Republic of Iraq
Ministry of Higher Education and Scientific Research
University of Anbar
College of Computer Science and Information Technology
Department of Computer Science

OCR Improvement for Unstructured Big Data Integration

A Thesis Submitted to the Department of Computer Science, College of Computer Science and Information Technology University of Anbar as a Partial Fulfillment of the Requirements for the Degree of Master of Science in Computer Science

By

Dalia Amir Abd Al Lattif

Supervised By

Prof. Dr. Murtadha M. Hamad

1442A.H. ***202 1 A.D.***

بسم الله الرحمن الرحيم

قالوا سُبْحَانَكَ لَا عِلمَ لنَا إلّا مَا عَلَّمْتَنَا
إنَّكَ أنتَ العَلِيمُ الحَكِيمُ
صدق الله العظيم
سورة البقرة - الآية 32

Supervisor Certification

*I certify that I read this thesis entitled **"OCR Improvement for Unstructured Big Data Integration"** that was done under my supervision at Department of Computer Science Information technology University of Anbar , by **"Dalia Amir Abd Al Latttif"** and that, in my opinion, it meet the standards of a thesis for the degree of Master of Science in Computer Science.*

Signature:

*Name: Prof .Dr .**Murtadha M. Hammed***

Date:7/9/2020

Certificate

*I certify that I have read this thesis entitled **“OCR Improvement for Unstructured Big Data Integration”** and I found it linguistically adequate.*

Signature:

Name:

(Linguist Authority)

Date: / /

Examination Committee Certification

We certify that we have read this thesis ***“OCR Improvement for Unstructured Big Data Integration” ,*** *and as an Examination Committee the student* ***“Dalia Amir Abd Al Latttif ”*** *in its contents and that in our opinion it is adequate to fulfill the requirements for the degree of* ***Master of Computer Science.***

Signature:

Name: (Chairman)

Date: / /

Signature: (Member)

Name:

Date: / /

Signature: (Member)

Name:

Date: //

Signature:

Name:

Signature: (Member)

Name:

Data: / /

Dedication

I would like to dedicate this work to :

The great teacher , Prophet ……..

(peace be upon him)

Soul of my father and soul of my mother

My husband……

My brothers and sisters…..

All my friends and lovers…….

The researcher

Dalia Amir Abd Al Latttif

Acknowledgment

First of all, I am grateful to the Almighty Allah, the owner of grace and favor who was and is still my support in preparing and achieving this thesis, and prayer and peace is upon his messenger Mohammed, and his own pure relatives and best companions. This research work reported in this thesis would not have been possible without the generous help of many persons, to whom I am grateful and wish to express my gratitude for them.

Praise is to Allah, who illuminated me the way of science and gave me patience to continue.

Also thanks, to all my beloved family members whose gave me the care and follow-up and guidance and tenderness throughout the period of study.

I wish to express my thanks and deep gratitude to my supervisor **Dr. Murtadha Mohamed Hamad** *for his invaluable advices, criticisms, encouragements, help, and supervision throughout this work to be in the best manner.*

I also wish to express my thanks and appreciation to the Dean of the College of Science ***Assist. Prof. Dr. Salah Awad Salman*** *,and the presidency of the Anbar University*

Finally, I wish to express my thanks to the head of the department of Computer Science and information technology ***Assist. Prof. Dr.Wissam Mohammed al-Rawi*** *, postgraduate Rapporteur* ***Dr. Ruqayah Rabeea Al-Dahhan ,****and all staff.*

Dalia Amir

ABSTRACT

The continues increasing in data that produced from different online systems and applications, has led to a fundamental problem related to how can managing and handling large volume of data. However, the most important point is the unstructured data storage method as it represents most of the data via internet management using the traditional methods is not suitable due to the availability of large and complex data. Hence, Hadoop was the suitable solution for the continuous increasing in data volumes and complexity, as well as dealing with and analyzing it as it is from any source, speed, size or quantity.

In this thesis, a system for analyzing big data is proposed. This system has the ability to identify repeated words (keywords) in a large number of image –based files(pdf) and files based text that have been scanned by (optical character recognition device or the so-called OCR). The system supports decision-making and developing the archiving process by providing an important entity to respond to keyword-based inquiries.

The goal of the project is to work on the (brain) of the (OCR) device using artificial intelligence (AI) techniques and mining techniques by making use of its ability to scan, read, analyze and convert texts and paper images into ASCII codes while at the same time solve its problems of inability to identify words one by one but rather reading full texts in addition to his inability to convert unstructured data into structured, and thus developing its capabilities to facilitate use of it in the business environment (business intelligence) and making his work close to the work of (Hadoop). This is done by making the device capable of identifying the required words by

adding rectangles around each word and giving it the ability to convert unstructured data into structured using the LSTM algorithm.

List of Abbreviations

Abbreviation	Meaning
ASCII	American Standard Code for Information Interchange
ANN	Artificial Neural Network
BD	Big Data
BDA	Big Data Analysis
BI	Business Intelligent
CED	Canny Edge Detector
DM	Decision Making
DFS	Distributed File System
DI	Data Integrity
HDFS	Hadoop Distributed File System
HLT	Hough Line Transform
HTML	Hyper Text Markup Language
IE	Information Extraction
IF	Intermediate Form
IR	Information Retrieval
IT	Information Technology
IoT	Internet of Things
KDD	Knowledge Discovery in Database
LDA	Levenshtein Distance Algorithm
LSTM	Long Short Term Memory

MBR	Minimum-area Bounding Rectangle
NLP	Natural Language Processing
OCR	Optical Character Recognition
OLAP	On Line Analytical Processing
PDF	Portable Document Format
PTM	Pattern Taxonomy Model
RNN	Recurrent Neural Network
RDBMS	Relational Database Management System
SQL	Structured Query Language
SVM	Support Vector Machine
SVR	Support Vector Regression
(Supa):	Absolute Support
(Supr):	Relative Support
SE	Structured Elements

Contents

Chapter One

General Introduction

Chapter Two

Theoretical Background

Chapter Three

Proposed Method Design

Chapter Four

Experiments and Results

Chapter Five

Conclusion and Future Works

List of Figures

List of Tables

Chapter One

General Introduction

Chapter One

General Introduction

1.1 Introduction

The term of Big Data (BD) refers to a very large datasets which its size growing continuously and in high speed. Therefore, it is very difficult to handle using any of software tools and database management which are available for normal structured data handling. These types of data have many categories and one of these categories is variety which include a textual contents i.e. structured, semi-structured, and unstructured .The contents of multimedia considered as big data which includes images , videos , and audio that are be on a multiple platforms (i.e. social media sites , Internet of Things [IoT], sensors networks, and cyber-physical systems). Dobre and Xhafa (2014) report that the world produces about 2.5 quintillion bytes of data (i.e. 1 exabyte equal 1 quintillion bytes or 1billion gigabytes) at every day [1]. Around 90% of these data being unstructured, Gants and Reinsel(2012) assert that by 2020, over 40 trillion gigabytes (Zettabytes) of data will be generated ,consumed, and imitated{Gantz, 2012 #437}[2].

