Deep Text Classification and Retrieval for Text Documents using Deep Learning Techniques

Thesis Submitted
In Partial Fulfilment of the Requirements for the Degree of

MASTER OF TECHNOLOGY

in

Software Engineering

by

Vaibhav Kansal
(2K22/SWE/21)

Under the Supervision of
Dr. Abhilasha Sharma
(Assistant Professor, SE, DTU)

To the
Department of Software Engineering
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042, India

May, 2024

ACKNOWLEDGEMENTS

I would like to express my deep appreciation to **Dr. Abhilasha Sharma**, Assistant Professor at the Department of Software Engineering, Delhi Technological University, for her invaluable guidance and unwavering encouragement throughout this research. Her vast knowledge, motivation, expertise, and insightful feedback have been instrumental in every aspect of preparing this research plan.

I am also grateful to **Prof. Ruchika Malhotra**, Head of the Department, for her valuable insights, suggestions, and meticulous evaluation of my research work. Her expertise and scholarly guidance have significantly enhanced the quality of this thesis.

My heartfelt thanks go out to the esteemed faculty members of the Department of Software Engineering at Delhi Technological University. I extend my gratitude to my colleagues and friends for their unwavering support and encouragement during this challenging journey. I have had some friends that I am thankful to be around. They made me feel truly at home. In particular, I would like to thank Aparna Arya whom I had such a great time. Last two years were great with such a lovely bunch of people around me. Their intellectual exchanges, constructive critiques, and camaraderie have enriched my research experience and made it truly fulfilling.

While it is impossible to name everyone individually, I want to acknowledge the collective efforts and contributions of all those who have been part of this journey. Their constant love, encouragement, and support have been indispensable in completing this MTech thesis.

Vaibhav Kansal
(2K22/SWE/21)

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CANDIDATE DECLARATION

I VAIBHAV KANSAL (2K22/SWE/21) hereby certify that the work which is being presented in the thesis entitled "**Deep Text Classification and Retrieval for Text Documents using Deep Learning Techniques"** in partial fulfillment of the requirements for the award of the Degree of Master of Technology submitted in the Department of Software Engineering, Delhi Technological University in an authentic record of my work carried out during the period from August 2022 to May 2024 under the supervision of Dr. Abhilasha Sharma.

The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other Institute.

Vaibhav

Vaibhav Kansal

This is to certify that the student has incorporated all the corrections suggested by the examiner in the thesis and the statement made by the candidate is correct to the best of our knowledge.

Abhilasha Sharma

Signature of Supervisor(s)

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CERTIFICATE BY THE SUPERVISOR

Certified that Vaibhav Kansal (2K22/SWE/21) has carried out their project work presented in this thesis entitled "**Deep Text Classification and Retrieval for Text Documents using Deep Learning Techniques**" for the award of **Master of Technology** from the Department of Software Engineering, Delhi Technological University, Delhi under my supervision. The thesis embodies results of original work, and studies are carried out by the student himself and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

Abhilasha Sharma

Date: 27/05/2024

Dr. Abhilasha Sharma
Assistant Professor
Department of Software Engineering,
DTU-Delhi, India

Deep Text Classification and Retrieval for Text Documents using Deep Learning Techniques

Vaibhav Kansal

ABSTRACT

Document Retrieval (DR) is pivotal in unlocking valuable insights from the ever-growing volume of medical literature. However, precise knowledge extraction from complicated clinical notes, discharge summaries, and research papers is still difficult. However, capturing these distinct dimensions of medical discourse and the complex interdependencies between entities with the currently used approaches becomes difficult. The classifications of Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and Convolutional Recurrent Neural Networks (CRNNs) for use in document retrieval are thoroughly discussed in this review paper. The study offers a comprehensive assessment of the individual models' performance by summarizing the conclusions from over forty studies that relate to the individual models' efficacy. It explores the potential outcomes of the group strategies in enhancing document retrieval accuracy. The yield of this research crystalizes the broad canvas encompassing the diverse features of CNNs, RNNs, and the ensemble technique CRNNs which are used to detect the complexities present in healthcare documents thereby presenting an analysis of their appropriateness included in various retrieval tasks and document types. The findings have imperative guidance for the researchers and practitioners looking to improve the retrieval system that documents have in healthcare to improve healthcare professionals' decision-making processes and promote access to vital medical information.

This work solves these shortcomings by introducing a convolutional recurrent neural network (CRNN) framework that merges the advantages of CNNs and RNNs to achieve very high accuracy in medical document extraction tasks. In the CRNN model, the spatial feature extraction capacity of CNNs is integrated with the sequential learning capabilities of RNNs to successfully find medical entities and understand their complex interrelationships, which leads to high-performance document classification. As shown in the case of CORD-19, our CRNN model significantly outperforms individual CNN and RNN models on its entity recognition and relationship extraction tasks (98.93% accuracy), demonstrating that using CRNN in medical documents can yield much better results, thus increasing opportunities for informed clinical decisions, advanced drug discovery, and improved public health interventions**.**

Keywords: Document retrieval, CNN, RNN, CRNN, CORD-19.

TABLE OF CONTENT

Title	Page No.
Acknowledgment	*ii*
Candidate's Declaration	*iii*
Certificate	*iv*
Abstract	*v*
Table of Contents	*vi*
List of Table(s)	*viii*
List of Figure(s)	*ix*
List of Abbreviation(s)	*x*
CHAPTER 1: INTRODUCTION	1-3
1.1 Overview	1
1.2 Motivation	2
1.3 Problem Statement	2
1.4 Precise Working	2
1.5 Thesis Outline	3
CHAPTER 2: LITERATURE SURVEY	4-17
CHAPTER 3: METHODOLOGY	18-29
3.1 Document Retrieval	18
3.2 Dataset	19
3.3 Software Requirements	20
3.3.1 Python Programming	20
3.3.2 Anaconda Distribution and Anaconda Navigator	20
3.3.3 Jupyter Notebook	21
3.4 Necessary libraries	22
3.4.1 NumPy	22
3.4.2 Pandas	22
3.4.3 Matplotlib	22
3.5 Data Cleaning	22
3.6 Pre-Processing Stage	23
3.6.1 Document Parsing	23
3.6.2 Lexical Analysis	23

3.6.3 Stemming and Lemmatization 24

3.6.4 Word Embedding 24

3.7 Dimensionality Reduction 24

3.7.1 Feature Selection 25

3.7.2 Feature Extraction 26

3.8 Doc2vec 27

3.9 Clustering Stage 27

3.9.1 K-Means Clustering 28

3.9.2 How the K-Means algorithm works 28

CHAPTER 4: DEEP LEARNING AND ENSEMBLE MODEL 30-34

4.1 Convolutional Neural Network 31

4.2 Recurrent Neural Network 32

4.3 Convolutional Recurrent Neural Network 32

4.4 Natural Language Processing 33

4.5 Long-Short-Term Memory 34

4.6 Gated Recurrent Unit 34

CHAPTER 5: DATASETS 35-36

5.1 Similar Datasets 35

CHAPTER 6: RESULTS AND DISCUSSION 37-38

CHAPTER 7: CONCLUSION, FUTURE SCOPE AND SOCIAL IMPACT 39-40

REFERENCES 41-45

LIST OF PUBLICATIONS 46-47

LIST OF TABLE(S)

2.1: Summarized review of literature papers 4

6.1: Performance of Neural models 37

LIST OF FIGURE(S)

1.1	Document Retrieval Process	1
3.1	Document Retrieval System	19
3.2	CORD-19 dataset	19
3.3	Anaconda distribution	21
3.4	Jupyter notebook	21
3.5	Embedding model	24
3.6	Dimensionality reduction technique	25
3.7	PCA	26
3.8	Learning model of DOC2VEC	27
3.9	K-Means clustering	29
4.1	Different types of deep learning models	30
4.2	Working of CNN	31
4.3	Recurrent Neural Network	32
4.4	Natural Language Processing	33
4.5	Gated Recurrent Unit	34
6.1	Comparison of Evaluation Measure	38
6.2	Accuracy of different models	38

LIST OF ABBREVIATION(S)

DR	Document Retrieval
NLP	Natural Language Processing
RNN	Recurrent Neural Network
PCA	Principal Component Analysis
CNN	Convolutional Neural Network
GUI	Graphical User Interface
AI	Artificial Intelligence
ML	Machine Learning
DNN	Deep Neural Network
OCR	Optical Character Recognition
NER	Named Entity Recognition
BOW	Bag of Words
TF-IDF	Term Frequency-Inverse Document Frequency
LDA	Linear Discriminant Analysis
GDA	Gaussian Discriminant Analysis
DBOW	Distributed Bag of Words
DM	Distributed Memory
DL	Deep Learning
CRNN	Convolutional Recurrent Neural Network
GPU	Graphics Processing Unit
MLP	Multilayer Perceptron
Re-LU	Rectified Linear Unit
GRU	Gated Recurrent Unit
LSTM	Long Short-Term Memory

CHAPTER 1

INTRODUCTION

This work focuses on developing a deep reinforcement learning model for medical document retrieval (DR). This study also demonstrated that, in practice, integrating numerous learning models should be multifaceted, and that flawless integration of a range of models can be therapeutically beneficial. Improved document searching and retrieval improves knowledge and allows for more tailored approaches to therapy, proper diagnosis, and hence better patient care and prognosis.

1.1 OVERVIEW

Document retrieval (DR) is one of the profoundly important ideas in medical care that can be utilized for empowering admittance to data from huge texts [1]. It concerns the technologies that can be utilized for looking and extracting the most important and helpful data from the unstructured information with respect to clinical accounts cases, and so on. During the COVID-19 pandemic, it has become crucial to efficiently extract necessary medical data from an ever-growing CORD-19 dataset [2]. This step is important for guiding clinical decisions, ensuring successful drug discovery, and serving as the underlying foundation of public health interventions. The typical manual technique of processing complex medical documents may become problematic, but utilization of NLP techniques seems to be a promising solution in this respect by automating the above procedure.

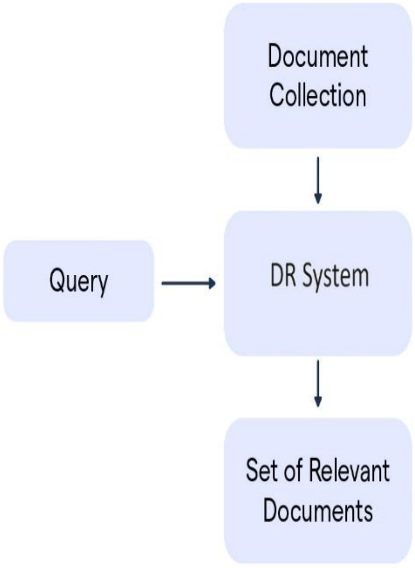


Fig.1.1 Document Retrieval Process

Consequently, the advancements in neural network models, especially the ones known as CNNs [3] and RNNs [4], have enabled them to extract information from reports as well as provide explanations. When it comes to assessing the relationship between words in a text, CNN is the favored method. On the other hand, RNN is better at processing sequential data. This, therefore, makes it suitable for the identification of names and terms used in interpreting relationships in medical texts. When using these neural network models in combination with NLP, the model can meet this research-based challenge of complexity and quantity inherent in medical documents. Research in this area has already shown that traditional CNNs and RNNs can extract information from medical documents.

1.2 MOTIVATION

To effectively manage digital archives, new technologies such as deep learning must be included for efficient document retrieval and data location tracking. Because traditional retrieval algorithms prioritize outcomes over text comprehension and contextual interpretation, they frequently provide incorrect outcomes. Furthermore, traditional approaches have various disadvantages, such as a lack of memory elements, hidden state, and so on, which causes them to provide inaccurate outcomes. This process is being altered by deep learning approaches, particularly when paired with natural language processing (NLP). Sentiment and other language complexity are all part of the NLP framework, which employs RNN, CNN, and other deep learning algorithms to detect complicated patterns. Thus, the data becomes more exact and useful.

Moreover, deep learning facilitates the immediate extraction of important features from the data, doing away with the tedious human feature-creation process and opening the door to more efficient and adaptable retrieval techniques. When access to relevant information is critical, deep learning-based document retrieval is a powerful tool that helps users quickly and reliably sort through massive volumes of data and extract meaningful information.

1.3 PROBLEM STATEMENT

The project's goal is to address knowledge gaps by developing effective clustering, classification, and deep learning models, which will improve document retrieval based on user demands. We must decide which model will be provided to our search engine so that it can provide the most relevant document in the future. It is the most significant issue we must address in our project. This increases the retrieval of the CORD-19 dataset [2] as well as the categorization of relevant scholarly publications. Although large datasets are employed, current models have not completely leveraged their possibilities to ensure improved information retrieval. The objective is to bridge this gap by developing new algorithms for precise document categorization, which will enable rapid retrieval of important data.

1.4 PRECISE WORKING

To categorize a text, the following steps will be performed:

- A deep neural network model and several classifications will be used to train a dataset of around three lakh query samples.
  1. First, PCA is used for data cleaning and dimensionality reduction.
  2. Next, the Doc2Vec method is used to extract the document.
  3. After this data is clustered using the k-means algorithm.

4. Lastly, we will feed the neural network model our dataset and test it by getting pertinent documents.

- The measurement of accuracy will be done with several neural models.

1.5 THESIS OUTLINE

To find out what has already been done in this specific domain which is covered in the upcoming chapter we have carried out a literature survey. Next, we described our project's workflow and methodology in Chapter 3. In Chapter 4, we have explored many deep learning algorithms and ensemble models to determine which model yields the best outcomes when applied to our dataset. The same task can be done on several datasets that have been mentioned in Chapter 5. Chapter 6 presents a series of calculations and comparisons that were made to determine the optimal accuracy. Ultimately, our project comes to an end with the scope and related future work that could be completed for further enhancements.

CHAPTER 2

LITERATURE SURVEY

This section demonstrates a thorough literature review conducted earlier on numerous classification and retrieval techniques for different systems that are now in use.

