iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.51443

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

Document Analyzing Using Deep Learning

Prajwal Akare[1], Rupal Wyawahare[2], Swaraj Bawankule[3], Sankalp Lanjewar[4], Arpit Nandanwar[5], Mrs. Surbhi Khare[6]
[1, 2, 3, 4, 5]*B.E -information technology, Priyadarshini College of Engineering, Maharashtra, India*
[6]*Assistant Professor, Department Of Information Technology, Priyadarshini College Of Engineering, Maharashtra, India.*

Abstract: ***Many businesses and large organizations have a large set of documents that need to be stored in various locations. cluster. In recent years, this task has become time consuming as the number of documents and articles has increased. Analysis of Documents are her one of the subjective research techniques analysts use to validate ideas. with some help. A visual technique that takes full advantage of layout and text formatting in a very clean output format. with the help of the model Most large and lots of architectural documents can be sorted. With the help of layout models and new interaction strategies Various layouts of any format within a single.***
Keywords: CNN, Deep Learning, Document Analyzer, Pre-Processing.

I. INTRODUCTION

Nowadays, world where there is an enormous amount of text data, digitization of documents is a technology used in different and so many and different types of fields. A domain with a large archive. Document Analyzer focuses on classifying documents based on their text. Document images and layout. Documents can usually be classified differently in many contexts. when we try In the task of analyzing text documents, document classification is an important procedure that must be followed. However, while recording Classification must address several and various challenges, including: B. High variability and low variability within the same document or class Between different classes or documents. Previous studies have shown structural similarity between classes and document.

II.DOCUMENT TYPE

All organizations, including universities, schools, corporations, etc., have data in various forms. Documents such as rating reports and tc cast certificates are evaluated using deep learning systems. CNN

Fig: types of documents

III.PURPOSE/OBJECTIVE

Identification and classification of Target documents is the purpose of this investigation. A form of qualitative research, known as document analysis, in which an analyst reviews documents to evaluate the subject of assessment.

A focus group or interview transcript, coding content into categories is a document analysis process.Prepare Your Paper Before Styling

IV. QUALITATIVE RESEARCH

Documentary analysis is a type of qualitative research in which documents are reviewed by the analyst assess an appraisal theme. Dissecting documents involves coding content into subjects like how focus group or interview transcripts are investigated. A rubric can likewise be utilized to review or score a document.

Similar to other methods of analysis in qualitative research, document analysis requires repeated review examination, and interpretation of the data in order to gain meaning and empirical knowledge oftheconstruct being studied.

©IJRASET: All Rights are Reserved

V. LITERATURE REVIEW

1) Analysis and Perceptions, ICDAR 2019 Analysis and Perceptions, ICDAR 2019, Sydney, Australia, 20-25. September 2019; pp. 726–731.

In 2019 edition of ICDAR, the International Conference on Document Analysis and Recognition ICDAR, which began in 1991 at St. Malo in France, is celebrating its 28th anniversary at this exciting conference, which was organized by me and Prof. Guy Lorette. ICDAR is now among the most significant international conferences in the pattern field. both artificial intelligence and recognition. The primary topics covered are document analysis and recognition, handwriting analysis and verification, text detection and processing, as well as other related subjects.

2) Papyri for author identification tasks. Mohammed, H. Marthot-Santaniello, I.; Margner,

It is crucial to showcase real research problems from academics through publishing datasets in order to come up with useful answers. Hence, for the purpose of writer identification, we suggest a dataset of handwriting on papyri. This datasets is based on research problems in the field of papyrology, and the samples were chosen by specialists in that area of study. This collection includes 50 Greek handwriting examples on papyri that date to around the sixth century A.D., representing the work of 10 distinct scribes. Together with their verified groundtruth data pertaining to the duty of writer identification, it is compiled and made freely available for non-commercial research.

3) Evaluation of Deep Convolutional Nets for Document Image Classification and Retrieval Adam W. Harley, A. U.

The features used in this research papper were learned using deep convolutional neural networks and represent a new state-of-the-art for document picture classification and retrieval (CNNs). Deep neural networks are capable of learning a hierarchical chain of abstraction from pixel inputs to succinct and descriptive representations in object and scene analysis. In the context of document analysis, the current work investigates this capability and finds that this representation method outperforms a number of common hand-crafted alternatives. Additionally, experiments demonstrate that I CNN features are resilient to compression, (ii) CNNs trained on non-document images perform well on tasks requiring document analysis, and (iii) with enough training data, it is not necessary to enforce region-specific feature learning. Also, a new tagged subset of the IIT-CDIP collection with 400,000 documents is made available through this study.