These huge amount of heterogeneous, complex , and valued data that generated from (any-where, any-device and at any- time), require strong analytic process and use a new analysis tools to analyze and extract potential insights thereby enhancing the process of Decision-Making (DM). Therefore the area of Big Data Analysis (BDA) has become a trending and required process increasingly adopted in many organizations to obtain these valuable information. As a result, it will be useful in providing a future predictions that support a decision-making process [3].

Storing these large number of files by using the same web address at the same server considers as a common big challenge facing all the organizations and enterprises. Hence, the Distributed File System (DFS) concept has been appeared as a very useful solution to deal with big problems by using a distributed file system to save such files. That solution is found by Google in (2002) by designing a mechanism or framework that have the ability to store and perform analysis algorithms in automatic and distributed way at a same time , therefore Google present a paper called it (Google MapReduce algorithms) which encourage yahoo to implements it. The project of Apache Hadoop saves files of terabyte and petabyte by using distributed file system as well as do analysis on it [4].

In this work, introduction an explanation about different Data Analytic tools and techniques such as Data Mining , Machine learning (ML) and OLAP will be presented in order to use these analysis considerations to provide demand -driven aggregation as well as achieve Big Data Integration . Although that the Big Data has the ability to provide insights at scale and leverage to the machine intelligence, but it suffer from context loss because it unstructured data. Although that the big data has the ability to provide insights at scale and leverage to the machine intelligence, but it suffer from context loss because it unstructured data. The collected big data considers useless for the organizations development because the difficulty of conform it to an existing data model like (structured data or even semi-structured data). Therefore we must rescue it from context loss and make it usable through convert it to structured data and apply the integration considerations by selecting an affected features and use these features to create a model by using Hadoop because it has the ability to bear the workload and in particular the queries that are difficult to deal with it by using the traditional available systems and make a simulation for Hadoop to do all that in big speed by using a supervised learning for training it to make a desired behavior after some training, Machine learning models can be used to automatic quickly move through categorize unstructured data and finally validate the results.

1.2 Differences Between Big Data and Traditional Data

There are a set of analytics tools which can be used to compare the emerging big data with traditional data. An example, in the February 2012 the reports of Facebook shows that its users daily make 2.7 billion comments and ‘‘like’’[5]. This comparison is summarized in table1.1.

Table (1.1).A Comparison Between Big Data and Traditional Data [5]

	Traditional-Data	**Big Data**
Volume	GB	Updated constantly (TB or PB-currently)
Rate of Generation	Slow	High Speed
Structure	Structured	Semi-structured or Unstructured
Data Source	Centralized	Fully distributed
Data Integration	Easy	Difficult

Data Store	RDBMS	HDFS, NoSQL
Access	Interactive	Batch or near real-time

1.3The Development of Data Analysis for Decision Making

Table 1.2 shows the evaluation of analyzing data for the purpose of decision making through the last 45 years and analytics and until reaching big data concept.

Table (1.2)Data Analysis Development [6].

Naming	Date	Key Points
Decision Support	1970 to 1985	It analysis only structured data in order to support the decision making .
Executive Support	1980 to 1990	The data analysis by big management to take the required actions.
Online Analytical Processing(OLAP).	1990 to 2000	A multidimensional data tables analysing application.
Business Intelligence (BI).	1990 to 2005	These applications are used to backing data driven decisions, with focus on reporting.
Analytic	2005 to 2010	The statistical evaluation and mathematical modelling are used for determination making.

1.4 Literature Review

A number of researchers who apply algorithms and models to achieve their assigned objectives, the research is summarized them, as below:

Table (1.3) Literature Review

No.	**Study Name**	**Author**	**Date**	**The Study Result**	**The Study drawbacks**
1	An Ontological Approach to Knowledge Transformation in Malay Unstructured Documents	Sidi et al.[7]	2018	The work showed a significant improvement and increases the understanding of Malay Unstructured Documents	The difficulty and variety of unstructured Malay documents have led to the define them as structured interrogative .

2	Using natural language processing to extract structured epilepsy data from unstructured clinic letters: development and validation of the ExECT (extraction of epilepsy clinical text) system.	Shadrach et.al. [8]	2019	The study result in proposed an algorithm called (Bio- YODIE) which linking pipeline with a traditional manual review of the clinician epilepsy heterogeneous data.	The big obstacle that faced the work is the limited number of messages that system was applied to, which reducing the validating chance.
3	Large Scale Product Categorization using Structured and Unstructured Attributes	Krishnan and Amarthalu ri et. al.[9]	2019	This proposed an idea that using the values of structured attributes in an unstructured fashion, so that the order of different features and attributes by product categories is very easy.	It is difficult to take into account all the product properties for modeling examination processes.

4	Implementation of Optical Character Recognition using Tesseract with the Javanese Script Target in Android Application	Abdul Robby et. al. [10]	2019	The implementation of the model achieved a high accuracy rate (97,50%) . This accuracy has been achieved by combining both a (single boundary box) for the whole parts of the character and the (separate boundary boxes) in main body .	Achieving less accuracy for groups of letters" in Javanese language, which leads to distinguishing them as groups of letters "I".
5	Computational modelling of an optical character recognition system for Yorùbá printed text images.	Olalekan Joseph ONI et. al. [11]	2020	The results show an CER reduction where the recorded error rate of The Times New Roman font was (1.182%) which achieve a better performance of the other font styles.	All font types showed less accuracy than Times New Roman font.

6	Handwritten Character Recognition from Images	Mayur Bhargab Bora [12]	2020	The result of work shown there was using CNN-ECOC conclude that network the AlexNet is the most suitable CNN among the implemented network that are implemented for combining with ECOC to handwritten characters recognition.	The LeNet architecture simulation needs to be modified by adding two additional layers in order to increasing the low resolution it has shown.

1.5 Problem Statement

The Big Data characteristics such as rapid continuous growth, the diversity of its forms and sources, imposed new challenges and requirements in terms of rapid analysis, extraction of valuable information, processing, ensures their integration and storage. Therefore, new models, formats of semi-structured and unstructured big data and mechanisms of storage and architectures have been created, used and integrated by the modern data warehouse. This thesis deals with the following problems:

1. Although that OCR is useful for reading whole pages of unstructured physical documents and converting it to long string of ASCII codes ,but it lack the ability to answer queries such as detect any element

inside the pages and converting the structuring text to structure one .

2. Getting an effective dealing with the physical documents that contains big data.
3. Discovering the useful patterns from digital text documents, and utilizing these recurrent patterns to develop the performance of the system.