Table 2.1: Summarized review of literature papers

No.	Ref	Methodology	Conclusion	Dataset Used
1.	[5]	The study looks into text classification using a hybrid CNN-LSTM model with TF-IDF. To represent documents, they first preprocess text input and calculate TF-IDF scores. LSTM network captures sequential relationship. When linked to old-style approaches, the model performs better in text classification tasks after integrating these variables for classification.	The authors provide a CNN-LSTM and TF-IDF-based text categorization model. While CNN concentrate on gathering local information, TF-IDF may specifically extract the most significant textual aspects, while LSTM can gather general information. By combining these three, they ran trials on datasets of short and long texts, respectively.	THUCNews, Taobao Review
2.	[6]	By presenting an exclusive Text-CNN classifier with a freshly constructed CE-MSERs detector, the authors of this study have produced a new scene text detection method. It makes use of binary text/non-text data, character class, text region mask, and other highly-supervised text information. They frame the Text-CNN training problem as a	The study concludes that scene text detection problems are well handled by the Text-Attentional Convolutional Neural Network (TACNN). TACNN performs better than conventional techniques in accurately detecting text in complicated settings by integrating both textual and visual information through attention mechanisms. TACNN's	ICDAR 2011

		multi-task learning problem that successfully combines multi-level supervisory interactions. They demonstrate that building a strong Text-CNN that can reliably distinguish ambiguous text from complex backdrop depends heavily on informative multi-level supervision.	potential for real-world scene text identification applications is showcase by experimental findings that indicate its superiority in terms of detection accuracy and robustness against fluctuations in text appearance and backdrop clutter.	
3.	[7]	An enhanced Bi-LSTM-CNN technique for news text classification is presented in this research. They first preprocess news texts and display them as token sequences. Then, they add residual connections and attention techniques to improve the conventional Bi-LSTM-CNN model.	In this research, the Bi-LSTM-CNN model uses a convolutional neural network to construct the left and right settings of each word, resulting in a literary articulation that more accurately conveys the meaning of the text. It accomplishes this by gathering the setting data using the circular design.	THUCNews
4.	[8]	The article presents the CNN-BiLSTM model for sentiment analysis at the document level. They represent documents as token sequences after preprocessing them. Lastly, the model outperforms convention techniques by combining these features for sentiment categorization at the document level.	The creators of this exploration introduce the joining of convolutional and bidirectional recurrent neural networks for report-document-level sentiment analysis with Doc2vec Embedding. Since it exploits both the CNN element extraction abilities and the BiLSTM's ability to get familiar with the drawn-out bi-directional conditions of the text, the consolidated CNN-BiLSTM model performs above and beyond huge texts.	French Articles
5.	[9]	In this research, a unique neural network-	In this publication, researchers present a	TCM, Hallmarks

		based approach to medical text classification is presented. They use a neural network architecture created especially for this task and preprocess medical texts. The efficiency of the suggested strategy in medical text classification tasks is demonstrated through experimental evaluation.	hierarchical neural network approach for medicinal manuscript classification. By segmenting the document and then aggregating those segments into the document representation, the approach creates sentence representations of sentences. It incorporates the attention mechanism, BIGRU, and the word-level convolutional layer.	
6.	[10]	Using a deep dense LSTM-CNN framework, the research provides an extractive text method for biological transcripts. Next, they employ a blend of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) to acquire both local and sequential data. They utilize a dense layer at the end to provide summary sentences	The study concludes that extractive text summarization for biomedical transcripts can be efficiently facilitated by the deep dense LSTM-CNN framework. Key information from the transcripts is reliably identified by the model, which makes use of both sequential and resident features. The suggested approach is superior to existing methods, as demonstrated by the experimental findings.	MT Data Samples
7.	[11]	The method described in this study for document-level text categorization is a convolutional neural network (CNN) with a single layer and several filter sizes. These filters are tools for extracting features. The model operates on document-level text classification tasks across several datasets, learning layers	The goal of this study is to draw attention to Urdu, a language with limited resources, by creating and making available a sizable, intricate, and versatile text collection. The SMFCNN performs well in classifying small, medium, and large-size datasets on the Urdu TC problem. Although determining	COUNTER, NPUU and naïve Documents

		of characteristics and classifying texts based on these qualities, which produces excellent performance.	the SMFCNN's optimized parameters takes a lot of time and resources, the classifier performs better as a result.	
8.	[12]	This research offers a novel technique for text classification that combines CNNs with topic-based word embeddings. Using topic information, they first create word embeddings, and then they use CNNs to identify local patterns in the text. Experiments described in the study show that this combination strategy delivers enhanced performance in text classification tasks and effectively learns representations of texts.	With word embeddings developed by Skip-gram, an amazing approach for learning subject-based semantic word embeddings for message categorization with CNNs, which was introduced in this release, scientists achieved incredibly ruthless results. While Skip-gram concentrates on setting data from neighboring word windows, the suggested Topic-based Skip-gram incorporates semantic information from records as well.	MEDLINE Citations Public Dataset
9.	[13]	This work most likely defined goals gathered relevant data, pre-processed it, and extracted features with rich metadata. They trained models, chose deep learning architectures, and assessed performance in comparison to baselines. Most likely, analysis ensued, evaluating the effect of deep learning on semantically rich representations for document classification.	The INFUSE dataset in the funding sector is used to demonstrate an ontology-based document categorization system in this paper. It integrates pertinent terminology into semantically rich text representations that are obtained from ontologies. Deep learning classifiers are used to accurately assign categories to documents by using semantic representations as input.	INFUSE
10.	[14]	In this paper, the authors suggested a deep neural network-based methodology for	They conclude that their deep neural network-based method successfully brings	Medical Diagnosis Dataset

		online medical diagnosis data extraction and named entity recognition. They most likely trained the deep neural network model, assessed its effectiveness, and pre-processed the data. Their strategy probably entailed using the network's capacity to identify complex patterns in the data to recognize and extract named entities accurately.	named items from online medical diagnosis data and recognizes them. The approach performs well and has potential uses in enhancing the precision and efficiency of medical data analysis and decision-making procedures.	
11.	[15]	The authors suggested using deep neural networks to automatically summarise documents. They probably trained the neural network model, assessed the quality of the summaries, and pre-processed the papers. Deep neural network technology made it possible to extract intricate relationships from the text, which aided in the creation of clear and useful summaries. The efficacy of the strategy was probably evaluated by contrast with other available techniques.	This study introduces the DNN to find the sentence semantic space to extract sentences with meaningful semantics and limit information duplication. Both qualitative and quantitative analysis show that this deep framework can produce encouraging outcomes.	DUC2006, DUC2007
12.	[16]	The use of word embeddings developed on closed-domain corpora in the paper improves the extraction of biomedical information. First, they use Word2Vec to train word embeddings using	The authors of this publication provided instructions on how to create an unannotated text corpus for the Italian biomedical domain. They used it to produce several biomedical WE exhibits	PubMed and PMC Articles

		domain-specific datasets. They then adjust these embeddings to the domain by fine-tuning them on a specific biomedical corpus. Lastly, they employ the refined embeddings in information extraction tasks and show better results than with general-purpose embeddings.	and then integrated it as the information layer for a B-NER DL engineering. By comparing the displays of these WEs with the behaviors of WE demonstrations made on broad area corpora, it was therefore possible to illustrate the convenience of WEs prepared on the biomedical shut space corpus when employed as the contribution of the B-NER DL design..	
13.	[17]	This evaluation presents a deep mind network-based method for Named Element Acknowledgment (NER) in Chinese clinical texts. They employ a Bidirectional Long Short Term Memory (Bi-LSTM) network with a Restrictive Irregular Field (CRF) layer to capture name information. The model is ready on described clinical text data in order to recognize and organize named objects appropriately.	The authors of this paper looked into using a deep neural network for NER from Chinese clinical texts. A subsequent study revealed that the automatic semantic evidence gathered by the DNN-based word embeddings was the source of the performance gain, demonstrating the value of unsupervised feature learning.	Chinese Clinical Corpora
14.	[18]	This article, Cde-iiith's method for SemEval-2016 Task 12—focuses on applying machine learning to extract temporal information from clinical documents. They use elements from syntactic and semantic analysis to create Conditional Random Fields (CRFs). They also use other	In this publication, they present their work on the Clinical TemEval task from the SemEval 2016 challenge. To accomplish the challenge's obstacles, researchers employed two different strategies: the first is based on CRF and SVM, and the second makes use of deep neural networks.	Cancer Patient Record

		resources, such as medical ontologies, to increase the precision of temporal information extraction.	The outcomes demonstrate that both strategies perform about the same on the challenges offered to train and test datasets.	
15.	[19]	The technique for classifying web pages using recurrent neural networks (RNNs) is proposed in this paper. They use an RNN architecture to encrypt the sequences of tokens created by preprocessing web page content. A softmax layer receives the encoded representations to do classification. Based on experimental results, web pages can be accurately classified using the RNN-based technique	In this study, a deep learning-based system has been created to classify web pages. The system that was designed took advantage of the data that Roksit provided. A webpage is categorized using the metadata that is present on it. The title, description, and keywords meta tags are employed. The crawler module created for this study is responsible for gathering this data. The tests were conducted using a deep-learning architecture based on RNNs.	Roksit's web classification database
16.	[20]	This work presents a development of making a message summing up a model that coordinates the extractive and the abstract methodology. To upgrade the nature of the rundowns delivered, they utilize document context vectors on the records and outlines, and they depend on Recurrent Neural Networks (RNNs) in the creation of the synopses. Starting there, it targets furnishing the peruser with a compact and clear comprehension of the substance of the first message by	The proposed system of using both extractive and abstractive rundown strategies essentially works on the nature of message outline as is obviously clear from the end part of the paper. It gives an outline of the topics covered in the first papers and it plans succinct and valuable synopses. As displayed in the trial results, the proposed technique outflanks regular outline procedures, and the outcomes likewise uncover its conceivable extent of use in different NLP undertakings.	eBay description

		introducing its significant ideas and key data.		
17.	[21]	This article offers an approach to catching classification as a model that is prepared to utilize recurrent Neural Networks (RNNs) combined with Natural Language Processing (NLP). In the first place, with the assistance of NLP, they separate highlights from the articles, and accordingly, apply the RNN system to characterize the articles. This RNN model considers the grouping of similitudes of the texts and consequently more accurate classification of texts instead of expressions should be possible. The appropriateness of the suggested technique has been demonstrated through tests, which are consequently characterized by document jobs.	The performance analysis of this proposed strategy from the tests uncovers that their work is plausible and compelling on a few datasets and even yields better outcomes looked at than a portion of the current procedures. After making a few tests, the strategy is determined to have almost 97.5% of the typical exactness of the proposed procedure. The examination utilized three sorts of grouping calculations to test the framework; discoveries show that the document object characterization framework utilizes a recurrent neural network to manage the overt repetitiveness peculiarity during multi-mark order more productively than the other classification algorithms.	PDF Dataset
18.	[22]	This research adds to planning an attention-based Recurrent Neural Network (RNN) model for the Document Picture Quality Assessment. They first transform the record pictures into machine-processable structures through preprocessing of the archives. Then, at that point, to find noticeable locales in the picture, they utilize an	The authors of this work describe a DIQA technique based on attention-based RNN. This technique uses a convolutional layer to extract features and a spatial peek to choose a prominent area of the document image. The locator module is trained using a reinforcement learning technique, and other network components are	SOC Dataset

		RNN model that applies attention components. As laid out in the assessment cycle, the attention component empowers dynamic highlighting of huge picture regions that empower work on the model's presentation in precisely surveying the nature of archive pictures.	trained using stochastic gradient descent. After that, they create a brand-new reward system and confirm its efficacy.	
19.	[23]	The technique for extracting key phrases from Twitter via Deep Recurrent Neural Networks (RNNs) is presented in this paper. Tweets are pre-processed and represented as token sequences. Subsequently, a deep RNN architecture is utilized to accurately identify key phrases by capturing contextual information and semantic linkages inside the tweet messages. The experimental results show the efficiency of the suggested method for obtaining pertinent key phrases from Twitter data.	To complete the key-phrase extraction challenge, researchers in this paper suggested a novel deep recurrent neural network (RNN) model that combines keywords and context information. The task of generating key phrases and ranking keywords can be handled simultaneously by the suggested model. Its two hidden layers which categorise key phrases and differentiate between keywords combine to form the final objective function. They used a dataset filtered from ten million tweets that had been crawled to test the suggested approach.	Twitter Dataset
20.	[24]	In this paper, a deep sentence embedding technique based on Long Short-Term Memory (LSTM) networks is presented. To obtain semantic information, they first preprocess text data and use LSTM networks to encode phrases.	To describe the long-range context information and integrate the important information of a sentence in a single semantic vector, the authors presented a model based on long short-term memory. The authors go on to	BookCorpus Dataset

		Subsequently, they utilize a ranking loss function to optimize language embeddings for information retrieval tasks during model training. The results of the experimental study demonstrate how well the suggested method works to produce meaningful language embeddings, which enhances information retrieval applications.	demonstrate how the semantic vector changes with time and only incorporates relevant data from fresh input. Input gates, which identify and attenuate unnecessary information, have made this possible. They suggested and used user click-through data from a for-profit web search service to train the model with a minimal supervision signal due to the overall lack of available human-annotated data.	
21.	[25]	The technique for extracting key phrases from Bengali documents using LSTM recurrent neural networks (RNNs) is presented in this paper. First, they preprocess Bengali text and store it as token sequences. The effectiveness of the LSTM-based method for extracting Bengali document key phrases is demonstrated through experimental evaluation.	Rather than using the conventional ranking method, the authors have used the LSTM recurrent neural network in our suggested system to generate the confidence score of candidate key phrases. Furthermore, the authors generated candidate key phrases using the conventional n-gram method. Additionally, stop words, verb suffix filtering, and stemming were employed to narrow the pool of potential phrases.	Bangla Dataset
22.	[26]	The approach for multilingual material retrieval with help of deep learning techniques is proposed in this paper. At first, they use word embeddings to represent text data that has been pre-processed from several languages.	Using deep learning, a multilingual information retrieval system was created. Results from multilingual searches have an accuracy of 91.19%. The deep learning model and similarity score were used to get these	Tourism Dataset

		They then use a deep learning architecture to train language-independent depiction of documents, maybe using a (CNN) or RNN	findings. Results from bilingual searches have a 70% accuracy rate. This value is the sole product of the cosine similarity score. Therefore, the method of retrieving information by deep learning is better.	
23.	[27]	An integrated model for information retrieval based on neural networks is presented in this research. The model efficiently ranks and retrieves relevant pages by integrating many retrieval strategies, such as semantic similarity measures and keyword matching. The effectiveness of the suggested integrated model in information retrieval tasks is demonstrated by experimental evaluation.	In this work, an integrated approach to CIA management was presented. The experimental findings demonstrated that the suggested RMSprop outperformed the earlier AdaGrad and Adam algorithms in terms of performance. Compared to the current methods, the suggested algorithm has numerous advantages in terms of accuracy, speed, detecting comparable words, etc. RMS prop is utilized to increase our model's accuracy and speed. Word similarity is determined using LSTM.	Proposed Corpus
24.	[28]	In this review, another technique in light of RNN for extraction of table fields in business records is presented. Before separating highlights, they pre-process the archives and convert the pictures of reports into text. Then, they utilize the RNN construction to lay out conditions and investigate grouping examples of the archive text. The prepared	They suggest an RNN strategy for table field extraction in business papers, which remarkably works on the earlier work, as deduced in the paper. Applying RNN to the assignment, the upsides of table fields are effectively recognized integrating the succession of the text input. As per the consequences of the trials, the utilization of	Business Document Dataset

		assignment of the model is to remember them and concentrate on their fields with high accuracy.	the RNN-based method gives improved outcomes than conventional procedure, while offering the course of this innovation which is fit for upgrading the report examination.	
25.	[29]	The use of visual word vectors and recurrent neural networks can be utilized to characterize word visuals during the time spent distinguishing their matching occasions. They are then utilized in the informational index to find consecutive examples of the watchword implanting in RNN structures. The appraisal results demonstrate that our process is fundamentally effective in the time it takes to plan relevant catchphrases from old archival photos.	This research proposes a novel representation strategy for word images. One way to learn embedding vectors is by using a set of visual words. For any word image, the average of the associated visual embedding vectors can be used to represent it. On the other hand, RNN models the character sequences in word pictures. Thus, in this approach, word pictures are converted into fixed-length embedding vectors. Lastly, the two types of embedding vectors stated above are combined to represent all word pictures.	Mongolin Historical Documents
26.	[30]	A multi-scale CRNN model for Chinese papery medical document recognition is proposed in this paper. The model successfully recognizes text in medical documents by incorporating features from several scales. The multi-scale CRNN model outperforms other models in experimental validation regarding Chinese medical document recognition tasks.	In this study, the researchers propose a multi-scale CRNN for Chinese medical document identification after a thorough examination. The suggested model is capable of combining and successfully extracting multi-scale characteristics from various convolutional layers. According to experimental results, the suggested approach performs better on	Chinese Medical Document Dataset