VI. ADVANTAGES

1) To analyze and classify the documents using CNN .
2) To extract features of the documents using algorithms.
3) Create a working model that classify the document on the basics of feature that are extracted.
4) The model will use image segmentation and CNN to determine the articles.

VII. DATASETS AND TRAINING

1) To train and evaluate the document classifier, we collected students' documents (the document given below). from Diverse backgrounds and departments. Each document class is used to classify the correct content in a document image when using the OCR method. We collect documents from students of various states. As these documents vary from state to state, different states in India have different document layouts. For the document parser, we have used different classes for document classification such as SSC Marksheet, HSC Mark-sheet, Cast Validity and Income Certificate.

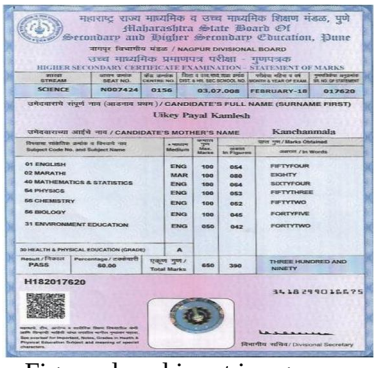

महाराष्ट्र राज्य माध्यमिक व उच्च माध्यमिक शिक्षण मंडळ, पुणे
Maharashtra State Board Of Secondary and Higher Secondary Education, Pune
नागपूर विभागीय मंडळ / NAGPUR DIVISIONAL BOARD
उच्च माध्यमिक प्रमाणपत्र परीक्षा – गुणपत्रक
HIGHER SECONDARY CERTIFICATE EXAMINATION – STATEMENT OF MARKS
SCIENCE | N007424 | 0156 | 03.07.008 | FEBRUARY-18 | 017620
CANDIDATE'S FULL NAME (SURNAME FIRST): Ukey Payal Kamlesh
CANDIDATE'S MOTHER'S NAME: Kanchanmala
01 ENGLISH ENG 100 054 FIFTYFOUR
02 MARATHI MAR 100 080 EIGHTY
40 MATHEMATICS & STATISTICS ENG 100 064 SIXTYFOUR
54 PHYSICS ENG 100 053 FIFTYTHREE
56 CHEMISTRY ENG 100 052 FIFTYTWO
56 BIOLOGY ENG 100 045 FORTYFIVE
31 ENVIRONMENT EDUCATION ENG 050 042 FORTYTWO
30 HEALTH & PHYSICAL EDUCATION (GRADE) A
Result / निकाल: PASS | Percentage / टक्केवारी: 60.00 | एकूण गुण / Total Marks: 650 390 | THREE HUNDRED AND NINETY
H182017620
Divisional Secretary

Fig: analyzed input image

2) Model Builder uses an automated process called training to train a model to respond to contextual requests. Once trained, the model can make predictions using completely new inputs. For example, you can estimate the price of a home and predict the sale price if a new home is on the market. Model Builder uses automatic machine learning (AutoML), so it requires no input or configuration during training.
3) *"How long should I train?"* To determine which model performs best, Model Builder uses AutoML to analyze different models.

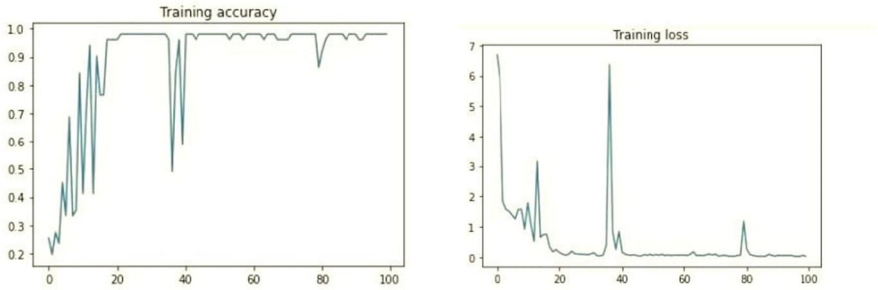


Fig: Model traing accuracy

VIII. METHODOLOGY AND FLOW DIAGRAM

The "Document Analyzer" proposed approach in this research tries to identify the class of a document by analyzing the content of the scrimmage. We used three approaches to address this challenge: layout identification, text classification, and image classification. In the pre training stage, we suggest using a multi-modal Transformer model to join the document's text, layout definition, and visual data. This model learns cross-modal interactions in a single framework.