1.6 Thesis Contribution

The main contribution of thesis is summarized as following:

- This work contributes to facilitate& development of the archiving process for documents and files, and thus contributes to accelerating work in institutions and developing DM &BI .
- Creating structured and accurate data from unstructured elements because this will increase data integrity in term of free from distortion and damage ,hence production increases in the business world through automation.
- The system support Decision Making (DM) by providing an important entity for response the queries that are based on keywords.

1.5 Thesis Goals

- Improving the OCR capabilities by giving it the ability to detect any text element in the pages by adding a bounding boxes around every word, and converting unstructured text to structured one , hence increasing query responding by automation.
- making a brain for OCR so that it can identify repeated words from thousands of image-files as well as ensure the integrity (free from distortion and damage) , and it will be a qualitative addition to support it and make it approximate and simulate the Hadoop framework work.

1.6 Thesis Structure

The following chapters consist of this thesis:

- **Chapter One: General Introduction ,** this chapter explains some concepts such as Big Data, Traditional data, Decision Making. It also introduced the Literature Review of the studies that dealt with big data and how to deal with it , and clarify the thesis problem statement , contributions , and goals.
- **Chapter Two**: **Theoretical Background,** this chapter gives the background and Data Analysis techniques , Big Data concepts, and text mining , N-gram and Levenshtein algorithm .
- **Chapter Three**: **The Proposed System**, this chapter describes the proposed pattern discovery system with its design and implementation.
- **Chapter Four**: **Experiential Results and Tests,** this chapter explains the results and evaluation that have been getting from the proposed system.
- **Chapter Five**: **Conclusions and Suggestions**, for future works .This chapter presents the conclusions of this work. Furthermore, it provides suggestions for future work.

Chapter Two

Theoretical Background

Chapter Two

Theoretical Background

2.1 Introduction to Big Data

The term Big Data means the datasets that characterized by 4V which refers to :Volume that growing continuously, Velocity which means vast speed , Varity in its forms (structure, semi structured ,and unstructured) and its sources (Social Media ,photos, PDFs, videos, business generated data , image based files and etc.) ,and finally Veracity (noise, abnormal). Unstructured data can be defined as a data unit in which the information has a concurrent representation nature and don't has predefined arranging or any numeric values[13]. which make it difficult to be perceived, acquired, managed and processed by traditional software tools such as DM techniques, DBMS tools, and traditional database systems such as RDBMS within a tolerable time to make it in structured format which is highly-organized and easy to process ,analyze, and store [14,15].

Nowadays, Business Intelligence (BI) and Big Data Analysis (BDA) are among the most important terms associated with each other for dealing with big data especially in term of analysis . The big data life cycle include data inception, collection, transport through inter-networks, saved into distributed storage around the world that offers the best quality price with a reliable network as shown in figure 2.1[16].

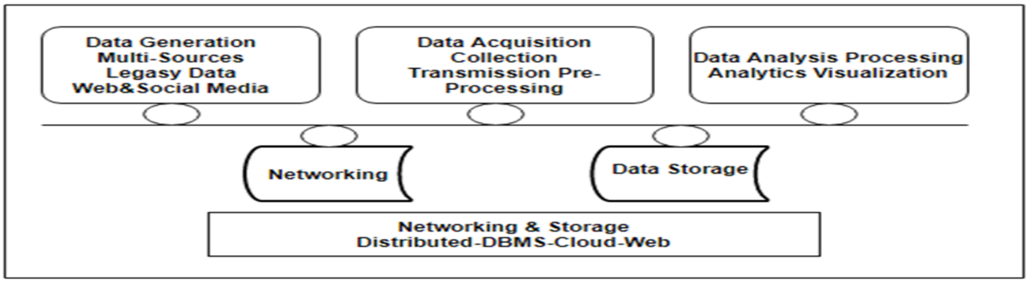


Figure (2.1) Big Data Life Cycle [16].

2.2 Big Data Integrity

Ensuring big data integrity consider as a big challenge .There are many terms that fall under the name of big data integrity such as security , validity (free from distortion ,altering, and damage), and availability (preserving it from loss), which can cause confusion .The thesis focused on big data integrity process that dealing with ensuring its validity in order to use it in improving archiving process [17].

2.3 Big Data Analytics

A process of discovering patterns from a large amounts of data in order to extract its value and correlations is a big data analytic. The data retrieval must pass the following stages [18]**:**

A. Data acquisition: Data obtained from the midst where data generation is growing at an exponentially rate unprocessed data. Selecting and discarding unneeded data can be quite challenging because the data that is being produced continuously mostly made up of unprocessed and in unstructured form .

B. Data extraction: There is an abundance of raw data that acquired and majority of them are not useful .Hence, deciding which data are needed to be kept and which one should be discarded is a difficult task to perform .

C. **Data collation**: There is a point at which the data can be analyzed more properly where data is retrieved from different sources and would be combined or superimposed so that a bigger and more detailed picture is formed .Therefore, utilizing data from one sample is inadequate to be used in analysis or prediction.

D. **Data structuring:** In order to be easier to retrieval of information it is important for the analyzed data to be organized in a structured form.

E. **Data visualization:** That is done by converting the information into visual format such as (map or graph) in order to make it easier to understand by the user.

F. **Data Interpretation :** In which a valuable information will be extracted. There are two types of information that can be acquired in this step: Retrospective involves gaining insights from the past events and actions .While Prospective Analysis is discovering trends for future based off the data that was recorded and distinguishing patterns.

There are two analytics approaches of big data as shown in figure(2.2). These approaches are: **Reactive** in which the knowledge is gained from the past which may have some usefulness or purpose in the future. The use of business intelligence in this approach ends up in generating standard business reports, ad hoc reports, OLAP, alerts and notifications.**Proactive** which means looking forward i.e. proactive big data analytics is required for proactive decision making such as optimization, text mining, modeling.

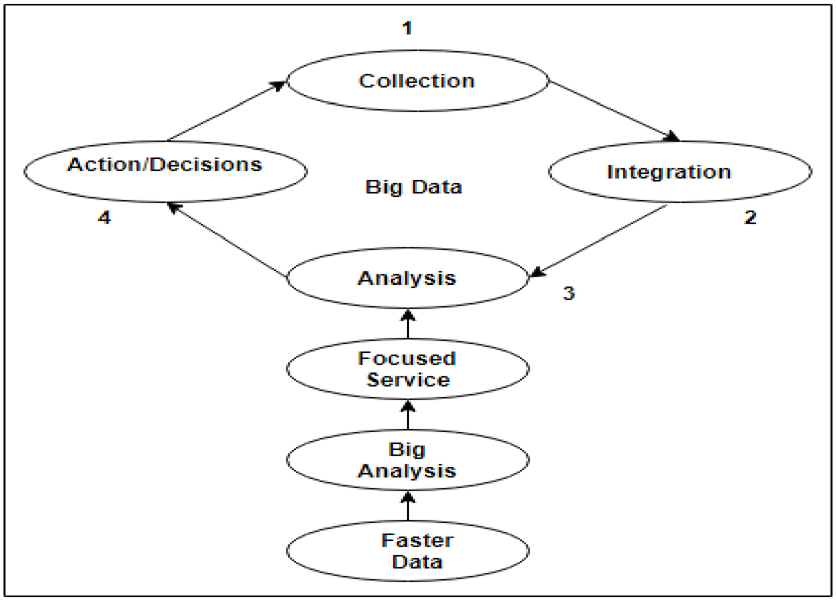


Figure (2.2) Big Data Stack [18].