			CMDD than the widely used approaches.	
27.	[31]	The technique for text line segmentation in ancient Indian handwritten writings using Faster R-CNN is presented in this research. Initially, they utilize the Quicker R-CNN model to preprocess archive picture and concentrate district suggestions. They then work on these ideas to unequivocally draw text lines. Finally, they employ present handling strategies to improve the nature of segmentation.	This study suggests a novel technique for TLS in photographs of old handwritten documents written in Devanagari script. The suggested approach, which is based on grouping methodology and projection profile, addresses several issues with current TLS systems. The indorsed technique can precisely recognize text lines even in situations when the reports are blurred, have dark patches on them, or have skewness and interline covering.	Historical Devanagari Documents
28.	[32]	A technique for table examination and data abstraction from clinical lab information is proposed in this exploration. they begin by practicing image processing techniques to preprocess record pictures and concentrate tables. The suggested approach shows its adequacy in overseeing organized information in clinical archives by accomplishing exact table examination and data extraction from clinical research center outcomes.	The authors of this work develop a method for information extraction from reports from medical laboratories. They segment and recognize texts using efficient algorithms, as per the document's structure. To be more precise, they identify multi-scale lines that are utilized to segment and des-kew table regions. They can comprehend the table structure more fully thanks to our system's top-down pipeline for text detection than they can with bottom-up techniques.	Medical Laboratory Reports
29.	[33]	The method for text discovery and recognition in photos of medical laboratory transactions is proposed in this paper. They start	This investigation offers a sophisticated method for word recognition and recognition from images of clinical research center	Chinese Medical Documents Dataset

		by applying methods like morphological operations and linked component analysis to preprocess photos and extract text sections. The model performs better than conventional techniques after being taught end-to-end to precisely identify and detect text in photos of medical laboratory reports.	outcomes. First, given an image of a clinical lab report, a finder that generates a number of leaping boxes with messages is trained using a fix-based preparation technique. Next, a link structure is fed into a recognizer, which uses the jumping enclosed sections of the source image as information sources to interpret messages.	
30.	[34]	An improved Convolutional Recurrent Neural Network (CRNN) for scene text recognition is presented in this paper. They start by extracting text regions from scene photos and preprocessing them. Then, to better capture textual information and increase recognition accuracy, they modify the CRNN architecture by adding extra layers and attention processes.	In this study, the author suggests an improved CRNN for scene text acknowledgment. Six public datasets are used to evaluate the enhanced CNN model for word arrangement recognition in images, which was trained on fictitious datasets. This paper's primary contribution is the enhancement of the original CRNN network and the introduction of a novel neural network-based approach to text recognition.	SVTP, CUTE
31.	[35]	CRNN architecture for document classification is presented in this research. They first preprocess text data and represent it using token sequences. Next, they use recurrent layers to capture sequential dependencies and convolutional layers to extract	This paper proposes a Convolutional Recurrent Neural Network model for text characterization by joining the current neural network, which is profitable in separating neighbor-hood. The results of the early analysis demonstrate that our model performs well in precisely classifying text data in both Chinese and English.	DBPedia

CHAPTER 3

METHODOLOGY

This chapter describes the research design, including the reason for the chosen strategies and their relation to the study's objective. It describes the sample procedures, data-gathering instruments, and analytical tools utilized to answer the study objective.

3.1 Document Retrieval

One of the most crucial concepts in healthcare for granting access to critical data from large text collections is document retrieval (DR) [1]. Several Search engines like Google, and Bing are highly reliant on on document retrieval algorithms so that they get the most relevant documents from large datasets. It describes the use of automated techniques to find and pick the most pertinent data from unstructured medical documents, like research papers, discharge summaries, and clinical notes. To manage this new and continually amassing clinical information, clinical information mining to remove significant well-being information has become significant, particularly during the Coronavirus pandemic [2]. This phase is essential for directing clinical judgments, guaranteeing effective medication research, and acting as the cornerstone of public health initiatives.

The complicated medical records that are acquired by conventional means might be difficult to handle. In order to address the problem, we combine the previously described approach with the help of natural language processing techniques, which seem particularly effective in this regard. As a result, they can now extract information and provide explanations from reports thanks to developments in neural network models, particularly CNNs [3] and RNNs [4]. When analyzing word relationships within a text, CNN is the preferred network, while RNN performs better when handling sequential data.

Document retrieval in healthcare research comprises the systematic process of gathering relevant information from different sources to solve specific research inquiries. Researchers start by unveiling their data requirements, which could be questions about treatment effectiveness to questions about disease patterns.

Once the initial search results are obtained, researchers filter and evaluate the retrieved documents based on relevance, credibility, and quality. After retrieving the important documents researchers extract important information from the retrieved documents and synthesize them to address the research objectives. This synthesis process may include statistical analysis, qualitative coding, or thematic synthesis, depending on the nature of the research.

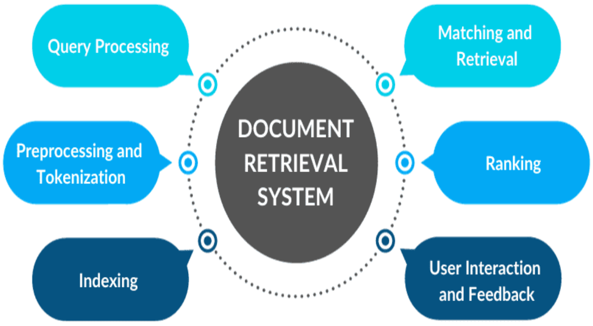

DOCUMENT RETRIEVAL SYSTEM
Query Processing
Matching and Retrieval
Preprocessing and Tokenization
Ranking
Indexing
User Interaction and Feedback

Fig.3.1 Document Retrieval System [53]

3.2 Dataset

Because of the Coronavirus pestilence, the White House and numerous eminent logical gatherings fostered the Coronavirus Open Logical Dataset (CORD 19) [2]. North of 400,000 over a million examination distributions on Coronavirus, SARS-CoV-2, and related COVID-19 containing CORD 19 are full-text accessible. This dataset, which is freely accessible to the world, is consequently made accessible to the worldwide examination local area so they can apply the latest normal language handling and other simulated intelligence exploration to dynamic illnesses, possibly prompting the disclosure of advancements that could be useful to in the battle against this disease. The interest in these strategies is ascending because of the fast speed increase of new COVID writing, and the clinical exploration local area is finding it challenging to stay aware of it.

The White House and an alliance of driving exploration establishments are pushing the world's computerized reasoning specialists to foster text and information mining instruments that will assist the clinical local area with tackling significant logical issues. The biggest machine-meaningful Covid writing assortment presently accessible for information mining is the CORD 19 [2] dataset.

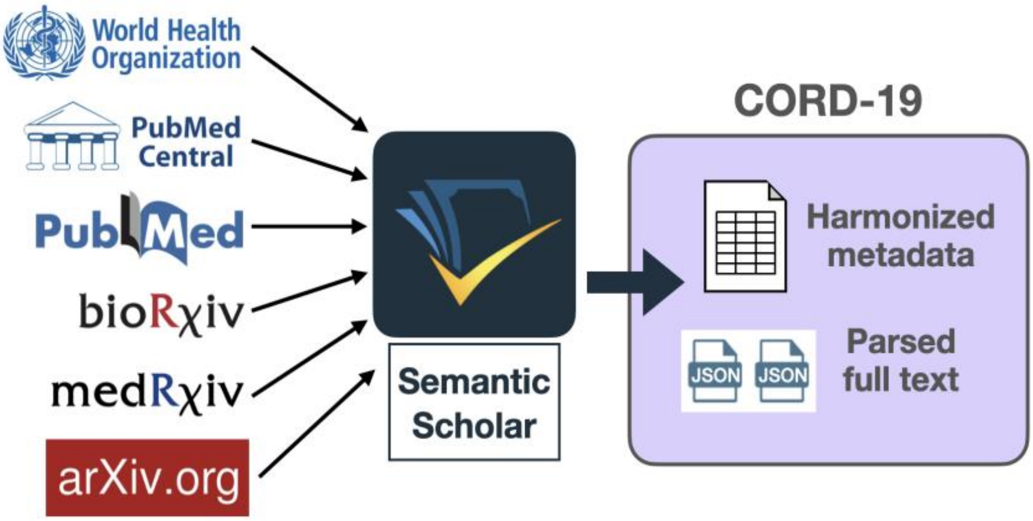


Fig.3.2 CORD-19 dataset [54]

This furnishes the global man-made intelligence research local area with the amazing chance to apply text and information mining apparatuses to interface experiences across and find answers for issues inside this substance, supporting the continuous Coronavirus reaction tasks around the world. The clinical local area finds it more challenging to keep awake to date with the developing criticalness of these medicines due to the Covid writing's quick turn of events.

3.3 Software Requirements

Python programming is required to construct this model, together with Jupyter Notebook and Anaconda Navigator tools.

3.3.1 Python Programming

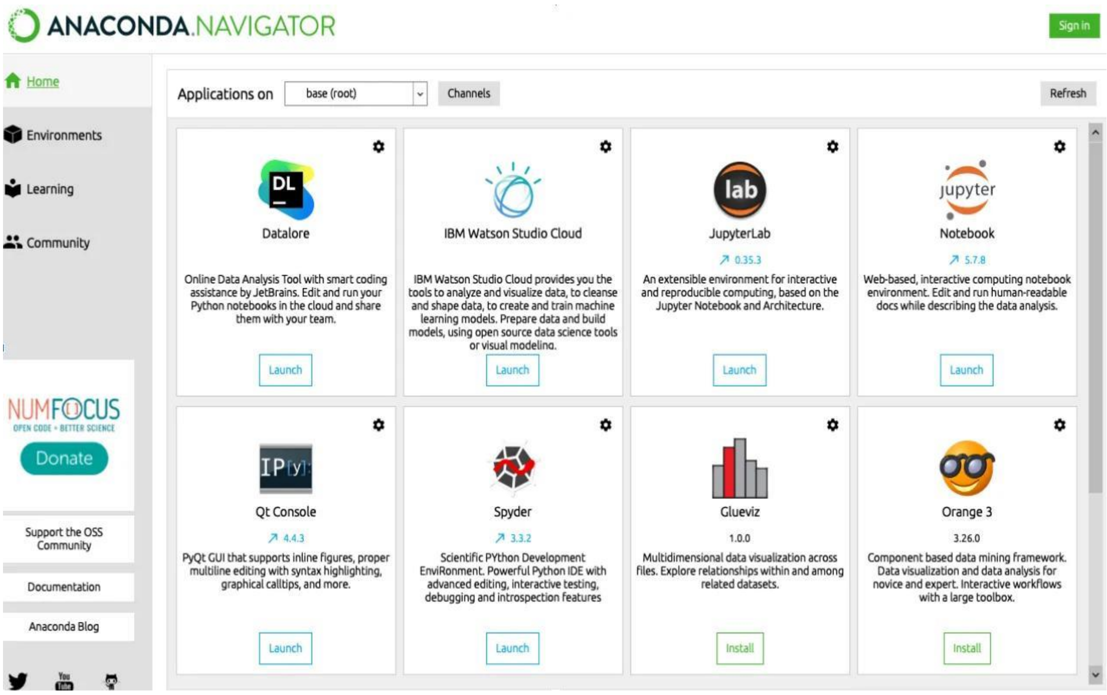
Python is a broadly material, reasonable, and significant-level programming language. Python was planned by Guido van Rossum and first made accessible in 1991. Its plan reasoning accentuates code clarity and huge utilization of whitespace. Its thing-based idea and phonetic headways are intended to help computer programmers create reliable, adequate code for both little and enormous-scope projects. Python utilizes trash assortment and dynamic creation.

It is viable with a few programming dialects, including object-arranged, utilitarian, and organized (particularly procedural). The ABC language was supplanted with Python in the last part of the 1980s. With the arrival of Python 2.0 in 2000, highlights like references including in the trash assortment framework and rundown understandings were incorporated. 2008 saw the arrival of Python 3.0, a critical update to the language that isn't completely reverse viable and requires changes for most Python 2 projects to work appropriately on Python 3.

3.3.2 Anaconda Distribution and Anaconda Navigator

Anaconda is an open-source dissemination of the logical processing programming languages R and Python. Bundles for information science that are viable with Windows, Linux, and macOS are remembered for the delivery made to make bundle the executives and arrangement more straightforward.

The Anaconda distribution dispersion accompanies a work area graphical UI (GUI) Anaconda Navigator, which empowers clients to oversee conda bundles, conditions, and diverts as well as sending off applications without requiring order line input. Navigator can search for bundles in a nearby Anaconda repository or on Anaconda Cloud. It can likewise introduce and refresh bundles in a climate. Linux, macOS, and Windows can all utilize it. Anaconda Distribution is an extensive tool for Python and R programming, especially appropriate for information science, AI, and huge-scope information handling. It incorporates a wide cluster of pre-introduced bundles, conda for bundle and environment management, and similarity with the Jupyter Scratch-pad. Anaconda Navigator, a graphical point of interaction, improves on bundling the executives, climate generation, and admittance to instruments like Jupyter Lab, Spyder, and Versus Code. It gives an easy-to-understand method for overseeing and sending off applications, increasing the transparency and efficiency of intricate information science job procedures.



Online Data Analysis Tool with smart coding assistance by JetBrains. Edit and run your Python notebooks in the cloud and share them with your team.

IBM Watson Studio Cloud provides you the tools to analyze and visualize data, to cleanse and shape data, to create and train machine learning models. Prepare data and build models, using open source data science tools or visual modeling.

An extensible environment for interactive and reproducible computing, based on the Jupyter Notebook and Architecture.

Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.

PyQt GUI that supports inline figures, proper multiline editing with syntax highlighting, graphical calltips, and more.

Scientific PYthon Development EnviRonment. Powerful Python IDE with advanced editing, interactive testing, debugging and introspection features

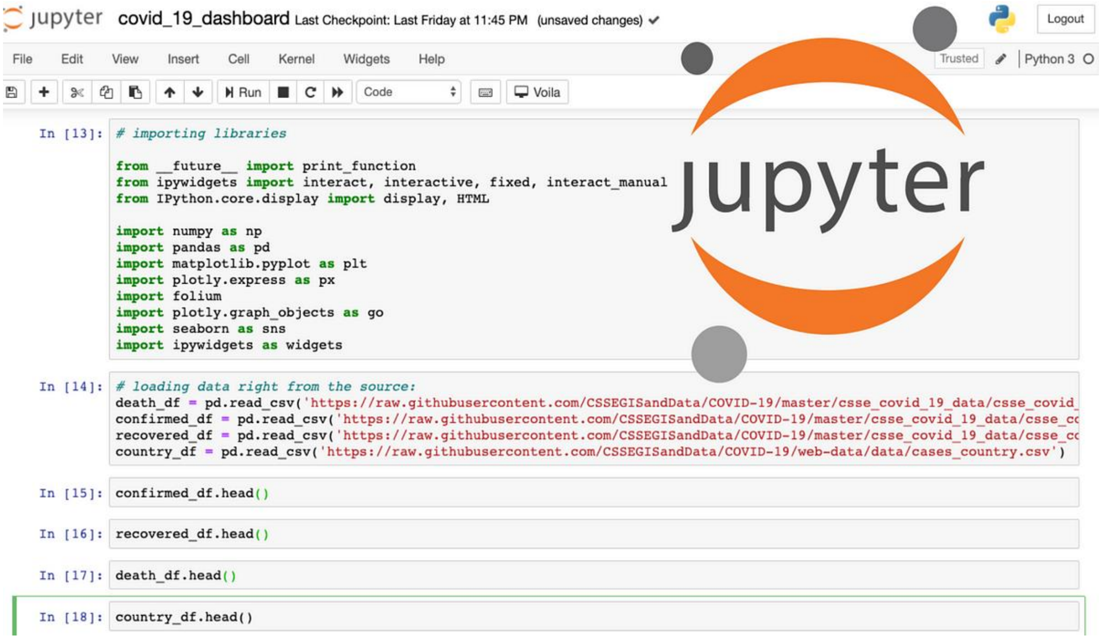
Multidimensional data visualization across files. Explore relationships within and among related datasets.

