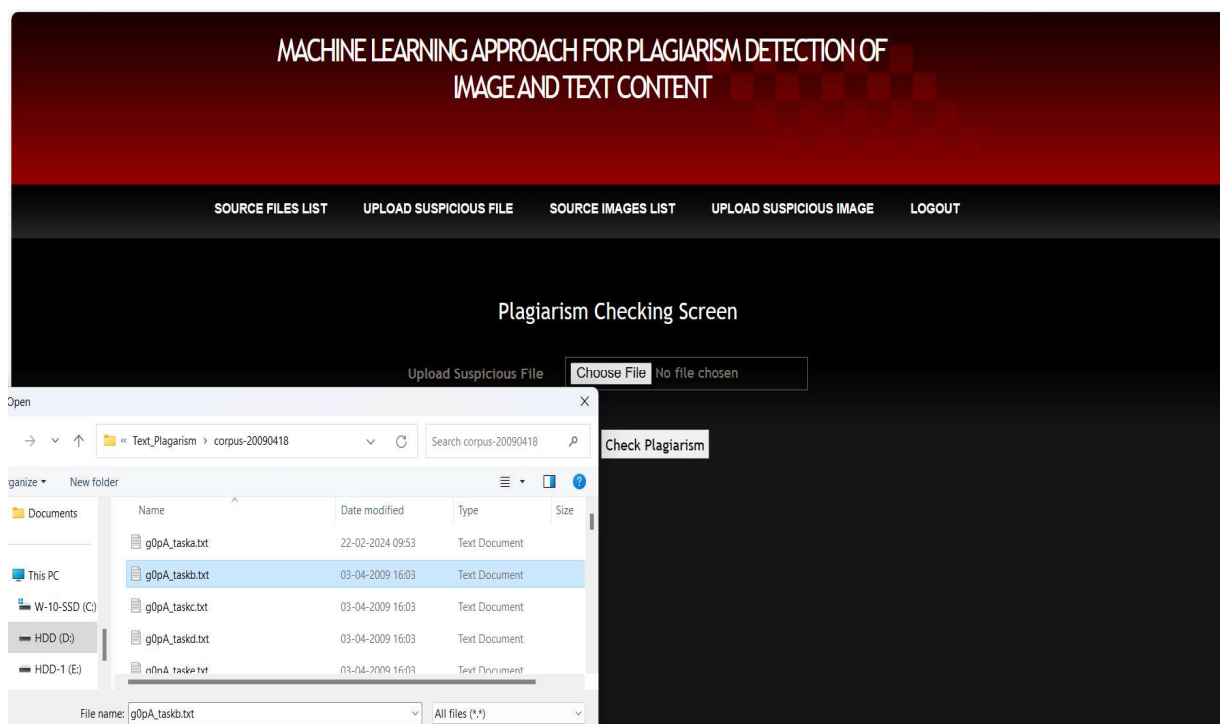


Fig 10.6: Upload Suspicious Files

PLAGIARISM DETECTION RESULTS

MACHINE LEARNING APPROACH FOR PLAGIARISM DETECTION OF IMAGE AND TEXT CONTENT				
SOURCE FILES LIST	UPLOAD SUSPICIOUS FILE	SOURCE IMAGES LIST	UPLOAD SUSPICIOUS IMAGE	LOGOUT
Source Original File Name	Suspicious File Name	LCS Score	Plagiarism Result	Plagiarism Percentage
g0pA_taskb.txt	g0pA_taskb.txt	1.0	Plagiarism Detected	100.0

Fig 10.7: Plagiarism Detection Results

PLAGIARISM DETECTION RESULTS

MACHINE LEARNING APPROACH FOR PLAGIARISM DETECTION OF IMAGE AND TEXT CONTENT				
SOURCE FILES LIST	UPLOAD SUSPICIOUS FILE	SOURCE IMAGES LIST	UPLOAD SUSPICIOUS IMAGE	LOGOUT
Source Original File Name	Suspicious File Name	LCS Score	Plagiarism Result	Plagiarism Percentage
sample.txt	t.txt	0.5	No Plagiarism Detected	49.999999999999986