The class of a document has been determined based on a number of features, including the document's layout, header and footer, body, or the content of the document, which is extracted using OCR techniques, and how the document is formatted. All of these features assist in determining the precise class of the given document.

Text embedding, visual embedding, and layout embedding are the three sections of the Layout LM architecture. Tokenizing the OCR, the text's order, and providing some segments are common tasks in this type of embedding. One approach used in natural language processing is the sub world.

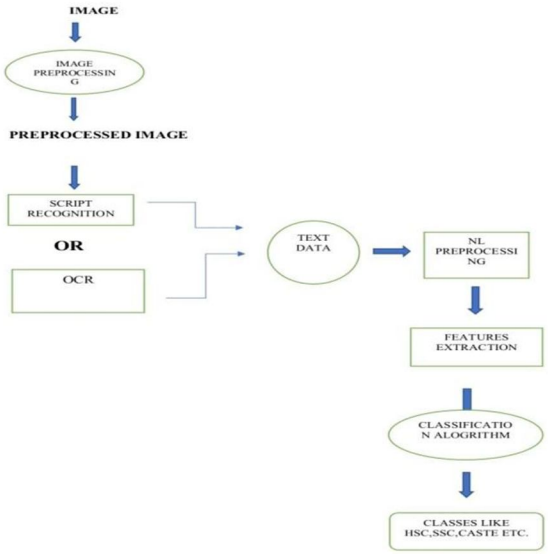


Fig: Flow chart diagram

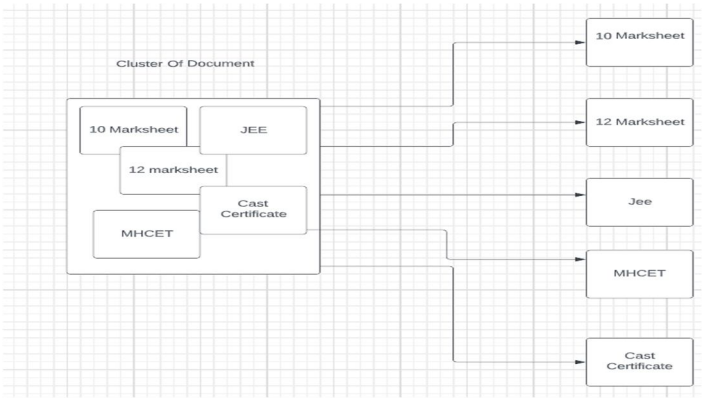


Fig: classification of image

The process classifies the document according to their categories and classes, as we already discuss in datasets and training section classes can be like mark sheets, caste certificate, tc etc. As shown second image above cluster of images are present and it has classified into classes.

The sentences are used to examine the characters in the aforementioned input and to identify the most frequent pairings of the different characters in the sentence.

Image Processing - This topic covers the fundamentals of image pre-processing. This aims to perform elementary picture pre-processing such as image scaling, resizing, and compression, as well as morphological image pre-processing such as erosion and dilation.

Pre-processed photos with a similar look will be this module's output.

OCR, often known as optical character recognition, is a method for extracting text from images. This module's objective is to extract image.

The design of the document, the headers and footers, the document's text, and the style of writing all contribute to the identification process and serve as criteria for determining a document type. Document kind A government certificate with a seal and/or a logo to assist classify the document is an example of a form of document that shares common characteristics with other sorts of papers.

A. Working Of Machine Learning(Ml) Using Python

A cross-platform, open-source machine learning framework called ML with Python enables machine learning for Python developers. Machine learning may be incorporated into Python programmers both online and off using ML-python. With the data your application has access to, you can create predictions automatically thanks to this functionality.

The Convolution Neural Network (CNN) algorithm used in its hidden layers

A CNN is a specific type of network design for deep learning algorithms used for tasks such as image recognition and pixel data processing. CNN plays very important in these process CNN are trained to perform classification tasks, but CNN trained for classification can also be used to perform search. Although these feature vectors are multidimensional, their dimensionality can be greatly reduced by principal components analysis without significantly affecting discriminating power. Ranking these training data images returns a sorted list of documents.