Component based data mining framework. Data visualization and data analysis for novice and expert. Interactive workflows with a large toolbox.

Fig.3.3 Anaconda distribution [55]

3.3.3 Jupyter Notebook

We might make and impart archives to live code, math, designs, and story text utilizing the open-source web instrument Jupyter Notepad. Information change and cleaning, factual displaying, mathematical reproduction, information perception, and numerous different applications are among the purposes.



```
In [13]: # importing libraries

         from __future__ import print_function
         from ipywidgets import interact, interactive, fixed, interact_manual
         from IPython.core.display import display, HTML

         import numpy as np
         import pandas as pd
         import matplotlib.pyplot as plt
         import plotly.express as px
         import folium
         import plotly.graph_objects as go
         import seaborn as sns
         import ipywidgets as widgets

In [14]: # loading data right from the source:
         death_df = pd.read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_
         confirmed_df = pd.read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_co
         recovered_df = pd.read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_co
         country_df = pd.read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/web-data/data/cases_country.csv')

In [15]: confirmed_df.head()

In [16]: recovered_df.head()

In [17]: death_df.head()

In [18]: country_df.head()
```

Fig.3.4 Jupyter notebook [56]

3.4 Necessary Libraries

3.4.1 NumPy

NumPy [36] is a common tool used by scholars, experts, and algorithm inventors in machine learning. NumPy is a programming language that is based on Python, which is widely used. It has several operations and functions, as well as multidimensional arrays, which are the basic data structures for scientific computing. It helps to store the array efficiently and helps to save a lot of space. The Python community has created libraries based on NumPy, including the machine learning package sci-kit-learn (Sklearn).

3.4.2 Pandas

Python developers may use pandas [37], a collection of high-level data structure designs and tools, to do robust information research. The main goal of pandas is to facilitate our ability to locate information fast.

3.4.3 Matplotlib

Matplotlib [38] is a Python 2D charting tool that produces production-excellence graphics. It is a graph plotting toolkit designed for visualization purposes. It is a tool that may be interactive or non-interactive, and it allows a variety of picture storage formats. Matplotlib works perfectly with other Python libraries like NumPy and Pandas, allowing users to effortlessly view data kept in arrays, matrices, and data frames. It supports a broad series of plot forms, including bars, pie charts, histograms, and lines. Furthermore, it is extremely versatile, user-friendly, and customizable.

3.5 Data Cleaning

Data cleaning is essential to discovery and abolishing any duplicate, irrelevant, or missing data. It is necessary because if we train our model with error data, it will not be properly trained and will not predict results with high frequency.

When it comes to data preparation, information purification is a critical stage that contributes to information quality assurance. It is critical that the fundamental information on the discovery be great. If the material is not cleansed to remove errors, exceptions, missing data, or irregularities, it is absolutely possible to make an error or even mislead about the results of an inquiry. Finally, with clear information, it is easier to display and identify instances since computation is most beneficial when provided with exact, high-quality data that is free of errors. Together, clear data helps in outcome translation and the maturation of execution ideas.

Stages to do to clean up data:

1. Removing Unwanted Discernments: Determine whether the insights are unimportant, and then delete them to reduce dimensionality. This cycle includes analyzing extra records, repeated entries, and fields that do not contribute to the investigation.
2. Correcting structural errors: Identify any faults with the information's architecture, such as factor kinds, variable names, information designs, or any combination of these that may exist in the current dataset. The final advantage will be the existence of consistent arrangements, modified names, and a worked-out structure for information depiction. Tackling primary issues

upgrades the dependability of the data which is useful for investigating and deciphering.

3. Taking care of Undesirable Anomalies: Circle or pick the perceptions that are far away from the mean or the normal of the dataset. Depending upon the unique situation, in which outliers impact the examination, it is prescribed to change or kill them. Discarding outliers is considered a drawback since it reduces the accuracy and power of data processing.
4. Overseeing Missing Information: There are different conceivable outcomes on how to deal with missing information, for example, for example, utilizing more perplexing techniques like ascription or barring inadequate and missing records and measurable attribution. Purifying the information is especially essential in any given dataset as well as in diagnosing issues in the examination.

3.6 Pre-Processing Stages

In this we understand the different phases of pre-processing and how it works.

3.6.1 Document Parsing

The extraction of organized information from unstructured materials presents a troublesome interaction known as document parsing. Unstructured records are those that contain incorporate significant information but don't adhere to the predefined system, like solicitations or structures. By looking at the substance of the report, tracking down relevant data, and sorting it out into an organization that can be utilized, document parsing is the way to get this information.

Once the data structure has been defined and the type of information to be extracted has been determined, then we need to compile all the documents to be extracted into one location. The next actions that must be performed are as follows:

1. Text Extraction: Our archives may be in several formats, such as Word, PDF, or HTML. The first step is to convert these record kinds into text documents that computers can read. Although optical character recognition (OCR) is typically used for this, there are other ways of text extraction available.
2. Tokenization: The message is divided into "tokens," which are essentially words or subwords, so that the machine can investigate the data.
3. Named Entity Recognition (NER): Likewise alluded to as natural language processing (NLP), this method recognizes and categorizes items in a text, such as names, dates, quantities, and locations. A syntactic analysis is then conducted to better understand the text's linguistic construction.
4. Organisation: This addresses inadequacies, contradicting text, and disturbance in video recordings. After the data has been cleaned up, it is grouped into tables using key-esteem matches or another structured format.

3.6.2 Lexical Analysis

A crucial stage in NLP is lexical analysis, sometimes known as scanning. The lexical analyzer, also known as a lexer or scanner, reads the source code character by character in programming languages to classify the characters into tokens, which are the smallest units of code that have meaning. These tokens usually belong to one of the following categories: operators (arithmetic, logical, relational), punctuation (commas, semicolons, braces), keywords (reserved words with predefined meanings like if, while, return), and constants (like integers, doubles, characters, and strings).

3.6.3 Stemming and Lemmatization

By removing the word's affixes, stemming creates the base word from the inflected term. A predetermined set of regulations governs the dropping of these affixes. It should be mentioned that stemmers may or may not produce base words with semantic meaning. Compared to lemmatizers, stemmers are less computationally expensive and speedier. Lemmatization is the process of combining similar words' inflected forms. In this manner, we can access any word's fundamental shape that has inherent meaning. The Lemma is the term for this basis.

3.6.4 Word Embedding

One method for representing words and documents is Word Embedding. A word is represented in a lower-dimensional space by a numerical vector input known as word embedding or word vector. It permits words to have comparable representations when they have similar meanings. Word embeddings are a way to take textual properties and turn them into machine-learning features that may be used with textual data. They make an effort to maintain semantic and syntactic information. The word count of a sentence is the basis for techniques like Bag of Words (BOW), Count Vectorizer, and TF-IDF; syntactical or semantic information is not saved. The amount of vocabulary elements determines the size of the vector in these methods. If the majority of the elements are 0, we can have a sparse matrix.

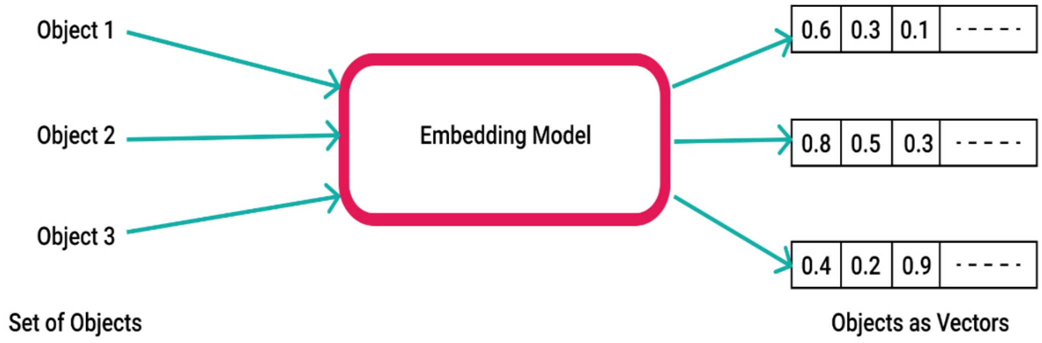


Fig. 3.5 Embedding model [57]

Large input vectors will result in an enormous number of weights, increasing the amount of computation needed for training. Word Embeddings provide an answer to these issues. But we use a neural approach to perform word embedding and the famous technique used to perform word embedding is DOC2Vec. It simply converts each word of the sentence into a vector representation. The underlying idea of using this is words that consist of the same meaning have the same vector representation.

3.7 Dimensionality Reduction

Reducing the number of items in a dataset while preserving the majority of the associated data is the aim of dimensionality reduction. One important area where high-layered information is handled is machine learning. The well-known problem in machine learning known as the "curse of dimensionality" states that a model's performance decreases with increasing element count. This is due to the fact that a model's complexity increases with the number of attributes it contains. This leads to a lack of proper training, which makes it more difficult to come up with a workable solution because our model is often confused by trying to comprehend an infinite number of variables. data. The two essential strategies for diminishing dimensionality are feature extraction and feature selection.

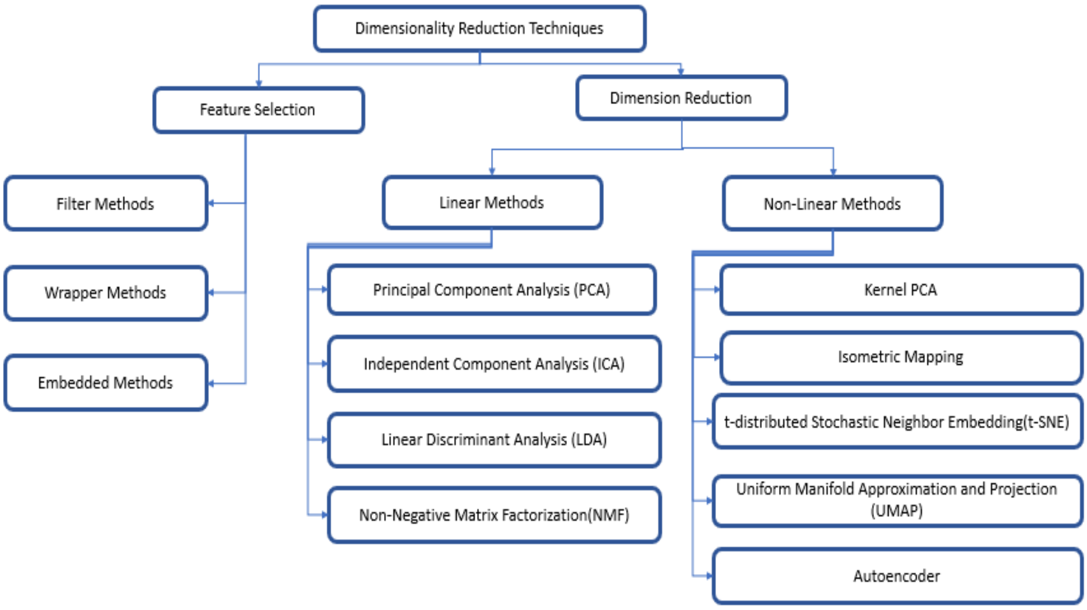

Dimensionality Reduction Techniques
Feature Selection
Dimension Reduction
Filter Methods
Wrapper Methods
Embedded Methods
Linear Methods
Non-Linear Methods
Principal Component Analysis (PCA)
Independent Component Analysis (ICA)
Linear Discriminant Analysis (LDA)
Non-Negative Matrix Factorization(NMF)
Kernel PCA
Isometric Mapping
t-distributed Stochastic Neighbor Embedding(t-SNE)
Uniform Manifold Approximation and Projection (UMAP)
Autoencoder

Fig.3.6 Dimensionality reduction technique [58]

3.7.1 Feature Selection

Feature Selection is the cycle that assists us with concluding which highlights are pivotal for the model. The groundwork of all AI methods is highlight designing, which comprises fundamentally two stages: feature extraction and feature selection. Processes for feature selection and extraction might have a similar objective, however they are very unmistakable from each other. The main contrast between the two is that feature extraction produces new elements, though feature selection centers around picking a subset of the underlying component assortment.

Algorithms for feature selection fall into three broad categories: filter techniques, wrapper methods, and embedding methods.

1. Filter Methods: Usually, these techniques are applied during the pre-processing phase. Regardless of whether machine learning algorithms are used, these methods choose features from the dataset. Although these techniques are excellent at eliminating redundant, correlated, and duplicate features, multicollinearity cannot be eliminated by them.
2. Wrapper Methods: It is also called greedy algorithms, and use an iterative process to train the algorithm using a subset of features. Features are added and removed according to the findings drawn from training the preceding model. The primary benefit of wrapper methods over filter methods is that they offer the best possible collection of features for model training, which improves accuracy compared to filter methods but comes at a higher computational cost.
3. Embedded method: The feature selection algorithm in embedded methods has its built-in feature selection methods because it is integrated into the learning process. The disadvantages of filter and wrapper techniques are met by

embedded methods, which combine their benefits. These techniques take into account a variety of features in addition to being quicker and more accurate than filter techniques.

3.7.2 Feature Extraction

Feature extraction is the technique in which the original features are combined or changed to generate new features. The feature extraction approach reduces dimensionality in data by selecting a subset of characteristics that provide similar information. This involves placing multiple locations with too many dimensions on a map with fewer dimensions. With it, we can solve the problem of overfitting. There are three significant kinds of element extraction procedures that we by and large use specifically Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Gaussian Discriminant Analysis (GDA).