Fig 10.8: Plagiarism Detection Results

SOURCE IMAGE LIST

MACHINE LEARNING APPROACH FOR PLAGIARISM DETECTION OF
IMAGE AND TEXT CONTENT

SOURCE FILES LIST

UPLOAD SUSPICIOUS FILE

SOURCE IMAGES LIST

UPLOAD SUSPICIOUS IMAGE

LOGOUT

Source File Name	Words in File
g0pA_taska.txt	1
g0pA_taskb.txt	113
g0pA_taskc.txt	129
g0pA_taskd.txt	104
g0pA_taske.txt	112
g0pB_taska.txt	155
g0pB_taskb.txt	122
g0pB_taskc.txt	176
g0pB_taskd.txt	107
g0pB_taske.txt	130
g0pC_taska.txt	104
g0pC_taskb.txt	97
g0pC_taskc.txt	86
g0pC_taskd.txt	88
g0pC_taske.txt	83
g0pD_taska.txt	101
g0pD_taskb.txt	45
g0pD_taskc.txt	93

Fig 10.9: Source Image List

UPLOAD SUSPICIOUS IMAGE

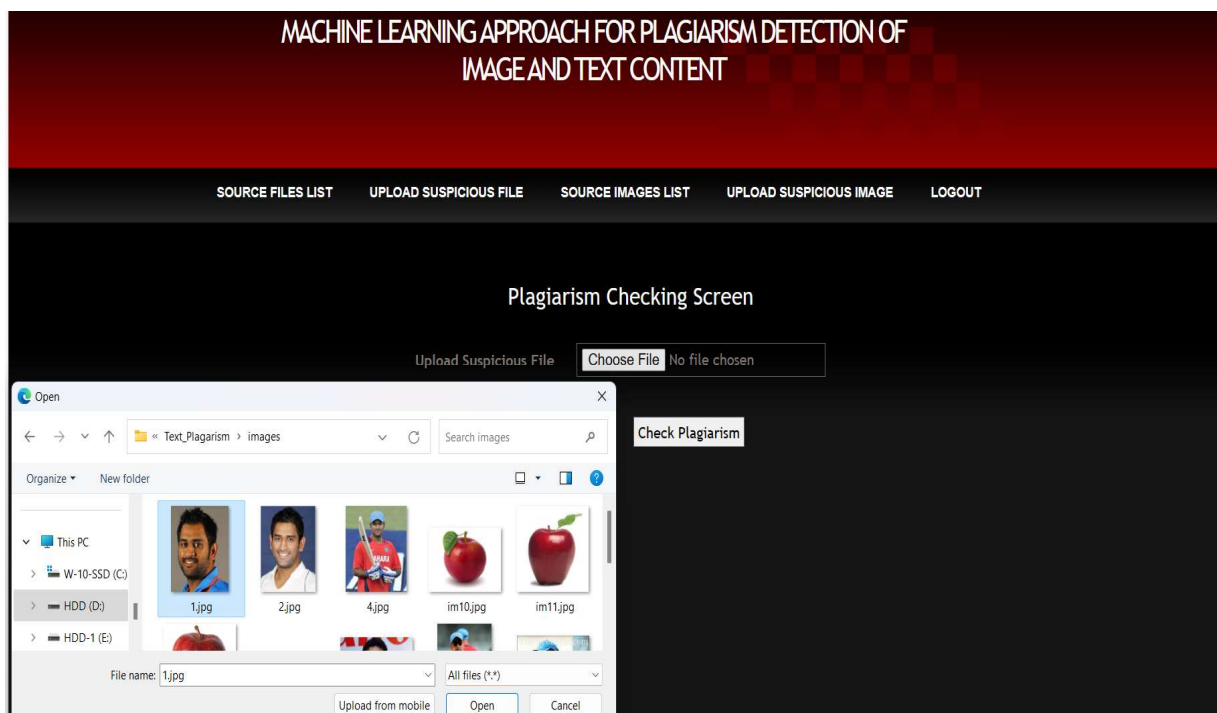


Fig 10.10: Upload Suspicious Image

PLAGIARISM DETECTION RESULTS

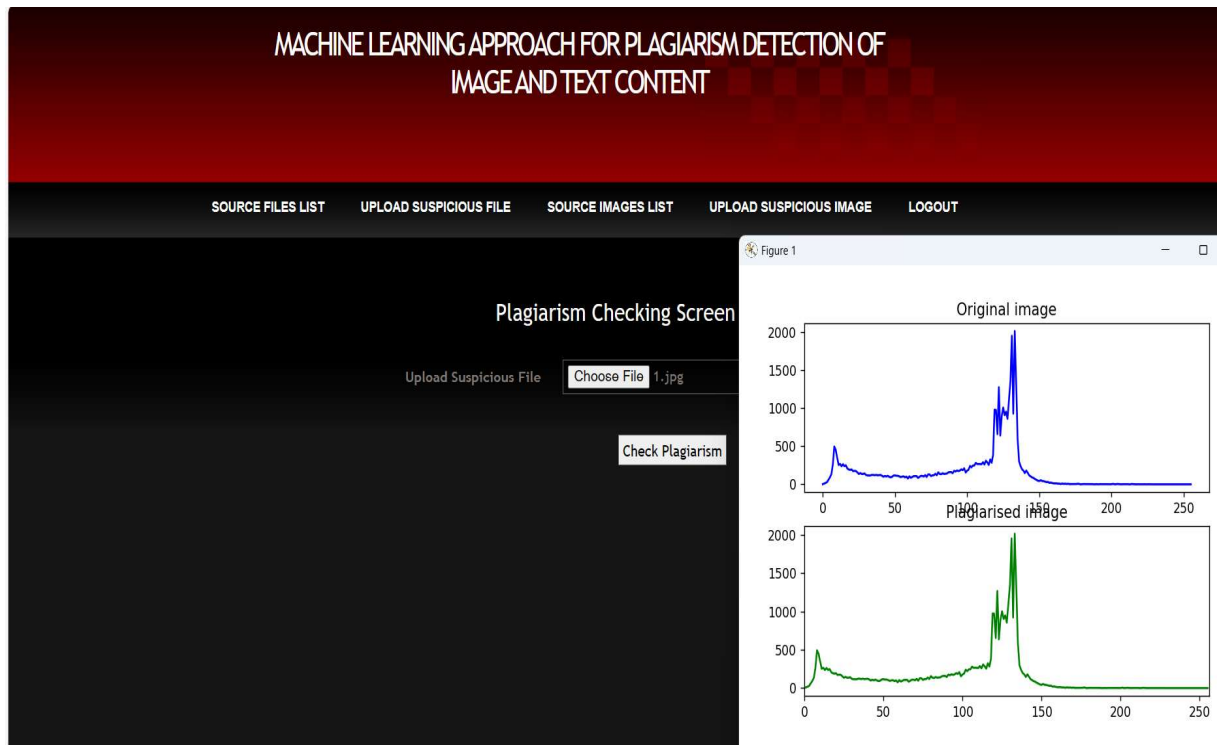


Fig 10.11: Plagiarism Detection Results

PLAGIARISM DETECTION RESULTS

MACHINE LEARNING APPROACH FOR PLAGIARISM DETECTION OF IMAGE AND TEXT CONTENT											
SOURCE FILES LIST	UPLOAD SUSPICIOUS FILE	SOURCE IMAGES LIST	UPLOAD SUSPICIOUS IMAGE								
LOGOUT											
<table border="1"> <thead> <tr> <th>Source Original Image Name</th><th>Suspicious Image Name</th><th>Histogram Matching Score</th><th>Plagiarism Result</th></tr> </thead> <tbody> <tr> <td>1.jpg</td><td>1.jpg</td><td>40000.0</td><td>Plagiarism Detected</td></tr> </tbody> </table>				Source Original Image Name	Suspicious Image Name	Histogram Matching Score	Plagiarism Result	1.jpg	1.jpg	40000.0	Plagiarism Detected
Source Original Image Name	Suspicious Image Name	Histogram Matching Score	Plagiarism Result								
1.jpg	1.jpg	40000.0	Plagiarism Detected								