2.4 Machine Learning and Big Data

As mentioned earlier the ability to extract value from big data depends on data analytics[19]. We can say that the analytics is the core of the Big Data revolution. Data analytics involves various approaches, technologies, and tools such as those from text analytics, business intelligence, data visualization, and statistical analysis , and add to it Machine Learning (ML) as a fundamental component of data analytics. ML will be one of the main drivers of the Big Data revolution that's what said by the McKinsey Global Institute. Because it has a great ability to learn from data and provide data driven insights, decisions, and predictions [20].

2.4.1Learning Categories

There are two main categories of learning tasks according to the nature of the available data [21]:

A - Supervised Learning

The learning will be supervised learning when both inputs and their desired outputs (labels) are known and the system learns to map inputs to outputs .The examples of supervised learning are classification and regression . In classification the outputs take discrete values (class labels) .Examples of classification algorithms are k-nearest neighbor, logistic regression, and Support Vector Machine (SVM) .In regression the outputs are continuous regression examples include Support Vector Regression (SVR), linear regression, and polynomial regression.

B- Unsupervised Learning

In which the system itself discovers the structure within the data when desired outputs are not known. Unsupervised learning includes clustering which groups objects based on established similarity criteria; k-means is an example of such algorithm. Some algorithms such as Neural Networks can be used for both, classification and regression. To attempt predict the future , predictive analytics develop models built using past data so depending on machine learning ; there are numerous algorithms can be used for this purpose including SVR, neural networks, and Naïve Bayes.

2.4.2 Deep Learning

An approach of the representation learning family of machine learning is Deep Learning. Representation learning is also often referred to as feature learning. The reason why this kind of algorithm is called Deep Learning is the fact that it uses data representations rather than explicit data features to perform tasks. Its job is to transforms data into abstract representations that enable the features to be learnt[22].

In the context of big data and due to the challenges associated with this process the ability to avoid feature engineering is regarded as a great advantage. Deep learning looks like neural networks where it is uses a hierarchical learning process to extract data representations from data that is mean it makes use of several hidden layers, and as the data pass through each layer, non-linear transformations are applied .These representations give high level of complexity to abstractions of the data. Each layer attempts to find the factors of variation within the data and separate it out the output of the last layer can be used as an input to other machine learning algorithms as well , because the output of the last layer is simply a transformation of the original input[23].

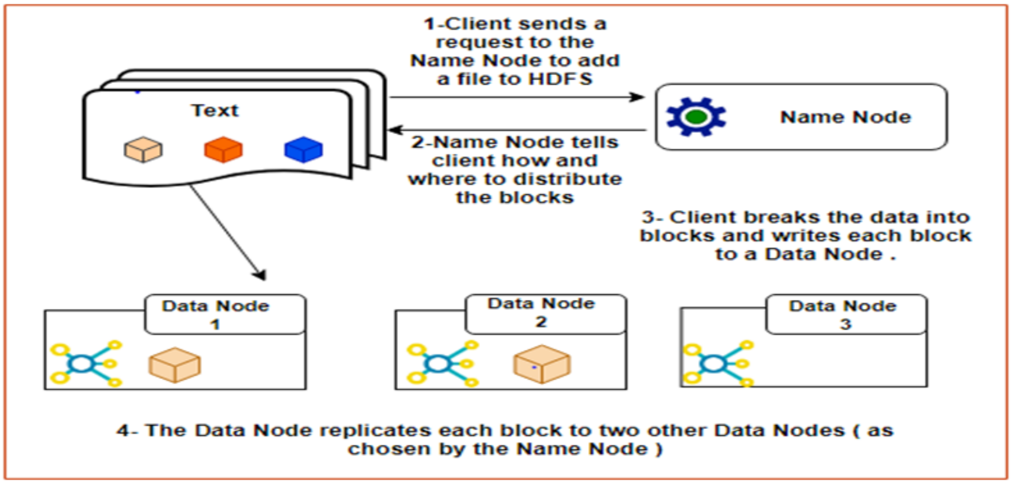
2.4.3 Deep Learning and Long Short-Term Memory (LSTM) Network

One of the most important classes of Artificial Neural Network (ANN) is Recurrent neural network (RNN) that has been developed through 1980s [24]. RNN is similar to traditional NN except that memory-state is add to the neurons, and its output back to itself number of times. Although that RNN consider as an repeating model of(NN) which uses many different connected loops which have the ability to give the present task an previous information, but it lack this ability in case of getting a gap between the output and the relevant information. In order to solve this problem a long short-term memory (LSTM) network has been designed as a type of RNN. LSTM shaped by adding some layers of neurons to the (RNN) chain model. It has the ability to get three decisions: removing the useless information, what the new information that can pass to the next layer, and finally decide the output of each layer [25]. That mean it has the ability to perform Deep

Learning by using training data, therefore it is broadly used in text data classification. The LSTM also used in (translation languages, image recognition, handwriting recognition, and translation languages) applications [26].

2.5 Hadoop and Big Data

Hadoop is an open-source cloud software designed to get rid of complexity the low performance of the traditional technologies when processing and analyzing big data .The basic reason that distinguishing them from the other traditional technologies is his ability to carry out tasks wherever data stored in contrary of the other traditional technologies which in order to perform calculation , they must first copying the whole distant data in memory . It is the most powerful dynamic analysis framework that deals with the problems of big data through the rapid processing , fault tolerance , query response, distribution, and other required operations with large storage capacity. Hadoop includes two main subcomponents that gave it all these capabilities and strengths to handle data storage and computation: Hadoop Distributed File System (HDFS) and Map Reduce framework. Hadoop architecture shown in figure (2.3)[27].


1-Client sends a request to the Name Node to add a file to HDFS
Text
Name Node
2-Name Node tells client how and where to distribute the blocks
3- Client breaks the data into blocks and writes each block to a Data Node .
Data Node 1
Data Node 2
Data Node 3
4- The Data Node replicates each block to two other Data Nodes (as chosen by the Name Node)

Figure(2.3) Hadoop Architecture[27].