1. Principal Component Analysis (PCA): It reduces the dimensionality of a data set by finding a new set of variables that are smaller than the original set of variables while retaining the majority of the information that characterizes the sample, and such variables are useful for data regression and classification. PCA is a solo straight dimensionality Reduction method used to resolve the issue of overfitting. PCA is an AI approach generally used to foster expectation models and conduct exploratory information investigations

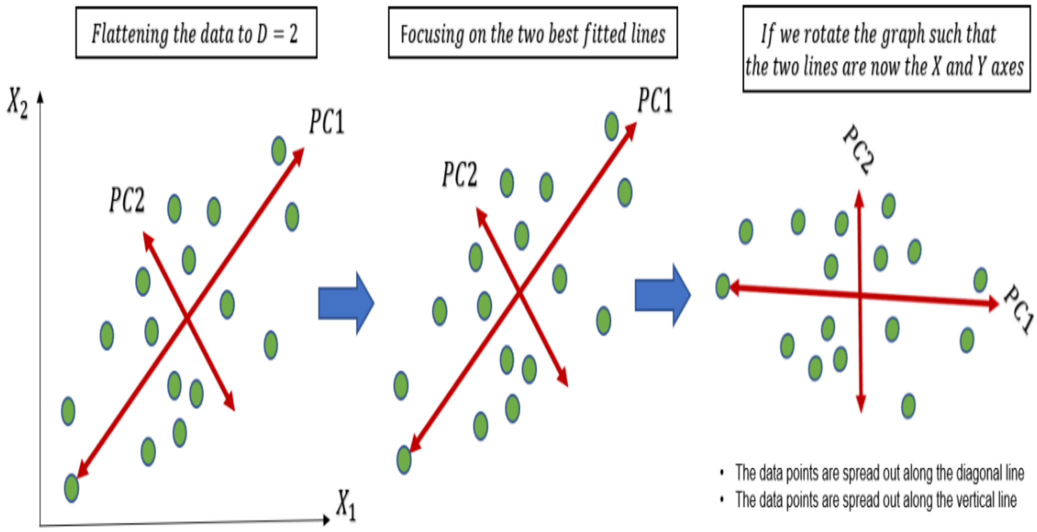

Flattening the data to D = 2
Focusing on the two best fitted lines
If we rotate the graph such that the two lines are now the X and Y axes
X_2
X_1
PC1
PC2
PC1
PC2
PC2
PC1
The data points are spread out along the diagonal line
The data points are spread out along the vertical line

Fig.3.7 PCA [59]

2. Linear Discriminant Analysis (LDA): It is a prominent multidimensionality reduction approach for supervised classification applications. It is also referred to as Normal Discriminant Analysis or Discriminant Function Analysis. It aims to maximize the separation of class with minimal variation. Its objective is to find a linear feature combination that maximally separates classes, resulting in a clear separation between them. LDA assumes that each document is created via a probabilistic process involving a collection of themes. It is assumed that papers are made up of a mixture of several themes, with each topic distinguished by its distribution over words. For example, we should efficiently categorize the two classes provided.

3. Gaussian Discriminant Analysis (GDA): It is a machine learning approach, that is used to solve classification issues. It is a Linear Discriminant Analysis (LDA) approach that does not impose strict equivalence on the covariance matrices across classes. GDA assumes that the properties of each class follow a Gaussian (normal) distribution. This assumption greatly simplifies both modeling and parameter estimates. To make GDA work, each class's mean and covariance matrix are computed assuming that the data in each class follows a Gaussian (normal) distribution.

3.8 Doc2 Vec

The Doc2Vec technique, based on neural networks, is used to learn the distributed representation of documents. The unsupervised learning technique allows us to map documents from a high-dimensional space to a fixed-length vector inside the input document space. The fundamental concept behind DOC2Vec is that words that are near to one other have a similar vector representation. Vector learning algorithms assign comparable texts to neighboring regions in the vector space. Thus, we may compare documents based on their vector representations, which allows us to perform tasks like document categorization, grouping, and similarity analysis.

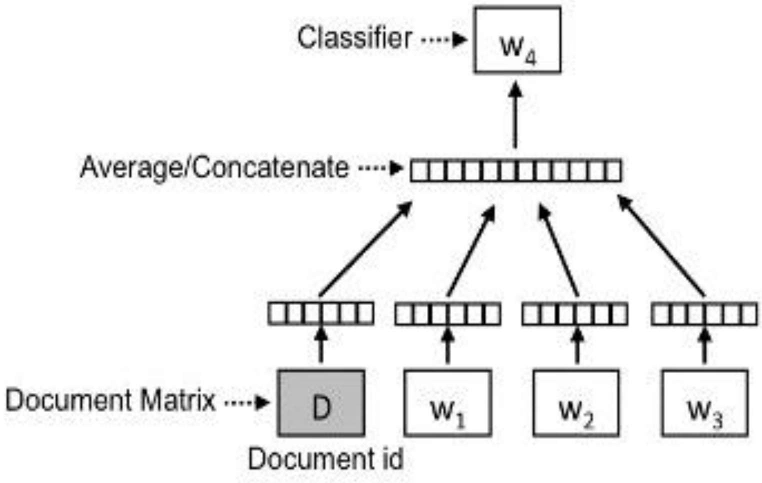


Fig. 3.8 Learning model of DOC2VEC [60]

The Doc2Vec approach has two primary variants. One is a distributed bag of words (DBOW) and the other is distributed memory (DM). The Distributed Memory model is a modification of the Doc2Vec model that expands on the well-known Word2Vec model. Distributed Memory works by learning a fixed-length vector representation depending on the context of each unique piece of text input (sentence, paragraph, or page).

DBOW, or Distributed Bag of Words, is a simplified version of the Doc2Vec technique that emphasizes word distribution within texts rather than word meaning. It ignores word order and records the frequency of each word.

3.9 Clustering Stage

Cluster analysis, or clustering, is the method of positioning data points into groups according to how similar they are to one another. The goal of clustering is to create homogeneous groups of data points from a diverse dataset. The points with the highest

similarity score are then grouped after the similarity is assessed using a metric such as Euclidean distance, Cosine similarity, Manhattan distance, etc.
To group related data points, there are generally two methods of clustering that can be done:

1. Hard Clustering: Every data point in this kind of clustering is either fully or partially a part of a cluster. Let's take an example where we need to cluster four data points into two clusters. Thus, every data point will be associated with either Cluster 1 or Cluster 2.

2. Soft Clustering: This kind of clustering assesses the chance or possibility that a given data point will be in a particular cluster rather than placing each point in a separate cluster. Let's take an example where we need to cluster four data points into two clusters. Thus, we shall assess the likelihood that a given data point will be included in one or both clusters. All of the data points are used to compute this probability.

3.9.1 K-Means Clustering

The most common way of preparing a calculation to work on unlabelled, unclassified information without human oversight is known as unsupervised machine learning. K is the number of groups in the bunching method. At first, the group centroid is doled out aimlessly. Then, information focuses are doled out to groups relying upon their separation from the centroid. After relegating each highlight to a group, a new bunch of centroids is made. The methodology is persistently more than once until a reasonable bunch is found. We expect that the quantity of bunches is fixed and that each point is relegated to a gathering.
There are conditions where K is ineffectively portrayed, in which case we ought to contemplate the best worth of K. K Means is the best gathering approach for data that is particularly segregated. It isn't suitable to utilize this social event when data centers are over. When diverged from other packing computations, K Means is quicker. It gives a tight connection between information things. K Means bunches don't give precise data about the group's quality. Various bunches might be framed because of introductory group centroid portions. The K Means technique is comparatively commotion delicate.

3.9.2 How the K-Means algorithm works?

We are furnished with a dataset of items with specific credits and values for those elements. Work includes classifying these things into gatherings. We will play out this work utilizing the K-means approach, an unsupervised learning apparatus. The letter "K" in the calculation's name addresses the ideal number of gatherings or groups into which we need to arrange our things. The strategy makes k groupings or similitude bunches of things. We use Euclidean distance to figure out which information focuses have a place wherein bunch.

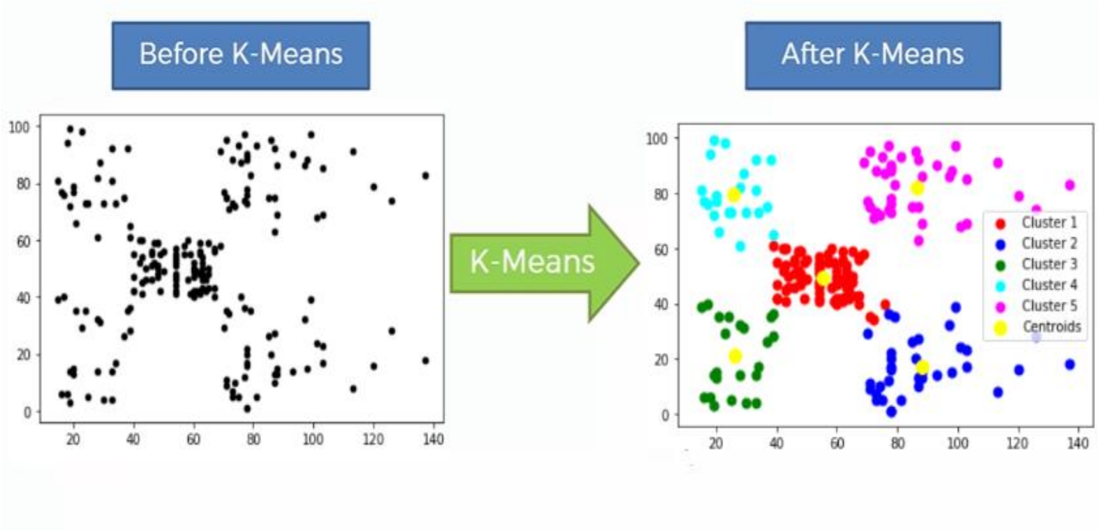


Fig.3.9 K-Means clustering [61]

The algorithm functions as follows:

1. At first, k focuses are randomly allotted and these k focuses are otherwise called group centroids or means.
2. Everything is gathered by the closest mean, and the directions of the mean which are the midpoints of the multitude of things ordered in that group hitherto are refreshed.
3. After a foreordained number of emphases, we again play out a similar cycle until we get our groups.

CHAPTER 4

DEEP LEARNING AND ENSEMBLE MODEL

Computational models containing a few handling layers can accomplish portrayals of information with several levels of abstraction through deep learning [39]. It ought to be underscored that these methodologies have currently essentially affected the cutting edge in a few fields, including drug improvement and hereditary qualities, voice acknowledgment and visual item ID, and article recognition. Deep learning handles various issues by portraying the backpropagation guideline, which presents alterations to the machine's inside loads in view of the portrayal marched in each layer as opposed to the portrayal strutted in each layer previously. Recurrent networks offer understanding into successive information, like voice and text, yet profound convolutional nets have made significant advances in picture, video, and sound handling.

It has been demonstrated to be exceptionally compelling at recognizing complex designs in high-layered information, making it helpful in an extensive variety of logical, business, and legislative settings. It has outflanked other AI procedures in foreseeing and dissecting particle smasher information, reconstructing brain circuits, and anticipating the impacts of changes in non-coding DNA on quality articulation and illness, as well as breaking records in picture and discourse acknowledgment.

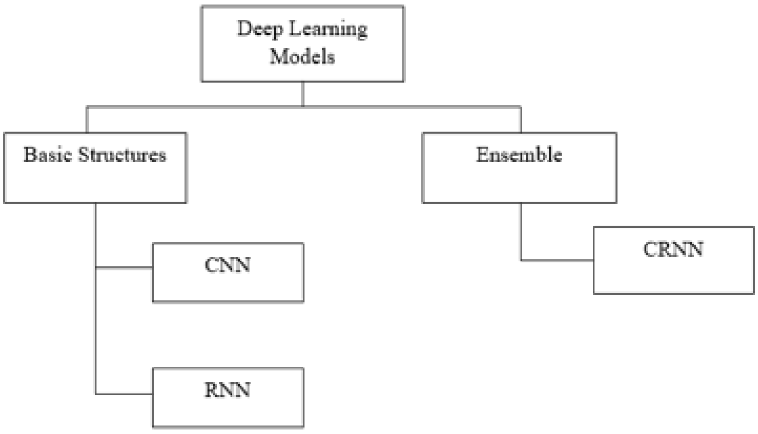


Fig.4.1 Different types of deep learning models

Progressive outcomes have been attained by DL models in NLP application domains. Ensemble deep learning models and Basic organized deep learning models can be used to group all deep learning models for document information retrieval. The term "Structure DL models" refers to deep learning models with their individual basic deep learning structures, such as CNN, RNN, and DNN. Ensemble DL models are those kinds of models that are constructed by integrating fundamental deep learning models, such as RMDL which helps to improve performance. The following is a brief explanation of those DL methods. These techniques provide a range of approaches to deal with difficult data problems and improve model performance, which characterizes the deep learning environment.

4.1 Convolutional Neural Network

Convolutional neural networks [40] apply a non-straight activation function as the result of a convolutional activity before pooling the information and characterizing it utilizing a full association layer. The channel, normally referred to as the kernel function is the focal part of a convolutional interaction. It travels through the first framework from left to right and from start to finish to complete element extraction. The estimation of the kernel function in regular language handling is many times equivalent to the width of the first network, and it just slides in the upper and lower headings, guaranteeing the trustworthiness of the word as the most minimal granularity in the language. There are two sorts of padding techniques utilized in the kernel function sliding cycle: valid padding and zero padding.

With its several layers, including embedding, convolutional, and pooling layers, CNN [41] offers the best feature extraction techniques. As the input layer, the embedding layer transforms every word from the pre-processed data into an embedding vector with embedding coordinates. The coordinates of the Convolution layer are decreased by the Pooling Layer. After receiving two-dimensional illustrations of words or features, the Flatten Layer transforms them into vector format that the Fully Connected Layer can use. One or more neural layers that employ activation functions and dropout rates to learn the model are referred to as the fully connected layer or dense layer. The model prediction is done using the Output Layer, whose activation function is SoftMax and whose nodes are equal to the dataset's class frequency.

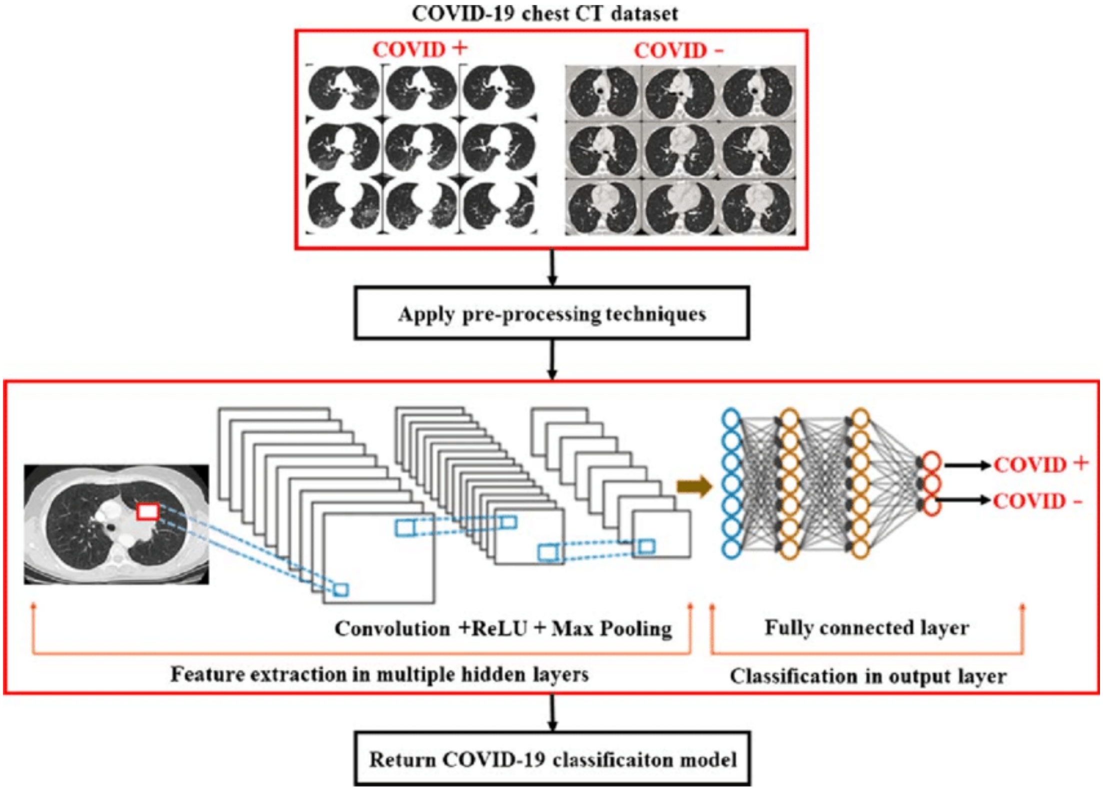

COVID-19 chest CT dataset
COVID +
COVID -
Apply pre-processing techniques
COVID +
COVID -
Convolution +ReLU + Max Pooling
Fully connected layer
Feature extraction in multiple hidden layers
Classification in output layer
Return COVID-19 classificaiton model

Fig.4.2 Working of CNN [62]

4.2 Recurrent Neural Network

When both the input and the output contain sequential structures, the order-to-order problem can be resolved using recurrent neural networks (RNN). Implicit relations typically exist between the structures. Nevertheless, the standard RNN model finds it challenging to thoroughly examine the relationships between the sequences [42].
Sequential data is a great fit for RNNs, which are essential for text mining and classification. They perform well when handling sequential data because they give preceding parts greater weights. RNNs are useful for several applications like image categorization because they can store temporal dependencies, which enables more sophisticated semantic analysis [43]. One of the most important features of the RNN paradigm is the way it collects the data from previous nodes, which helps to increase the semantic analysis of a given dataset. The input layer of the RNNs model is the implanting layer, which translates each word in the pre-processed data into embedding vectors with embedding coordinates. In addition to the dense layer, RNN architectures like GRU or LSTM are utilized to train the model.