Fig 10.12: Plagiarism Detection Results

PLAGIARISM DETECTION RESULTS



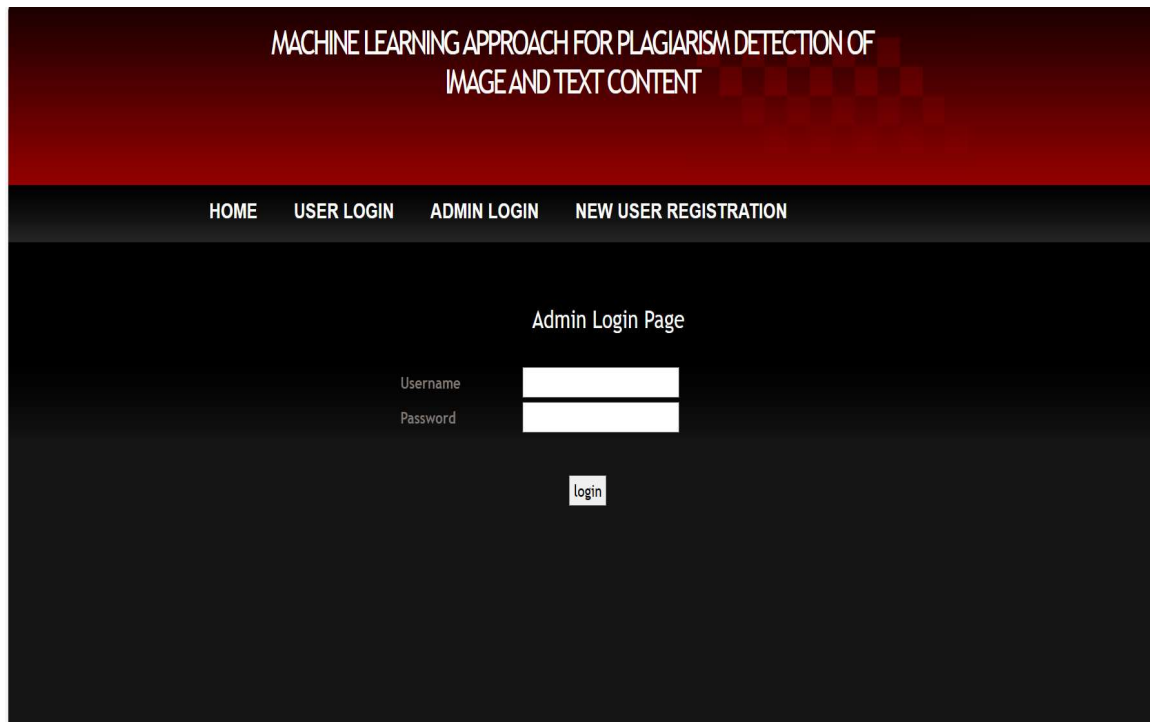
Fig 10.13: Plagiarism Detection Results

PLAGIARISM DETECTION RESULTS

MACHINE LEARNING APPROACH FOR PLAGIARISM DETECTION OF IMAGE AND TEXT CONTENT			
SOURCE FILES LIST	UPLOAD SUSPICIOUS FILE	SOURCE IMAGES LIST	UPLOAD SUSPICIOUS IMAGE
LOGOUT			
Source Original Image Name	Suspicious Image Name	Histogram Matching Score	Plagiarism Result
im11.jpg	Login.jpg	30402.0	No Plagiarism Detected

Fig 10.14: Plagiarism Detection Results

ADMIN LOGIN PAGE



The screenshot shows the Admin Login Page of a web application. At the top, a dark red header contains the text "MACHINE LEARNING APPROACH FOR PLAGIARISM DETECTION OF IMAGE AND TEXT CONTENT". Below the header is a black navigation bar with white links: "HOME", "USER LOGIN", "ADMIN LOGIN", and "NEW USER REGISTRATION". The main content area is black and features the title "Admin Login Page" in white. Below the title are two white input fields labeled "Username" and "Password". A white "login" button is positioned below the password field.