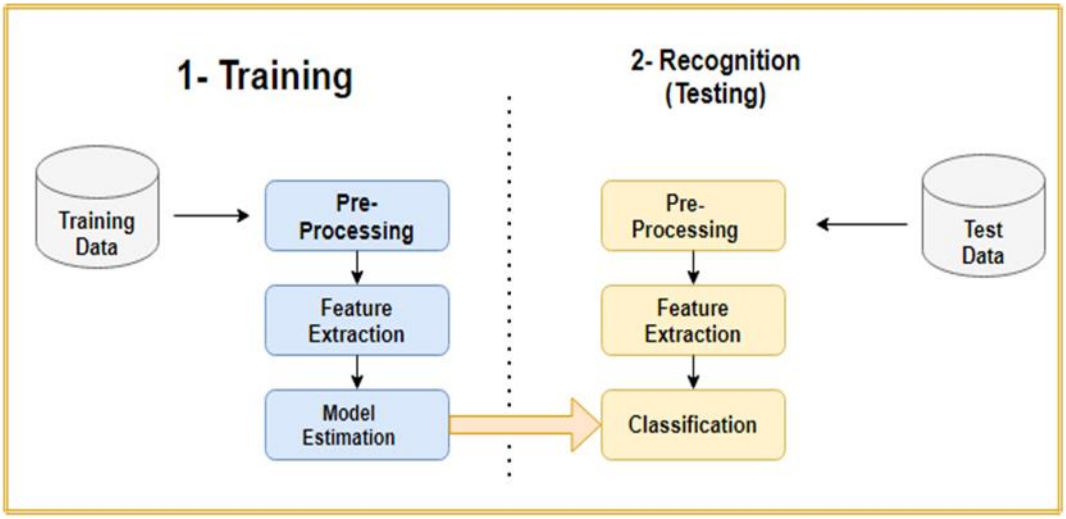
2.6 A Statistical Language Models (Language Processing)

The statistical language model is a probability distribution over sequences of words. The language model provides context to distinguish between words and phrases. The context refers to the objects or entities that surround the event and provides resources for its appropriate interpretation. The language models are used in (information retrieval) in the query likelihood model, where estimation the relative likelihood of different phrases is useful in many natural language processing applications. Language Modeling is used in many applications such as [28] :

1. Speech Recognition.
2. Parsing.
3. Handwriting.
4. Optical Character Recognition (OCR).

2.7OCR

OCR (optical character recognition) is the recognition of printed or written text characters (that are founded physically) by a computer. This involves photo-scanning of the text character-by-character, analysis of the scanned-in image, and then translation of the character image into character codes, such as ASCII, commonly used in data processing as shown in figure(3.3). In OCR processing, the scanned-in image or bitmap is analyzed for light and dark areas in order to identify each alphabetic letter or numeric digit. OCR is a type of software (program) that can automatically analyze the printed text and turn it into a form that computer can process more easily, it is a field of research in pattern recognition, Artificial Intelligence (AI), and computer vision . It is a common method of digitizing printed text , so that they can be electronically edited, searched, stored more compactly, displayed on-line and used in machine processing such as cognitive computing, machine translation, key Data, Text Mining [27].



Figure(2.4) ***Block Diagram of Character Recognition Process[27].***

Although that OCR is useful for reading the whole pages of unstructured elements , but it has some problems such as lack the ability to detect simple elements in the page and it also lacks in structuring the data that it read because it prints it in one long string. this thesis focused on unstructured text which is a more complex problem to solve.

2.8 Text Mining

Because the text is the generality normal form to store the information, it is believed that text mining represents a higher commercial probability than data mining. Actually, the recently existing studies referred that 80% of the company’s information is included in text documents. But, the task of text mining is more complicated than data mining since it deals with texts which are inherently fuzzy and unstructured [29]. The process of text mining works on discovering useful knowledge in text documents. The main challenge is to acquire accurate knowledge in text documents for helping employers to get their needs. Text mining is a fundamental stage in the knowledge discovery process in dataset, it has the whole techniques of the process of knowledge discovery and giving modeling step which is the application of techniques and algorithms for calculation of search pattern or models. So, it is very necessary to provide a proper model of text mining with relevant efficiency which is capable of retrieving the information that employers need[30].

Generally, text mining framework includes two distinct portions:

1-Text refining which transforms free form into an Intermediate Form (IF) text documents.

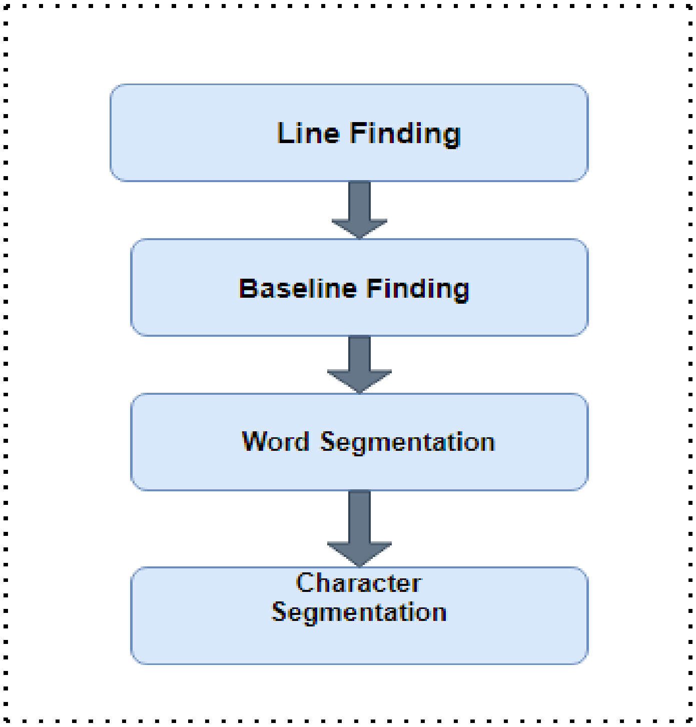
2-Knowledge distillation which produces knowledge or patterns from IF. There are two types of IF:

1- Structured IF like the relational data representation.

2- Semi-structured IF like the conceptual graph representation.

2.9Tesseract and Its Architecture

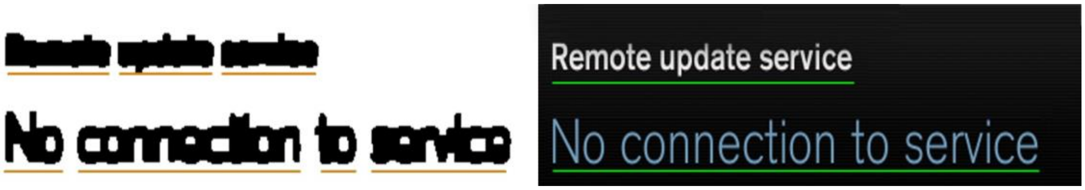
Tesseract is an OCR engine with open source code, it is the most popular qualitative OCR library . It is an analysis architecture that is built in an iterative pipeline process, except it revisits old steps. The recognition is done twice; during the first recognition run a static classifier is used and for an second the adaptive classifier is used. Tesseract is designed to recognize text even if it have a small skew without having to deskew the image, even though it is preferable to have the text horizontal for better recognition. The first part of the recognition process is the connected component analysis. It includes line finding, baseline fitting, and character and word segmentation. Then every word passed in the static classifier is fed the adaptive classifier for training. During the second recognition the adaptive classifier is used and words not previously recognized by the static classifier could now possible be recognized. To avoid lowering the output resolution, Tesseract needs to make sure that the image is appropriate ,so it perform image preprocessing ,this includes noise cancellation, refinishing and so on as shown in figure(2.5) [31].