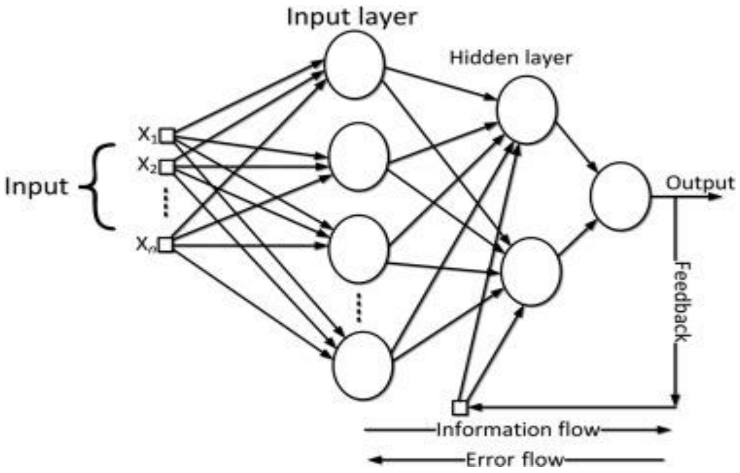


Fig.4.3 Recurrent Neural Network [63]

4.3 Convolutional Recurrent Neural Network

CRNN is an amalgamation of CNN and RNN, which is a novel framework for document extraction, that enhances the capability and efficacy by using the advantages of both models. CNNs help in detecting local patterns whereas RNNs focus on context [44] therefore, a hybrid of these two types of neural networks allows the development of exact spatial and temporal connections more effectively thereby improving document extraction accuracy. CNN mainly uses a convolutional layer to capture hierarchical features along with that RNN captures sequential dependencies which we can help to capture both the local and global context of text. This involves detecting complicated document structures because of RNN's “sequential handling” and CNN's ability to acquire “hierarchical visual signs [45]. It allows the model to focus on the crucial sections of the given dataset from which information needs to be retrieved. CNN is an already trained model that is built on the (ImageNet dataset) and RNN is based on the (Language Corpora) and helps to capture typical patterns and features for a given application. This helps to accelerate that consists of lesser “labeled data”. This property of CRNN is very useful for applications like scene text recognition, which requires knowing both the spatial arrangement of characters and the sequential flow of language.

Here's how CRNNs work:

1. Input: A CRNN accepts a sequence of data as input, such as audio samples or pictures.
2. Convolutional Layers: Convolutional layers are fed the input sequence to extract features from the input. These layers work especially well with image-based inputs and resemble those found in CNNs.
3. Recurrent Layers: One or more recurrent layers, which are especially useful for processing sequential data, are fed the output of a convolutional layer. Layers that are repeated preserve a concealed state that contains details about earlier entries in the sequence.
4. Relationships between recurrent and convolutional layers: Typically, a convolutional layer's output is sampled before being fed into a recurrent layer. In doing so, the computational complexity of the network is decreased without sacrificing crucial input properties.
5. Output: After passing through the fully linked final layer, the result of the former iterative layer generates a forecast for the input sequence. This forecast may take the form of a string of letters, words, or other task-relevant outputs.

4.4 Natural Language Processing

Full-text databases are growing quickly, and this has led to advances in natural language processing (NLP) technology. People working in NLP have suggested that NLP could be practically applied to text retrieval, mainly used for indexing, but maybe also for somewhat related tasks, such as file 'abstracting' or extracting; it could be used for user display or database formation, and it could be used at both shallow and deep content levels. The retrieval itself comes in a variety of forms, such as one-time searching, filtering, or routing; it is used in a variety of information and text refining activities, like labeling for different goals; and the content it covers, for instance, extends into hypertext.

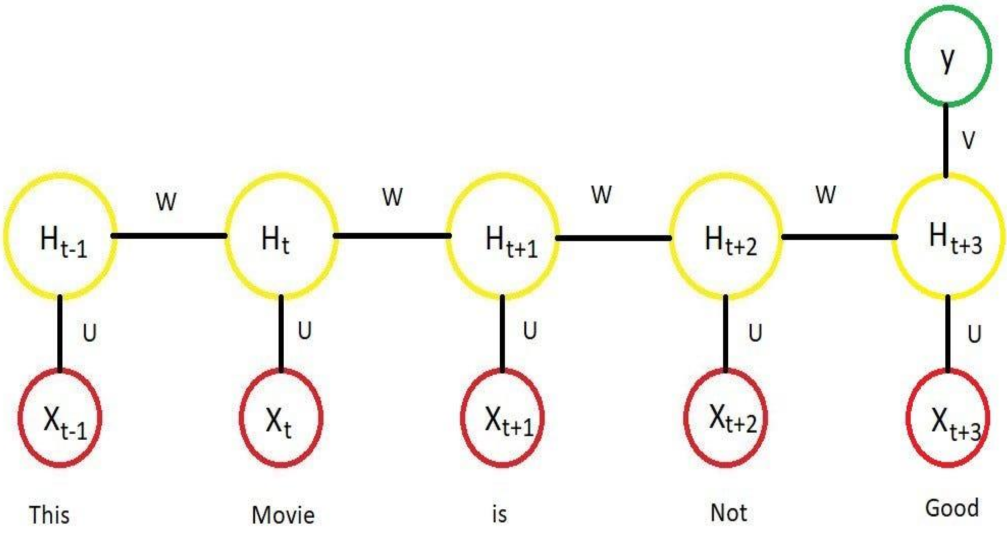


Fig.4.4 Natural Language Processing [64]

It is argued that linguistically driven analysis, and therefore natural language processing (NLP), is necessary not only for information extraction but even for more basic tasks like document retrieval given the present demanding environment in which large amounts of machine-readable content are becoming available.

Additionally, NLP might be required to create meaningful connections between different types of data and tasks within the information-selection task family by offering concept representations that could be used for related goals (like data and document queries) and to provide the user with more palatable information (like a display of important document concepts).

4.5 Long-Short-Term Memory

Contrasted with a standard RNN, Long Short-Term Memory (LSTM) [46] is a unique sort of RNN that jelly long haul dependence all the more effectively. This is particularly useful in settling the vanishing gradient issue. While LSTM and RNN share a chain-like construction, LSTM utilizes many entryways to painstakingly control how much data is allowed into every hub state.

4.6 Gated Recurrent Unit

The GRU is an architecture used in RNNs to handle extended-term relationships in consecutive data [47]. The GRU has gating mechanisms that GRUs are particularly useful in document extraction, where understanding context and relationships in sequential data is crucial.

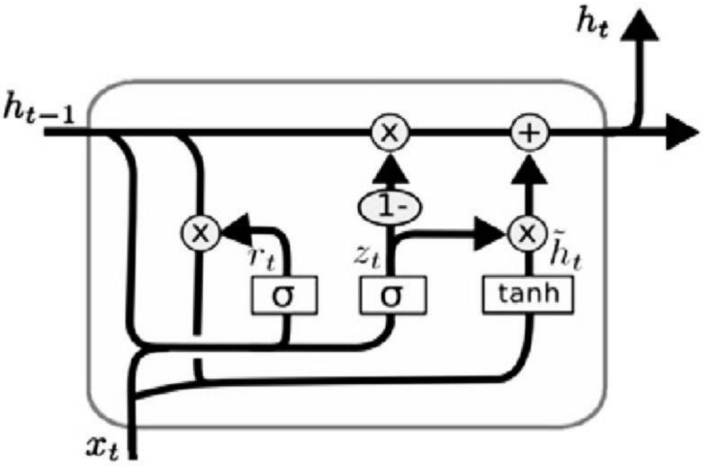


Fig.4.5 Gated Recurrent Unit [65]

They can selectively retain and discard information. GRUs feature two critical gates: the reset gate as well as the update gate. These gates function as intelligent filters, determining which information should be maintained from before and which is important to the subsequent phase. The reset gate decides what to discard, whereas the latest gate enables the selective transmission of important information.

CHAPTER 5

DATASETS

The COVID-19 pandemic prompted the White House and a group of eminent scientific institutions to establish the COVID-19 Open Scientific Dataset (CORD-19) [2]. More than 400,000 scholarly articles about COVID-19, SARS-CoV-2, and related coronaviruses are included in CORD-19 are fully text-accessible. This openly accessible dataset is made accessible to the global research communal so that they can utilize the latest advancements in natural language processing and other AI techniques to provide novel understandings that will support the ongoing fight against this infectious virus. The demand for these methods is rising due to the quick acceleration of new coronavirus literature, and the medical research community is finding it problematic to keep up with it.

These exploration establishments are pushing the world's artificial intelligent consciousness specialists to propel manuscript and information mining apparatuses that will uphold clinical foundations to take care of significant logical issues. The biggest machine-meaningful COVID writing assortment at present existing for information mining is the CORD- 19 [2] dataset. This furnishes the global artificial intelligence research local area with the event to apply text and information mining instruments to associate experiences across and find answers for issues inside this substance, supporting the continuous Coronavirus reaction activities around the world.

5.1 Similar Datasets

1. LitCovid [48]: The 2019 new coronavirus is the subject of several recently published PubMed papers that make up the LitCovid dataset. With over 23,000 items in total and about 2,000 new ones published each week, the dataset provides an extensive tool for researchers to stay informed about the COVID-19 pandemic.
2. Chest X-ray [49]: There are 30,805 unique patients' X-ray images (112,120) with ailment tags included in this NIH Chest X-ray Dataset. The disease classifications from the associated radiological reports were text-mined by the authors using Natural Language Processing to produce these labels. The labels are intended to be appropriate for weakly-supervised learning and to be more than 90% correct.
3. CONCORD [50]: An extensive collection of numerical claims taken from scholarly articles published on COVID-19-related research is available as an open-source dataset called the COVID-19 Numerical Claims Open Research Dataset (CONCORD). About 203k numerical claims relevant to COVID-19 are contained in CONCORD; these claims were taken from over 57,000 scientific research articles that were issued between January 2020 and May 2022. For further investigation, they used the raw dataset from the CORD-19 repository.

4. COVIDx CXR-4[51]: Researchers created the publicly available benchmark dataset COVIDx CXR-4, which consists of 30,882 CXR pictures from 17,026 patient cases. To enhance the dataset, more images might be contributed in the future. They are using this dataset to test and improve our COVID-19 identification models from CXR pictures.
5. COVIDSenti [52]: A massive manually interpreted COVID sentiment data set called COVIDSenti was created. It included 90,000 tweets that were crawled between February and March of 2020. Each tweet is allotted one of the three sentiment classes either it is optimistic, undesirable, or unbiased.

CHAPTER 6

RESULTS AND DISCUSSION

The CORD-19 dataset, which includes the COVID-19 documents, is used in our study. We reduce the initial dataset of 3 lakh queries to 10,000 entries to ensure a thorough analysis.

Important factors like sensitivity, F1-score, precision, and accuracy are carefully evaluated. Our study compares these measures amongst various neural network models to identify the best strategy.

1. Precision: The percentage of genuine positive predictions to the total of false positives and true positives is measured by the precision performance metric.
   Precision = True Positives / (False Positives + True Positives)
2. Recall: Analysing a model's capacity to prevent false negatives is essential. A high recall score indicates that there is less chance of false negatives because the model is good at identifying a significant percentage of pertinent positive cases.
   Recall = True Positives / (False Negatives + True Positives)
3. F1 Score: The F1 score is an average of recall and accuracy that is balanced. It assesses how well an algorithm can identify positive circumstances while reducing false negatives and possible positives.
   F1 Score = 2 * (Recall * Precision) / (Recall + Precision)

Table 6.1: Performance of Neural models

MODEL	Accuracy	Recall	Precision	F1-Score
CNN	98.84	98.3	97.1	97.7
RNN	82.65	85.7	83.4	84.5
CRNN	98.93	99.1	95.6	99.1

The CNN model obtains an accuracy of 98.84% with precision, sensitivity, and F1-score values of 97.1%, 98.3%, and 97.70%, respectively. On the other hand, the RNN model yields metrics for precision, sensitivity, and F1-score of 83.4%, 85.7%, and 84.50%, respectively, and an accuracy of 82.65%. CRNN, the fusion model, outperforms the two independent models with an accuracy of 98.93% and metrics of 95.60%, 99.1%, and 99.10% for F1-score, precision, and sensitivity, respectively. These results illustrate the classification performance of the CRNN model.

The differences in model performance that have been found point to the special qualities and benefits of the CNN, RNN, and CRNN architectures. When it comes to extracting spatial features from image data, RNNs do better than CNNs at capturing temporal correlations in sequential data.

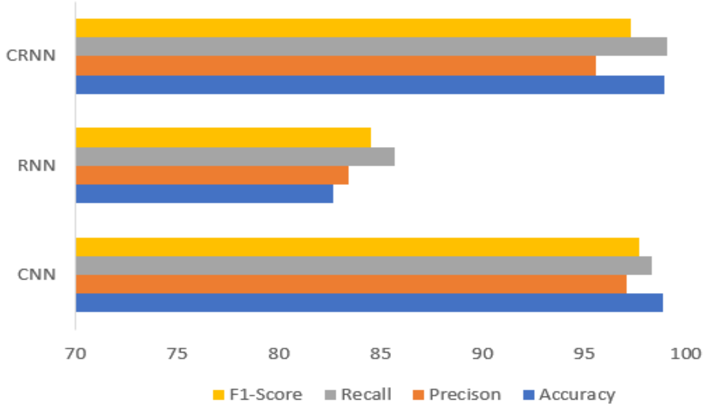


Fig.6.1 Comparison of Evaluation Measure

Improved classification accuracy and resilience are demonstrated by the CRNN model through the integration of geographical and temporal data. This emphasizes how important it is to use hybrid designs to take advantage of complimentary benefits and improve the model's overall performance.