Applications for machine learning employ data patterns to produce predictions without the necessity of explicit programming. Machine learning models are the heart of ML-python. The procedures needed to convert input data into predictions are specified by a model. By describing an algorithm, machine learning using Python enables you to train a specific model. You can include the model into your application to make predictions once you have it.

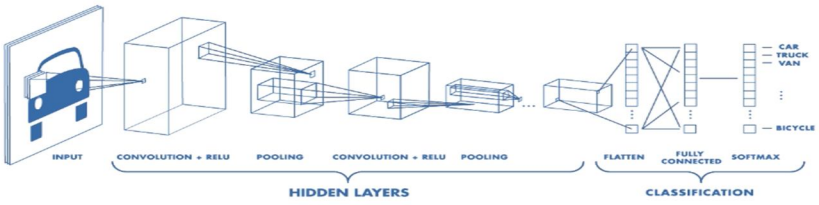


Fig: processing analyzing using algorithm CNN

IX. FUTURE SCOPE

People are moving to digital documents for authentication, but most areas are used, such as land registries, contracts between parties, legal certificates, and identification cards. Document verification is important because counterfeit documents affect the true owner so much. Therefore, awareness of authentic documents is necessary to avoid these scams. This paper uses deep learning for document recognition because it provides the highest accuracy and requires no preprocessing. Deep learning models based on convolutional neural networks (CNNs) are primarily used for image processing, classification, and segmentation. Since CNN algorithms learn more than KNN, SVM, etc., we use CNN in this work for better classification. CNN-based models such as VGG16, Inception v3, and CNNs with 3 and 4 convolutional layers have been trained for this classification. The data set is created by collecting documents from 10 different users. Among these four models, Inception V3 showed the highest accuracy of 95% with preprocessed images, while the same model only achieved 88% with raw images as input.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

X. RESULTS AND CONCLUSION

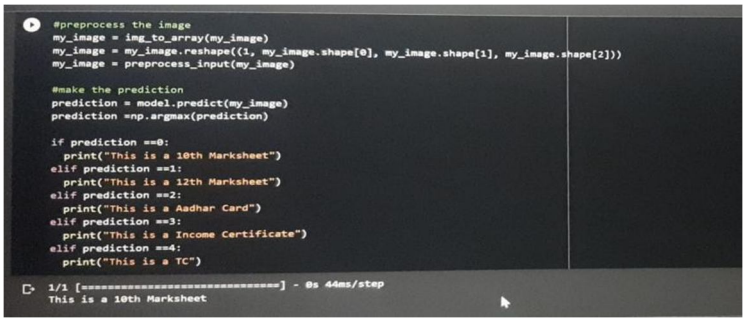


Fig: result of the document analyzing

1) The above image is showing the result of the model that has been train using deep learning algorithm and as we get accurate prediction of the document the document has analyzed successfully.
2) Our motive solution is a model that will correctly categorise and classify documents and articles. The model was created using CNN and image feature extraction, and results were improved even further by fine-tuning these features that were taken from document pictures.
3) The CNN method of representing document images is more effective than hand-made alternatives.

A. *Gui Output Of Document Analyser*

1) GUI is an interface that allows users to interact with different electronic devices using icons and other visual indicators.
2) The graphical user inteface were created beacause command line interfaces were quite complicated and it was difficult to learn all thew commands in it.
3) In Our Gui we have given proper page for login and home page where after login home page will open.
4) Home page consists of Home, Teams/About us, Contact and logout.
5) Here in home page we provided a button for uploading file or image to show output.
6) After uploading image or file we'll get outoput that which document it is and Its image.
7) For every task or project Gui is most important thing for user, as the user nothing know about backend, Frontend helps the user to get the work done according their instructions.
8) Below images shows our frontend outputs:

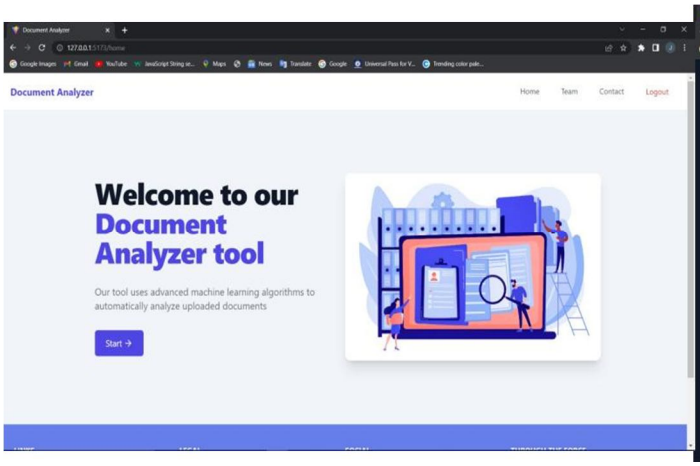


Fig: Document analyser tool home page

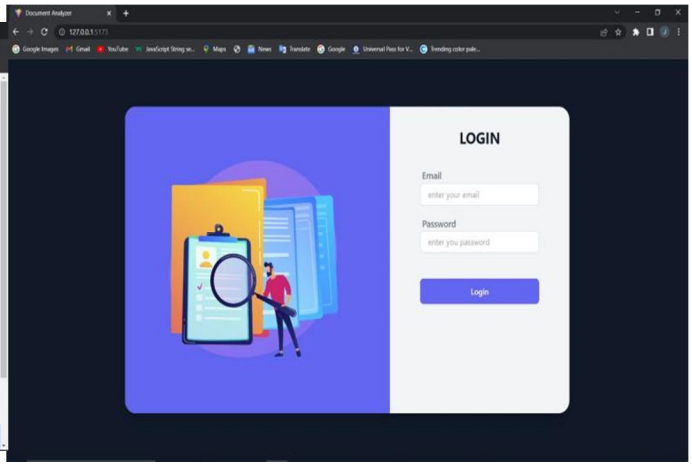


Fig: document analyser login page

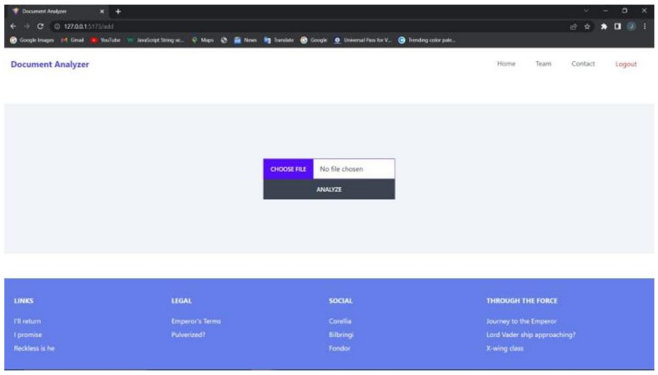


Fig: document anaylyser home page

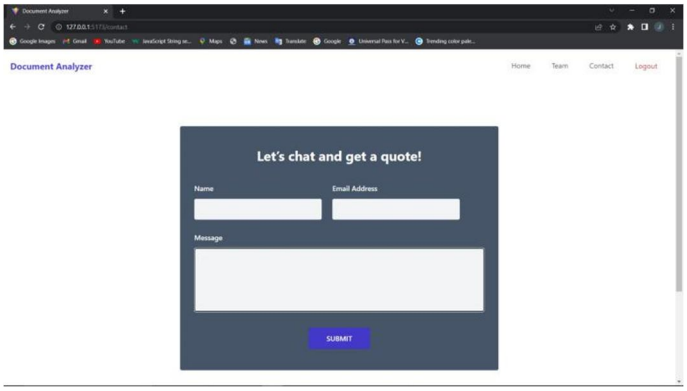


Fig: contact/ feedback page

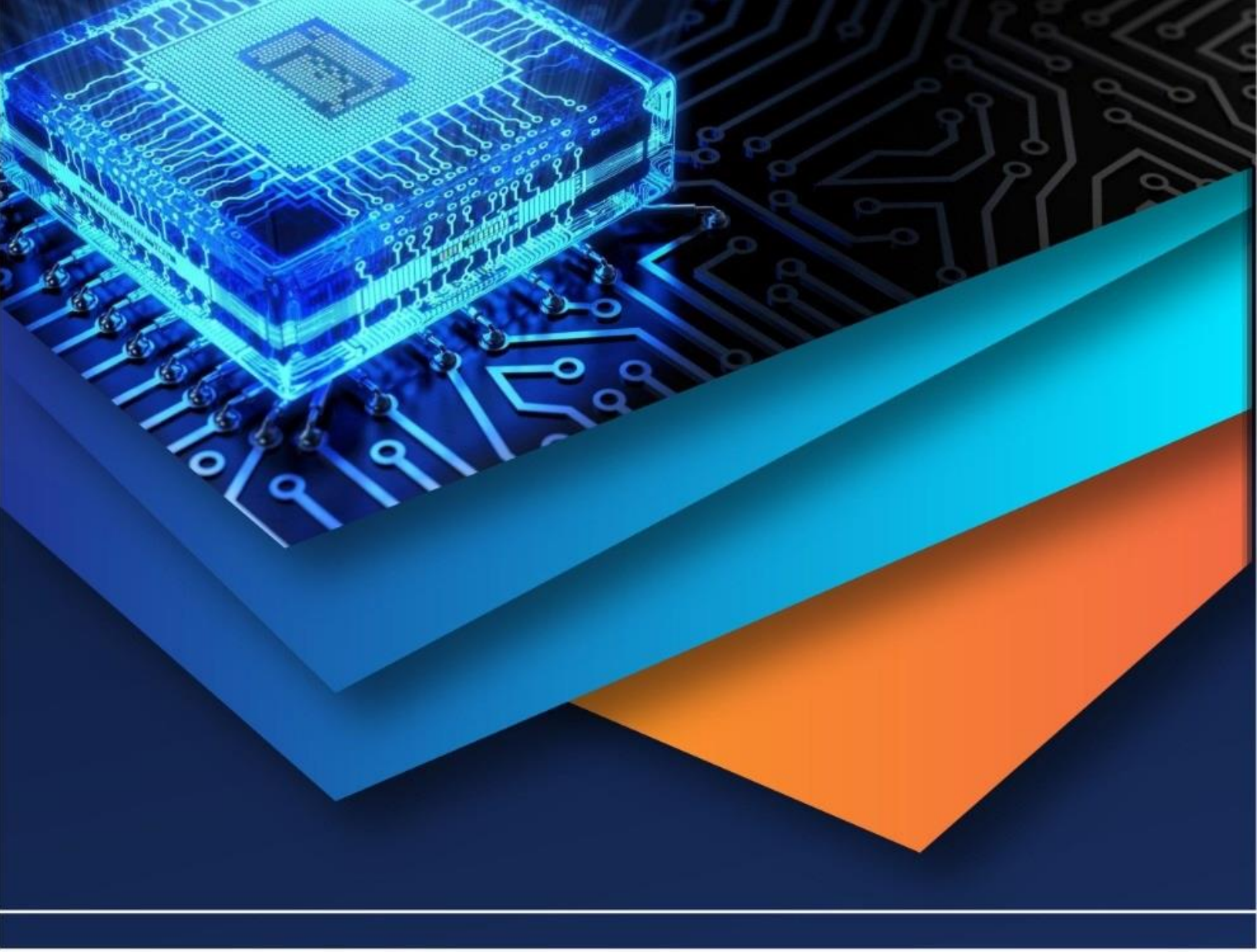
Fig: document analysing final output

XI. ACKNOWLEDGEMENT

We would like to thank Mrs. Surbhi Khare, our Professor-in-charge and our , HOD Mrs. Pallavi Choudhari for their support and guidance in completing our project on the topic Document Analysing Using Deep Learning. It was a great learning experience.
I would like to take this opportunity to express my gratitude to all of my group members . The project would not have been successful without their cooperation and inputs.

REFERENCES

[1] Emerson, S., Kennedy, R., O'Shea, L., & O'Brien, J. (2019, May). Trends and Applications of Machine Learning in Quantitative Finance. In 8th International Conference on Economics and Finance Research (ICEFR 2019).
[2] Siami-Namini, S., & Namin, A. S. (2018). Forecasting economics and financial time series: Arima vs. lstm. arXiv preprint arXiv:1803.06386.
[3] Heaton, J. B., Polson, N. G., & Witte, J. H. (2017). Deep learning for finance: deep portfolios. Applied Stochastic Models in Business and Industry, 33(1), 3-12.
[4] Moritz, B., & Zimmermann, T. (2016). Tree-based conditional portfolio sorts: The relation between past and future stock returns. Available at SSRN 2740751.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)