Figure(2.5) Tesseract Component Architecture [32].

A. Line Finding

The key ingredient to baseline finding is blob filtering and line construction. A blob is a word, or a symbol or any content not connected to the rest of the image (Figure 2.6). The mean height of the blobs are approximated, helping the engine to filter out and remove all small blobs, typically punctuation or eventual noise[32].



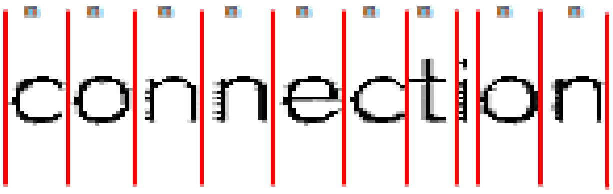
Figure(2.6) An Example of Finding Two Baselines From 7 Blobs[32].

B-Baseline Finding

When the lines of blobs have been found Tesseract examines them a little more closely. Baselines are fitted more precisely with a quadratic spline, i.e. four parallel lines that analyses the blob. This feature is very useful to help Tesseract handle curved words; e.g. scanned books where the words most often is curved in the center near the book bindings.

C-Word Segmentation

Segmentation means splitting an image into subsections for further processing . Segmentation include segmentation at the word level and segmentation at the letter level . In word segmentation the phrases with characters that sharing the same width (fixed pitch) are handled as a particular case, the phrases is sliced primarily based on the pitch and marked for recognition. However, most frequently do characters in phrases hold totally different pitches and need to be handled individually which will be seen in figure (2.7).

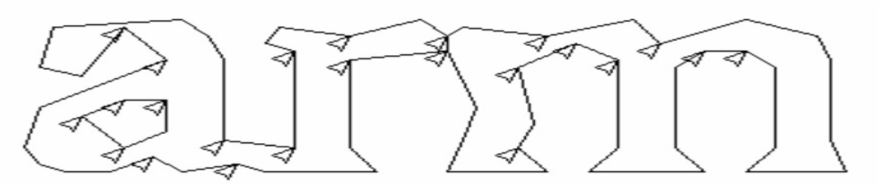


Figure(2.7) An Example of a Word Measurements[31].

Tesseract can dealing with different pitches by calculating the gaps in the limited vertical range between the both (baseline and mean line) [32].

D-Character segmentation

Tesseract tries to resolve the segmentation of characters by chopping the blob that was assigned the worst confidence by the character classifier. Potential candidate chop points are found from concave vertices of polygonal approximations of the outlines as shown in figure (2.8) .

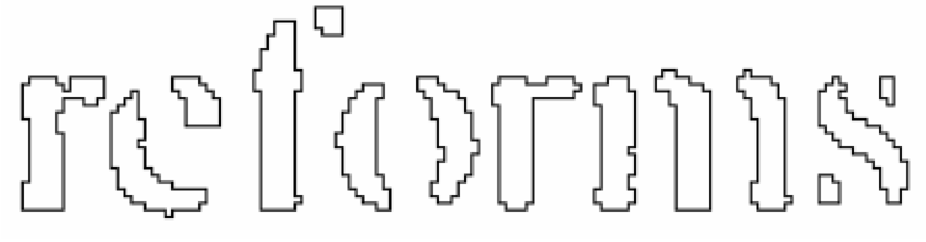


Figure(2.8) Candidate Chop Points In The Blob[31].

The chops are classified in a prioritized order, any chop failing to improve the confidence of the result will be undone, but not completely discarded. It will be analyzed once more by Associator. Identifying texts in images of printed handwritten documents is a difficult task due to the discrepancy between the background and the foreground. These degraded documents can be historical books, secret letters, or anything of value associated with it, so finding text information becomes the most important and difficult issue, especially if the text is deteriorating or overlapping. All of this requires dividing the text presented at the level of the word into letters, and this in itself is a great challenge due to the presence of letters connected with each other and cannot be separated easily or broken.Therefore there is need for associating broken character and character classification.

1. Associating broken characters

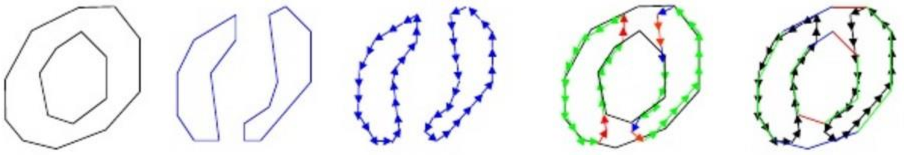
The associator try to filtering the candidate chops from a prioritized queue and evaluating them by classifying unclassified combos of fragments , an example of can be seen in Figure (2.9).



Figure(2.9) Word Recognition by The Associator Approach[31].

2.Character Classification

If all of the potential chops has been tried and the phrase continues to be not ok it's given to the associator. The associator also try to filtering the candidate chops from a prioritized queue and evaluating them by classifying unclassified mixtures of fragments as shown in figure(2.10) .



Fig(2.10) The Static Character Classifier Recognition of Complete and Broken Character [31].

2.9.1 Training

A great functionality of Tesseract possesses is the ability to learn new fonts and new languages. The training fundamentals of Tesseract consists of [28]:

- Character samples must be segregated by a font.
- Few samples are required; of each combination is good but one should be sufficient.
- Not many fonts are required.
- The number of different characters is limited only by memory.

The first bullet explains that the engine require a font to be trained. Thus, to be able to recognize symbols with Tesseract an conversion of vectorized symbols into a font is required. The process is done by the following steps:

1. Convert symbols into vectorized symbols.
2. Combine all vectorized symbols into a true type font.
3. Create a tif image with the symbol font exposing the different symbols.
4. Create a box file containing the data representation of each symbol from the tif image. The box file is essential for the training process to understand which symbol correspond to what data.