Fig 10.15: Admin Login Page

ADMIN HOMEPAGE

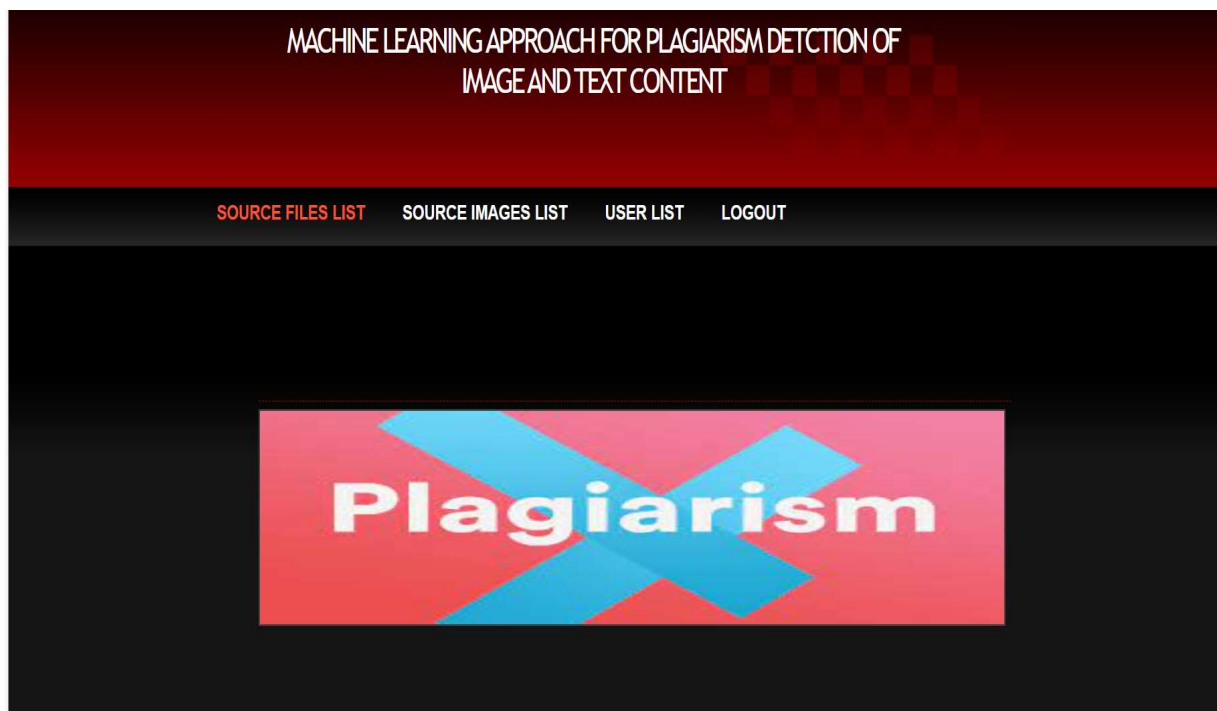


Fig 10.16: Admin Home Page

USER LIST

MACHINE LEARNING APPROACH FOR PLAGIARISM DETECTION OF IMAGE AND TEXT CONTENT			
SOURCE FILES LIST	SOURCE IMAGES LIST	USER LIST	LOGOUT
Users List			
Username	contact No	Email	Address
sindhu	9632587410	sindhu@gmail.com	Nizamabad
Meghana	8965749606	meghana@gmail.com	Hyderabad
Harshitha	9874536210	harshitha@gmail.com	Nizamabad
Akanksha	8965741230	akanksha@gmail.com	Hyderabad
Kumari	7896541305	kumari@gmail.com	Karimnagar

Fig 10.17: User List

CONCLUSION

Corpus is the first standardized corpus dedicated to the evaluation of automatic plagiarism detection and was successfully employed in the First International Competition on Plagiarism Detection. We believe that our corpus and the performance measures will become an effective means to evaluate future plagiarism detection research. Currently, an improved version of the corpus is being constructed.