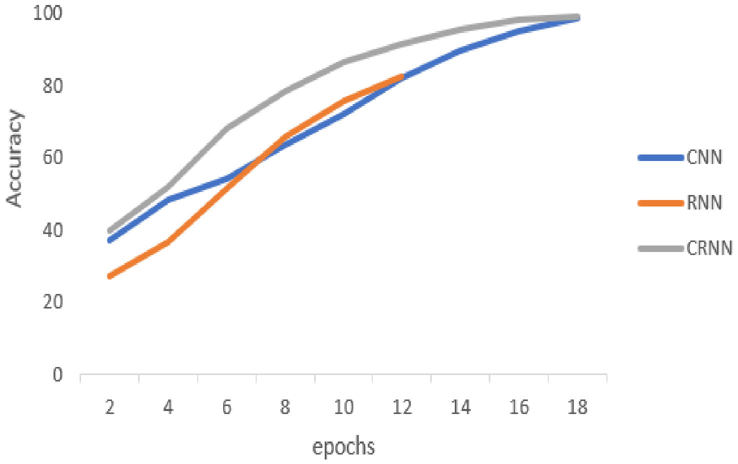


Fig.6.2 Accuracy of different models

CHAPTER 7

CONCLUSION, FUTURE WORK AND SOCIAL IMPACT

The following section explains in detail how deep learning models, particularly recurrent neural networks (RNNs), convolutional neural networks (CNNs), and their combinations, of convolutional recurrent neural networks (CRNNs), can help improve information retrieval from hospital patient documents. A systematic assessment of this research reveals that these models can track illness, particularly through a more comprehensive database of patient records. The fundamental distinction between RNN and CNN is the sequential usage of medical texts, with CNN determining local characteristics and spatial correlations. CRNNs stand out as an excellent approach to health care document categorization because they can capture the sequence of documents while maintaining location sensitivity at the same time, which can be done by the strong sequence processing of CNN combined with the sensitivity to the locality of RNN.

Researchers and practitioners can employ deep learning-based approaches to enhance document retrieval systems in healthcare through comparative analysis. It highlights the differences in these areas by demonstrating how specific healthcare information retrieval tasks are carried out. We can conclude from the study that CNNs, RNNs, and CRNNs are suitable for use in health document retrieval. As these documents become more widely available, patient outcomes may improve.

The comparison study will reveal how CNN, RNN, and CRNN models perform during the retrieval of relevant documents. CNN and RNN models perform well, but CRNN outperforms them all because of their greater resilience and accuracy. The next research effort, which will focus on architectural hybrid upgrades and optimization strategies, may push classification performance even higher. The CRNN model certainly has a higher potential for overall classifier performance. Future studies should assess the model's interpretability, scalability, and robustness against a variety of unexpected events and its actual utility for business purposes. Investigating these components may aid in the development of the model's broader application across other areas.

As important as it is to meet present project objectives, it is just as important to consider the project's future scope. Future advances are anticipated. The potential for overcoming these limitations is built into the advancement of innovation.

Below is a list of possible future scopes:

1. Multimodal Retrieval: Examine methods that have advanced beyond text extraction to data processing using pictures or audio information. Together with the visual audio data processing technology, we can modify document retrieval systems to be more suited for a wide range of distinct consumers
2. Enhanced Retrieval Algorithms: Create more complex, profound learning models for archive recovery that use cutting-edge tactics such as support learning, center systems, and BERT transformers.

3. Protecting privacy Retrieval: Use federated learning techniques or query encryption to effectively retrieve documents while leveraging technologies that protect user privacy.
4. Semantic Understanding: To increase conviction in the searched content, strengthen the deep learning model's capability to entirely realize the semantic background of the request.
5. Evaluation Metrics: Define new evaluation benchmarks that are improved at measuring the effectiveness of deep learning-based retrieval systems and account for the diversification of documents, user satisfaction, and results context.

Following are some social impacts of our project in the healthcare domain:

1. Better Patient Care: Medical workers will be able to access a variety of medical writing, treatment recommendations, and research publications online. With this system in place, medical workers may approach circumstances more thoughtfully, improving the patient's condition and, eventually, identifying answers to key health challenges.
2. Quicker Diagnosis and Treatment: In instances when time is crucial in healthcare, such as emergency and urgent care, deep learning-based retrieval systems allow medical professionals to retrieve data as rapidly as possible. This drastically lowers the time required for diagnosis, allowing for speedier treatment decisions.
3. Advancement of Medical study: Improvements in medical literature and research data enable scientists to conduct more quick and thorough research nowadays. Furthermore, new medical instruments are being developed, genetic engineering is being found, and medical research and innovation are moving more rapidly.
4. Patient Empowerment: Thanks to advancements in healthcare systems, patients may now obtain timely and reliable medical information, enabling them to take an active role in choosing their treatments and taking care of chronic illnesses.
5. Decreased Medical Errors: The attempt is anticipated to reduce clinical errors, incorrect diagnoses, and unpleasant situations while improving patient well-being and treatment quality. This aligns with the organization's mission of improving access to evidence-based medical knowledge and best practices.

REFERENCES

[1] Yang, Y., Wu, Z., Yang, Y., Lian, S., Guo, F., & Wang, Z. (2022). A survey of information extraction based on deep learning. Applied Sciences, 12(19), 9691.

[2] Kanakia, A., Wang, K., Dong, Y., Xie, B., Lo, K., Shen, Z., ... & Wu, C. H. (2020). Mitigating biases in CORD-19 for analyzing COVID-19 literature. *Frontiers in research metrics and analytics*, *5*, 596624.

[3] Zaman, R., Bashir, R., & Zaidi, A. R. (2021). Image Classification and Text Extraction using Convolutional Neural Network. *Journal of Computing & Biomedical Informatics*, *2*(01), 89-95.

[4] Behera, B., & Kumaravelan, G. (2021). Text document classification using fuzzy rough set based on robust nearest neighbor (FRS-RNN). Soft Computing, 25(15), 9915-9923.

[5] Zhou, H. (2022). Research of text classification based on TF-IDF and CNN-LSTM. In journal of physics: conference series (Vol. 2171, No. 1, p. 012021). IOP Publishing.

[6] He, T., Huang, W., Qiao, Y., & Yao, J. (2016). Text-attentional convolutional neural network for scene text detection. *IEEE transactions on image processing*, *25*(6), 2529-2541.

[7] Li, C., Zhan, G., & Li, Z. (2018, October). News text classification based on improved Bi-LSTM-CNN. In 2018 9th International conference on information technology in medicine and education (ITME) (pp. 890-893). IEEE.

[8] Rhanoui, M., Mikram, M., Yousfi, S., & Barzali, S. (2019). A CNN-BiLSTM model for document-level sentiment analysis. *Machine Learning and Knowledge Extraction*, *1*(3), 832-847.

[9] Qing, L., Linhong, W., & Xuehai, D. (2019). A novel neural network-based method for medical text classification. Future Internet, 11(12), 255.

[10] Bedi, P. P. S., Bala, M., & Sharma, K. (2023). Extractive text summarization for biomedical transcripts using deep dense LSTM-CNN framework. Expert Systems, e13490.

[11] Akhter, M. P., Jiangbin, Z., Naqvi, I. R., Abdelmajeed, M., Mehmood, A., & Sadiq, M. T. (2020). Document-level text classification using single-layer multisize filters convolutional neural network. *IEEE Access*, *8*, 42689-42707.

[12] Xu, H., Dong, M., Zhu, D., Kotov, A., Carcone, A. I., & Naar-King, S. (2016, October). Text classification with topic-based word embedding and convolutional neural networks. In Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (pp. 88-97).

[13] Kastrati, Z., Imran, A. S., & Yayilgan, S. Y. (2019). The impact of deep learning on document classification using semantically rich representations. Information Processing & Management, 56(5), 1618-1632.

[14] Liu, X., Zhou, Y., & Wang, Z. (2019). Recognition and extraction of named entities in online medical diagnosis data based on a deep neural network. *Journal of Visual Communication and Image Representation*, *60*, 1-15.

[15] Yao, C., Shen, J., & Chen, G. (2015, December). Automatic document summarization via deep neural networks. In *2015 8th International Symposium on Computational Intelligence and Design (ISCID)* (Vol. 1, pp. 291-296). IEEE.

[16] Silvestri, S., Gargiulo, F., & Ciampi, M. (2019, June). Improving biomedical information extraction with word embeddings trained on closed-domain corpora. In *2019 IEEE symposium on computers and communications (ISCC)* (pp. 1129-1134). IEEE.

[17] Wu, Y., Jiang, M., Lei, J., & Xu, H. (2015). Named entity recognition in Chinese clinical text using deep neural network. *Studies in health technology and informatics*, *216*, 624.

[18] Chikka, V. R. (2016, June). Cde-iiith at semeval-2016 task 12: Extraction of temporal information from clinical documents using machine learning techniques. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)* (pp. 1237-1240).

[19] Buber, E., & Diri, B. (2019). Web page classification using RNN. *Procedia Computer Science*, *154*, 62-72.

[20] Khatri, C., Singh, G., & Parikh, N. (2018). Abstractive and extractive text summarization using document context vector and recurrent neural networks. arXiv preprint arXiv:1807.08000.

[21] Ghumade, T. G., & Deshmukh, R. A. (2019). A document classification using NLP and recurrent neural network. *Int. J. Eng. Adv. Technol*, *8*(6), 632-636.

[22] Li, P., Peng, L., Cai, J., Ding, X., & Ge, S. (2017, November). Attention based RNN model for document image quality assessment. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* (Vol. 1, pp. 819-825). IEEE.

[23] Zhang, Q., Wang, Y., Gong, Y., & Huang, X. J. (2016, November). Keyphrase extraction using deep recurrent neural networks on twitter. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 836-845).

[24] Palangi, H., Deng, L., Shen, Y., Gao, J., He, X., Chen, J., ... & Ward, R. (2016). Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*(4), 694-707.

[25] Meem, N. T. A., Chowdhury, M. M. H., & Rahman, M. M. (2018, September). Keyphrase extraction from bengali document using lstm recurrent neural network. In *2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT)* (pp. 461-466). IEEE.

[26] Dodal, S. S., &amp; Kulkarni, P. V. (2018, July). Multi-lingual information retrieval using deep learning. In 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-6). IEEE.

[27] Nimmani, P., Vodithala, S., & Polepally, V. (2021, May). Neural network based integrated model for information retrieval. In *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 1286-1289). IEEE.

[28] Sage, C., Aussem, A., Elghazel, H., Eglin, V., & Espinas, J. (2019, September). Recurrent neural network approach for table field extraction in business documents. In *2019 International Conference on Document Analysis and Recognition (ICDAR)* (pp. 1308-1313). IEEE.

[29] Wei, H., Zhang, H., & Gao, G. (2017, July). Representing word image using visual word embeddings and RNN for keyword spotting on historical document images. In *2017 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 1368-1373). IEEE.

[30] Zhao, Y., Xue, W., & Li, Q. (2018, September). A multi-scale CRNN model for Chinese papery medical document recognition. In *2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM)* (pp. 1-5). IEEE.

[31] Jindal, A., & Ghosh, R. (2023). Text line segmentation in indian ancient handwritten documents using faster R-CNN. *Multimedia Tools and Applications*, *82*(7), 10703-10722.

[32] Xue, W., Li, Q., Zhang, Z., Zhao, Y., & Wang, H. (2018, August). Table analysis and information extraction for medical laboratory reports. In *2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)* (pp. 193-199). IEEE.

[33] Xue, W., Li, Q., & Xue, Q. (2019). Text detection and recognition for images of medical laboratory reports with a deep learning approach. *IEEE Access*, *8*, 407-416.

[34] Yu, W., Ibrayim, M., & Hamdulla, A. (2023). Scene text recognition based on improved CRNN. Information, 14(7), 369.

[35] Wang, R., Li, Z., Cao, J., Chen, T., & Wang, L. (2019, July). Convolutional recurrent neural networks for text classification. In 2019 international joint conference on neural networks (IJCNN) (pp. 1-6). IEEE.

[36] Nishino, R. O. Y. U. D., & Loomis, S. H. C. (2017). Cupy: A numpy-compatible library for nvidia gpu calculations. *31st confernce on neural information processing systems*, *151*(7).

[37] Heydt, M. (2017). *Learning pandas*. Packt Publishing Ltd.

[38] Tosi, S. (2009). *Matplotlib for Python developers*. Packt Publishing Ltd.

[39] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436-444.

[40] Luan, Y., & Lin, S. (2019, March). Research on text classification based on CNN and LSTM. In *2019 IEEE international conference on artificial intelligence and computer applications (ICAICA)* (pp. 352-355). IEEE.

[41] Jaderberg, M., Simonyan, K., Vedaldi, A., & Zisserman, A. (2016). Reading text in the wild with convolutional neural networks. *International journal of computer vision*, *116*, 1-20.

[42] Wang, F., & Tax, D. M. (2016). Survey on the attention based RNN model and its applications in computer vision. *arXiv preprint arXiv:1601.06823*.

[43] Behera, B., & Kumaravelan, G. (2021). Text document classification using fuzzy rough set based on robust nearest neighbor (FRS-RNN). *Soft Computing*, *25*(15), 9915-9923.

[44] Sutskever, I., & Hinton, G. (2010). Temporal-kernel recurrent neural networks. *Neural Networks*, *23*(2), 239-243.

[45] Wang, R., Li, Z., Cao, J., Chen, T., & Wang, L. (2019, July). Convolutional recurrent neural networks for text classification. In *2019 international joint conference on neural networks (IJCNN)* (pp. 1-6). IEEE.

[46] Kowsari, K., Heidarysafa, M., Brown, D. E., Meimandi, K. J., & Barnes, L. E. (2018, April). Rmdl: Random multimodel deep learning for classification. In *Proceedings of the 2nd international conference on information system and data mining* (pp. 19-28).

[47] Alqahtani, F., Abotaleb, M., Kadi, A., Makarovskikh, T., Potoroko, I., Alakkari, K., & Badr, A. (2022). Hybrid deep learning algorithm for forecasting SARS-CoV-2 daily infections and death cases. *Axioms*, *11*(11), 620.

[48] Gutierrez, B. J., Zeng, J., Zhang, D., Zhang, P., & Su, Y. (2020). Document classification for covid-19 literature. *arXiv preprint arXiv:2006.13816*.

[49] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2097-2106).

[50] Shah, D., Shah, K., Jagani, M., Shah, A., & Chaudhury, B. (2023). CONCORD: Numerical Claims Extracted from the COVID-19 Literature using a Weak Supervision Approach. *Available at SSRN 4222185*.

[51] Wu, Y., Gunraj, H., Tai, C. E. A., & Wong, A. (2023). COVIDx CXR-4: An Expanded Multi-Institutional Open-Source Benchmark Dataset for Chest X-ray Image-Based Computer-Aided COVID-19 Diagnostics. *arXiv preprint arXiv:2311.17677*.

[52] Naseem, U., Razzak, I., Khushi, M., Eklund, P. W., & Kim, J. (2021). COVIDSenti: A large-scale benchmark Twitter data set for COVID-19 sentiment analysis. *IEEE transactions on computational social systems*, *8*(4), 1003-1015.