2.9.2 Hough Line Transform (HLT)

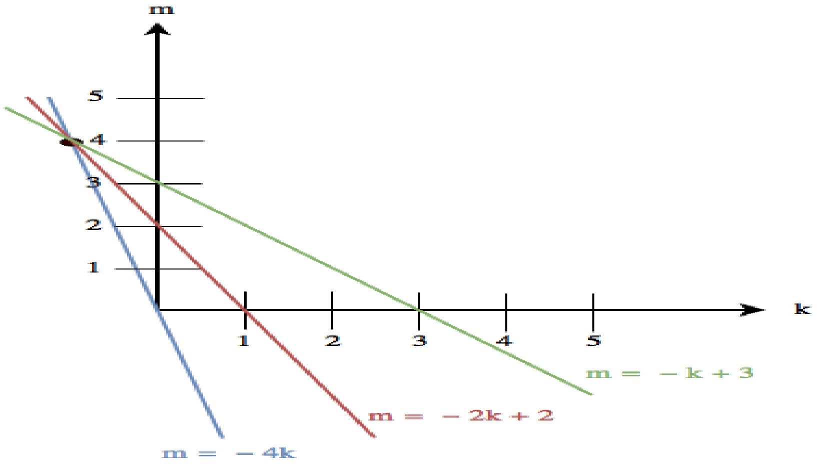
The theory of Hough Line Transform (HLT) is that any point in a binary image could be a part of a set of possible lines, together creating a distinct shape. Before applying the HLT the most common preprocessing stage is to apply an edge detector where each point is mapped in an image. The edge detector can miss points, or add disjoint points elsewhere from pixel noise making shapes “unconnected”. The HLT addresses this problem by grouping edge points into possible object candidates by a voting procedure. Hough Lines in the Cartesian coordinates can be expressed as[31]:

$y = kx + m$ (2.1)

where k is the gradient, m is the intersection point in y-axis. Consider a single isolated edge point (x, y) in the image plane, meaning there is an infinite number of lines that could pass through this point. Each of these lines can be characterized as the solution to some particular equation. Now let x and y be constants (i.e. a fixed point) instead and let m and k be the variables. One can think of it as a transformation from (x, y) − space to (k, m) − space. The equation 2.1 above can now be expressed as:

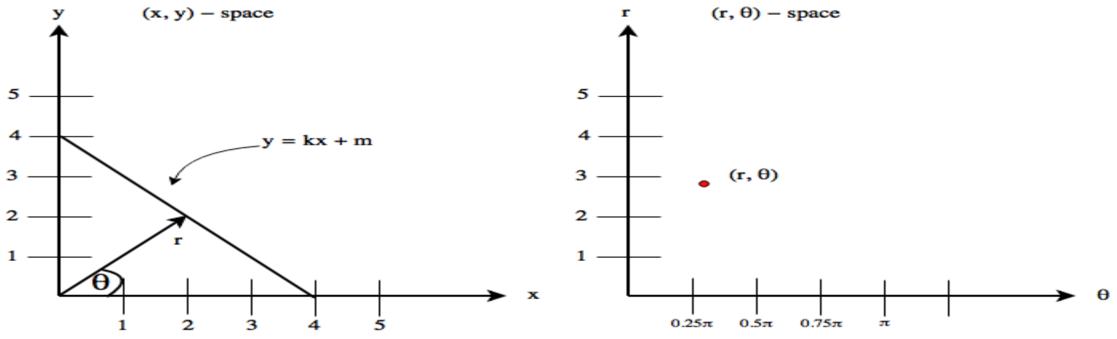
$m = -kx + y$ (2.2)

A visual example of how the equation is used is displayed in Figure (2.11).



Figure(2.11))*Three Lines Plotted In The* (*k, m*) *– space All Passing Through The Fixed Point(−1, 4) [31].*

A line the $(x, y) - space$ in Cartesian coordinate system can also be represented by a point in the $(r, \theta) - space$ from the Polar Coordinate system which can be seen in Figure (2.12).



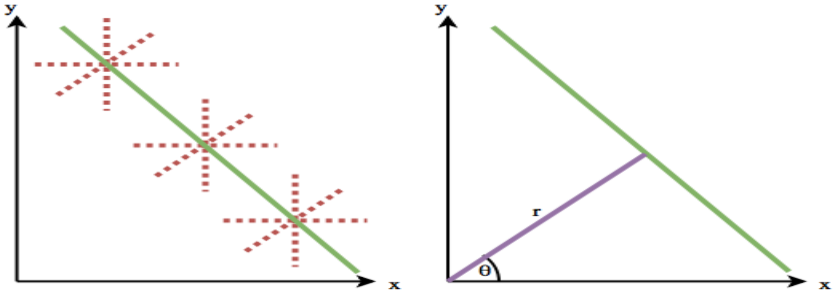
Fig(2.12) A Line In The (x, y) − Space Represented by a Point in The (r, θ) – Space[31].

The problem with Hough transform in *Cartesian coordinates* is that vertical lines cannot be represented as $k \longrightarrow \infty$. *Polar Coordinates* is used to solve this issue, where each line represents an angle θ and the perpendicular distance to the origin r. Where θ is bound by $[0, 2\pi]$ and r is bounded by the diagonal of the image.

However, another problem remains: for any single point, infinitely many lines can pass through it if the $(r, \theta) - space$ is continuous. To restrict the endless possibilities of different sets of angles, the angles is addressed in specific amount to be computed: $\theta = \theta 1, \theta 2, \ldots .\theta n$ the parameter can now be described as:

$$r_i = xcos\theta_i + ysin\theta_i \qquad (2.3)$$

Each (r, θ) point in the image compute the r, θ values and increase the values of the corresponding element of the accumulator matrix. Once the process is completed for every point, the elements in the accumulator matrix having the highest values will correspond to the lines in the image. A visual example can be seen in Figure (2.13).

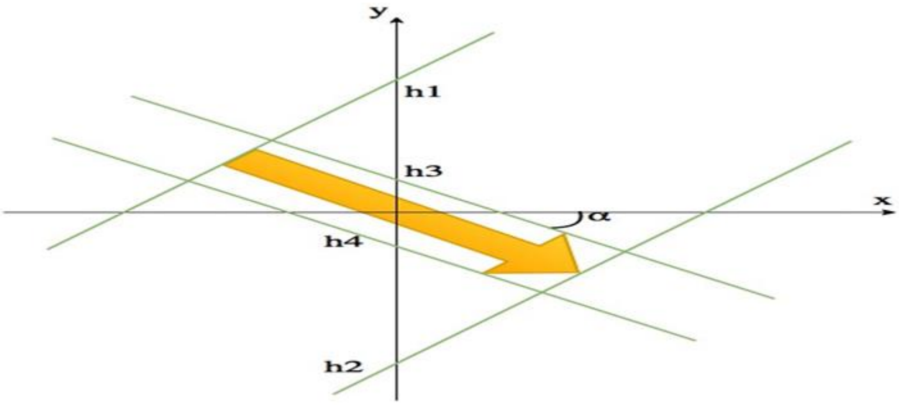


Figure(2.13) ***All Three Points Have One (r, θ) Pair In Common[31].***

This (r,θ) entry of the accumulator matrix is 3, remaining lines in the accumulator only have the value one is therefore ignore.

2.9.3 Minimum-area Bounding Rectangle (MBR)

The idea of the Minimum-area bounding rectangle(MBR) is to encapsulate every component from the image into a bounded area. If the bounded rectangle contains an angle in respect to the (x, y) − axis then the image most probably contain a skew. Figure(2.14) show an example of MBR.