The consumption of news is increasing day by day in cyberspace than the traditional media. Due to its increasing popularity and user-friendly access it leaves a huge impact on individuals and society. Therefore, in this model we have found a way to detect such fake news in both the forms of text and image by using the Logistic regression model. By redirecting the fake news to the authorized website (cyber- crime department), we hereby frame a high social impact and thus it reduces the spreading of false news distinctly.

BIBLIOGRAPHY

1. G. Mohamed Kandahar, "100 Social Media Statistics You must know," [online] Available at: <HTTP://blog.statutorily/social-media-statistics-2018-for-business/> [Accessed 02 Mar 2019]
2. Damian Radcliffe, Amanda Lam, "Social Media in the Middle East," [online] Available: https://www.researchgate.net/publication/323185146_Social_Media_in_the_Middle_East_The_Story_of_2017 [Accessed 06 Feb 2019].
3. GM_BLOGGER, "Saudi Arabia Social Media Statistics," G MI_blogger. [online] Available at: <HTTP://WWW.Globalmediainsightful/blog/Saudi-Arabia-social-statistics/> [Accessed 04 May 2019].
4. Smith, "49 Incredible Instagram Statistics," Brand watch. [online] Available at: <HTTP://www.brandwatch.com/blog/Instagram-stats/> [Accessed 10 May 2019].
5. Selling Stock (2014). Selling Stock. [online] Available at: <HTTP://WWW.selling-stockroom/Article/18-billion-images-uploaded-to-the-web-every-day> [Accessed 12 Feb 2019].
6. Li, W., Prada, S., Fowler, J.E., & Bruce, L.M. (2012). Locality-preserving dimensionality reduction and classification for hyper spectral image analysis. IEEE Transactions on Geo science and Remote Sensing, 50(4), 1185–1198.
7. A. Krizhevsky, I. Sutskever, & Hinton, (2012). Image net classification with deep Convolutional Neural Networks. In Advances in Neural Information Processing Systems, 1097–1105.
8. K. Ravi, (2018). Detecting fake images with Machine Learning. Harkuch Journal .
9. L. Zheng, Y. Yang, J. Zhang, Q. Cui, X. Zhang, Z. Li, et al. (2018). TI-CNN: Convolutional Neural Networks for Fake News Detection. United States.
10. R. Raturi, (2018). Machine Learning Implementation for Identifying Fake Accounts in Social Network. International Journal of Pure and Applied Mathematics, 118(20), 4785–4797.
11. J. Bunk, J. Bappy, H. Mohammed, T. M. Nataraj, L. Flenner, A., Manjunath, B., et al. (2017). Detection

and localization of image forgeries using resampling features and deep learning. University of California, Department of Electrical and Computer Engineering ,USA.

12. S. Aphiwongsophon & P. Chongstitvatana, (2017). Detecting Fake News with Machine Learning Method. Chulalongkorn University, Department of Computer Engineering, Bangkok, Thailand.

13. M. Villan, A. Kuruvilla, K.J. Paul, & E.P. Elias, (2017). Fake image detection using Machine Learning . IRACST—International Journal of Computer Science and Information Technology & Security (IJCSITS).

14. S. Shalev-Shwartz, & S. Ben-David, (2014). Understanding Machine Learning: From Theory to Algorithms. New York: Cambridge University Press.

15. D.-H. Kim, & H.-Y. Lee, (2017). Image manipulation detection using Convolutional Neural Network. International Journal of Applied Engineering Research, 12(21), 11640-11646.