[53] Otten, N. V. (2023, October 30). Document retrieval made simple & practical how to guide in Python. Spot Intelligence. https://spotintelligence.com/2023/10/18/document-retrieval/.

[54] Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Burdick, D., & Kohlmeier, S. (2020). Cord-19: The covid-19 open research dataset. *ArXiv*.

[55] What is Anaconda?: Domino data science dictionary. (s.d.). Domino Data Lab. Recuperato 20 maggio 2024, da https://domino.ai/data-science-dictionary/anaconda.

[56] Tyagi, H. (2020, aprile 6). The Complete Guide to jupyter notebooks for data science. Medium. Towards Data Science. Recuperato 20 maggio 2024, da https://towardsdatascience.com/the-complete-guide-to-jupyter-notebooks-for-data-science-8ff3591f69a4

[57] Tripathi, R. (2023, June 23) What are vector embeddings, Pinecone. Available at: https://www.pinecone.io/learn/vector-embeddings/ (Accessed: 20 May 2024).

[58] Khandelwal, R. (2022) A comprehensive guide to dimensionality reduction, Medium. Available at: https://arshren.medium.com/a-comprehensive-guide-to-dimensionality-reduction-851624b7377d (Accessed: 20 May 2024).

[59] W., M. (2022, Oktober 11). A gentle introduction to principal components analysis. LinkedIn. Abgerufen 20. Mai 2024, von https://www.linkedin.com/pulse/gentle-introduction-principal-components-analysis-michael-wynn.

[60] Aubaid, A. M., & Mishra, A. (2020). A rule-based approach to embedding techniques for text document classification. Applied Sciences, 10(11), 4009.

[61] Kumar, R. (2022, March 27). target particular segment of customers using K means clustering. Medium. Medium. Abgerufen 20. Mai 2024, von https://medium.com/@ravikumar10593/target-particular-segment-of-customers-using-k-means-clustering-834a7faa0da7

[62] Singh, D., Kumar, V., Vaishali, & Kaur, M. (2020). Classification of COVID-19 patients from chest CT images using multi-objective differential evolution–based convolutional neural networks. European Journal of Clinical Microbiology & Infectious Diseases, 39, 1379-1389.

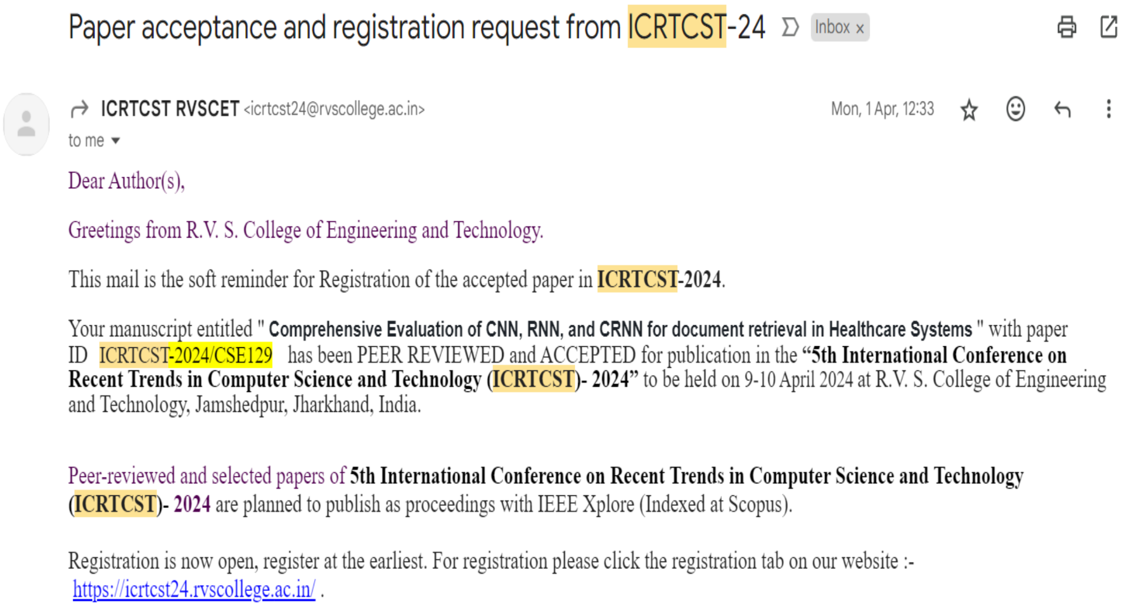
[63] R. DiPietro and G. D. Hager, “Deep learning: RNNS and LSTM,” Handbook of Medical Image Computing and Computer Assisted Intervention, pp. 503–519, 2020. doi:10.1016/b978-0-12-816176-0.00026-0.

[64] Nigam, V. (2021b, Januar 4). Natural language processing: From basics, to using RNN and LSTM. Medium. Analytics Vidhya. Abgerufen 20. Mai 2024, von https://medium.com/analytics-vidhya/natural-language-processing-from-basics-to-using-rnn-and-lstm-ef6779e4ae66.

[65] Alqahtani, F., Abotaleb, M., Kadi, A., Makarovskikh, T., Potoroko, I., Alakkari, K., & Badr, A. (2022). Hybrid deep learning algorithm for forecasting SARS-CoV-2 daily infections and death cases. Axioms, 11(11), 620.

LIST OF PUBLICATION(S)

1. Abhilasha Sharma, Vaibhav Kansal, "Comprehensive Evaluation of CNN, RNN, and CRNN for document retrieval in Healthcare Systems". The paper was Accepted at the 5th International Conference on Recent Trends in Computer Science and Technology **(ICRTCST 2024)** in April 2024. Indexed by **Scopus**. Paper ID: ICRTCST-2024/CSE129.



Paper acceptance and registration request from ICRTCST-24

ICRTCST RVSCET <icrtcst24@rvscollege.ac.in>

Mon, 1 Apr, 12:33

to me

Dear Author(s),

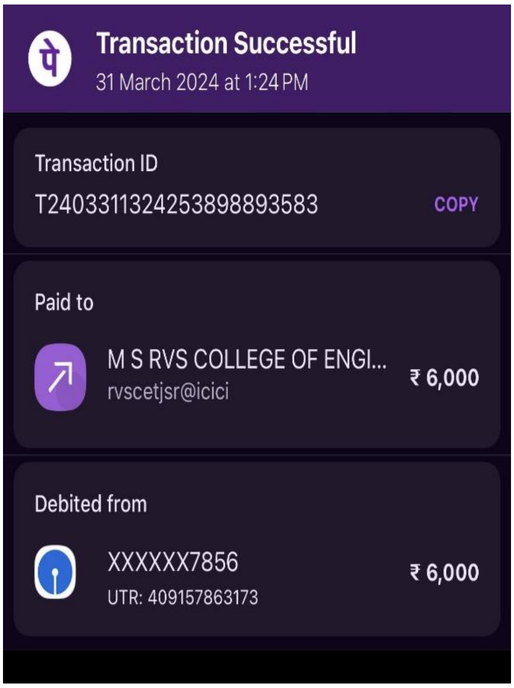
Greetings from R.V. S. College of Engineering and Technology.

This mail is the soft reminder for Registration of the accepted paper in **ICRTCST-2024**.

Your manuscript entitled " **Comprehensive Evaluation of CNN, RNN, and CRNN for document retrieval in Healthcare Systems** " with paper ID ICRTCST-2024/CSE129 has been PEER REVIEWED and ACCEPTED for publication in the **"5th International Conference on Recent Trends in Computer Science and Technology (ICRTCST)- 2024"** to be held on 9-10 April 2024 at R.V. S. College of Engineering and Technology, Jamshedpur, Jharkhand, India.

Peer-reviewed and selected papers of **5th International Conference on Recent Trends in Computer Science and Technology (ICRTCST)- 2024** are planned to publish as proceedings with IEEE Xplore (Indexed at Scopus).

Registration is now open, register at the earliest. For registration please click the registration tab on our website :- https://icrtcst24.rvscollege.ac.in/ .



Transaction Successful

31 March 2024 at 1:24 PM

Transaction ID

T2403311324253898893583

Paid to

M S RVS COLLEGE OF ENGI... ₹ 6,000

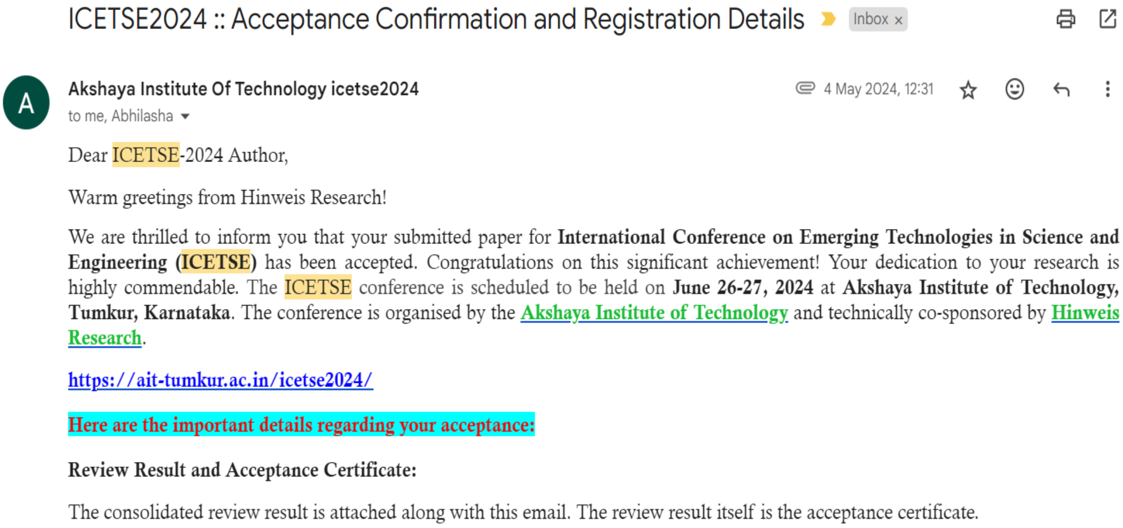
rvscetjsr@icici

Debited from

XXXXXX7856 ₹ 6,000

UTR: 409157863173

2. Abhilasha Sharma, Vaibhav Kansal, “A Survey On CNN, RNN, And CRNN For Document Retrieval In Healthcare”. The paper was Accepted at the International Conference on Emerging Technologies in Science and Engineering **(ICETSE 2024)** in May 2024. Indexed by **Scopus**. Paper ID: 356.



ICETSE2024 :: Acceptance Confirmation and Registration Details

Akshaya Institute Of Technology icetse2024

to me, Abhilasha

4 May 2024, 12:31

Dear ICETSE-2024 Author,

Warm greetings from Hinweis Research!

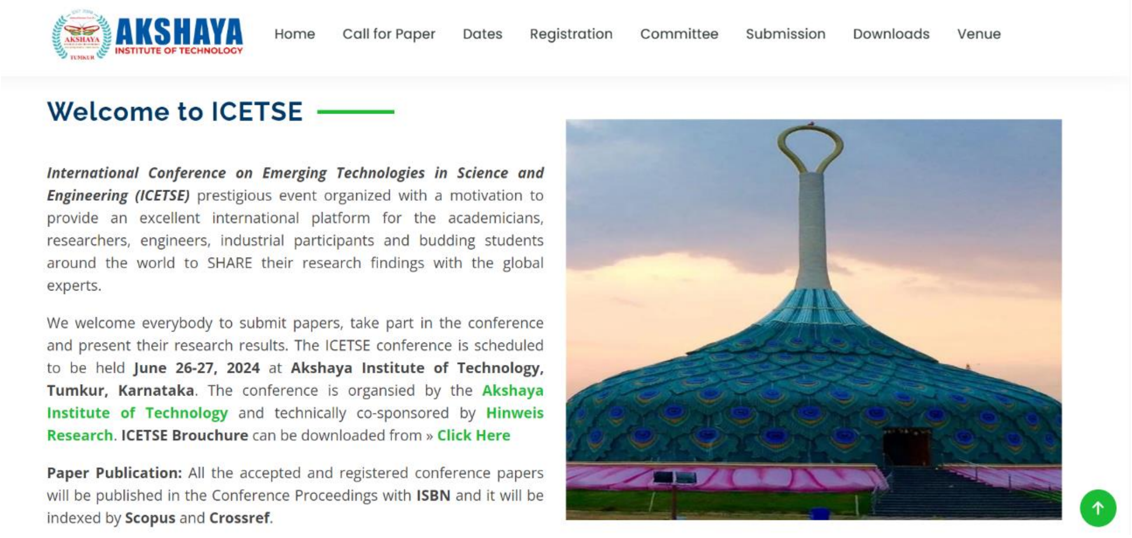
We are thrilled to inform you that your submitted paper for **International Conference on Emerging Technologies in Science and Engineering (ICETSE)** has been accepted. Congratulations on this significant achievement! Your dedication to your research is highly commendable. The ICETSE conference is scheduled to be held on **June 26-27, 2024** at **Akshaya Institute of Technology, Tumkur, Karnataka**. The conference is organised by the Akshaya Institute of Technology and technically co-sponsored by Hinweis Research.

https://ait-tumkur.ac.in/icetse2024/

Here are the important details regarding your acceptance:

Review Result and Acceptance Certificate:

The consolidated review result is attached along with this email. The review result itself is the acceptance certificate.

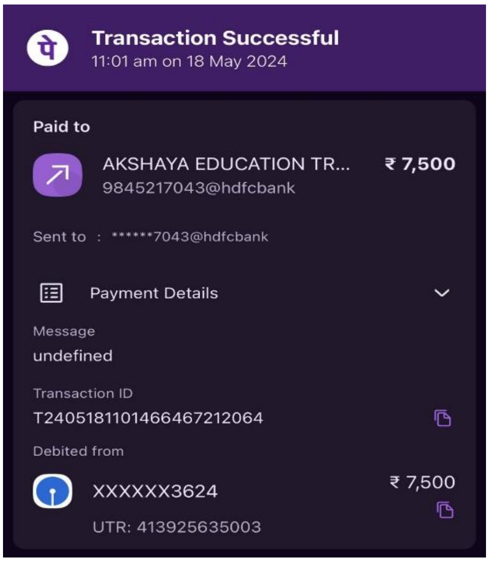


Welcome to ICETSE

International Conference on Emerging Technologies in Science and Engineering (ICETSE) prestigious event organized with a motivation to provide an excellent international platform for the academicians, researchers, engineers, industrial participants and budding students around the world to SHARE their research findings with the global experts.

We welcome everybody to submit papers, take part in the conference and present their research results. The ICETSE conference is scheduled to be held **June 26-27, 2024** at **Akshaya Institute of Technology, Tumkur, Karnataka**. The conference is organsied by the **Akshaya Institute of Technology** and technically co-sponsored by **Hinweis Research**. **ICETSE Brouchure** can be downloaded from » **Click Here**

Paper Publication: All the accepted and registered conference papers will be published in the Conference Proceedings with **ISBN** and it will be indexed by **Scopus** and **Crossref**.



Transaction Successful

11:01 am on 18 May 2024

Paid to

AKSHAYA EDUCATION TR... ₹ 7,500

9845217043@hdfcbank

Sent to : ******7043@hdfcbank

Payment Details

Message

undefined

Transaction ID

T2405181101466467212064

Debited from

XXXXXX3624 ₹ 7,500

UTR: 413925635003