Figure(2.14) Calculation of The Rotation Angle MBR Algorithm[31].

The MBR algorithm is defined by following steps [32]:

Algorithm (2.1) MBR Algorithm

Step1: Initialize the angle θ to zero, and the minimum area A_{min} to the current area of the bounding rectangle of the connected component at current angle θ.

Initialize $\alpha_{min,}$, α_{max} and $\Delta\alpha$ to their desired values. Consider the origin at the center of the rectangle

Step2: For different values of α, from α_{min} to α_{max} with a resolution of $\Delta\theta$ and α ǂ 0 repeat step 3 and 4.

Step3: The area A of the bounding rectangle in the α direction can be calculated by:

$$A=(h_1 - h_2) . (h_3 - h_4) ./\cos\theta . \sin\theta/$$

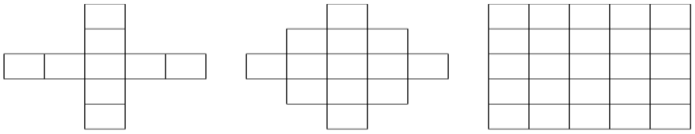
Step4:If $A< A_{min}$, set $A_{min} = A$ and $\alpha =\theta$

Step5: The rotation angle of the connected component equals to the skew angle α

Where h1 and h2 represents both maximum and minimum intercepts from origin of the lines having a slope of $\tan \alpha$, and passing through any boundary pixel of the connected component. h3 and h4 are the maximum and minimum intercepts from origin of the lines having a slope of $-1/\tan \alpha$ and passing through any boundary pixel of connected component. component.

2.9.4 Morphological Filtering

Morphology means “study of shape” in greek. In image processing it is used as a tool for extracting useful image components by a describing image shape. The binary images may contain noise or imperfections and the goal is to remove them by accounting a reference form. The technique utilizes a structuring element, a small shape. The element will be positioned on every pixel in the image and compared with neighboring-pixels, if it “fits” inside the neighborhood changes apply on the image or if it “hits”, meaning that the SE intersects with the neighborhood changes is rejected and nothing changes.The thesis implementation only encounters binary morphology. However, it can be applied to gray scale images as well as shown in figure (2.15)[31]



Figure(2.15) ***Different Structuring Elements, Cross, Diamond and Square[31].***

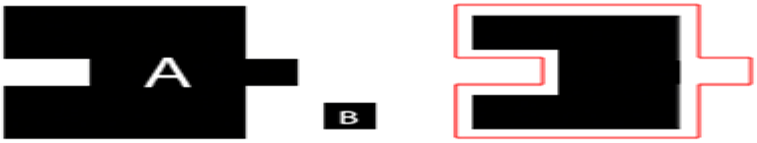
A- The fundamental operators: Erosion and Dilation

Erosion shrinks the contour in an image by stripping away pixels from the inner and outer boundaries. Holes between regions become larger, and small details gets imitated as shown in figure (2.16). The erosion of an binary image *A* by the structuring element *B* is defined [33] by:

$A \ominus B = \{z \in E | Bz \subseteq A\}$ (2.5)

where *Bz* is the translation of B by vector z:

$Bz = \{b + z | b \in B\}, \forall z \in E$ (2.6)



Figure(2.16) ***Binary Erosion of an Image A by a Structuring Element B[31].***

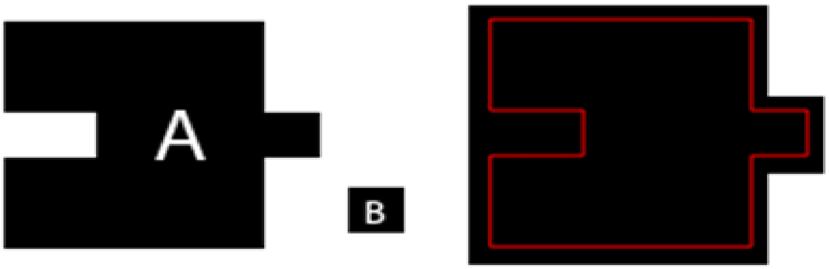
Dilation is the opposite to erosion, it increases the size of the contours by adding layers of pixels on the inner and outer boundaries. Holes between regions now become smaller, and small details is enlarged as shown in figure (2.17).

The Dilation of an binary image *A* by the structuring element *B* is defined [33] by:

$$A \oplus B\{z \in E|(Bs)z \cap A /= \emptyset \tag{2.7}$$

where B^S denotes the symmetric of B:

$$Bs = \{x \in E| - x \in B \tag{2.8}$$



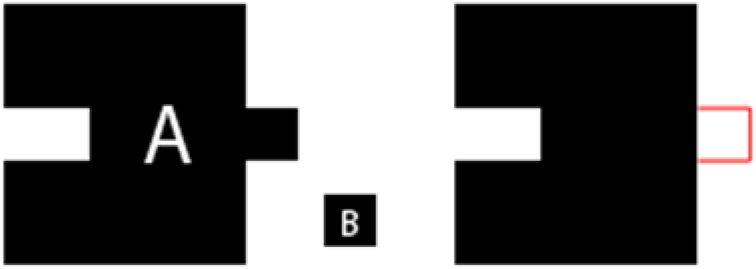
Figure(2.17) ***Binary Dilation of an Image A by a Structuring Element B[33].***

B- Opening

1. Opening generally smooths the contours, breaks narrow areas and eliminates protrusions. The parts that will remain in the image is where the SE hits properly, else when the element fits, content will be removed. The opening of A by B is obtained by erosion of A by B,

followed by dilation of the resulting image by B as shown in figure (2.18):Opening of an binary image A by the structuring element B is defined [33] by:

$$A \circ B = (A \ominus B) \oplus B \qquad (2.9)$$



Figure(2.18) Binary Opening of an Image A by a Structuring Element B [32].

A- Closing

Closing also tend to smooth the contours, but as opposed to opening, it generally fuses narrow breaks and long thin gulfs, eliminates small holes and fill gaps in the contour. The closing of the binary image A by the structuring element B is done by first dilate set A by B, then erode the result by B as shown in figure(2.19). Closing of an binary image *A* by the structuring element *B* is defined [33] by:

$$A \bullet B = (A \oplus B) \ominus B \qquad (2.10)$$

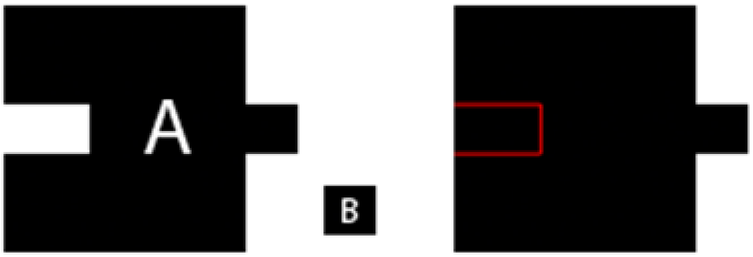


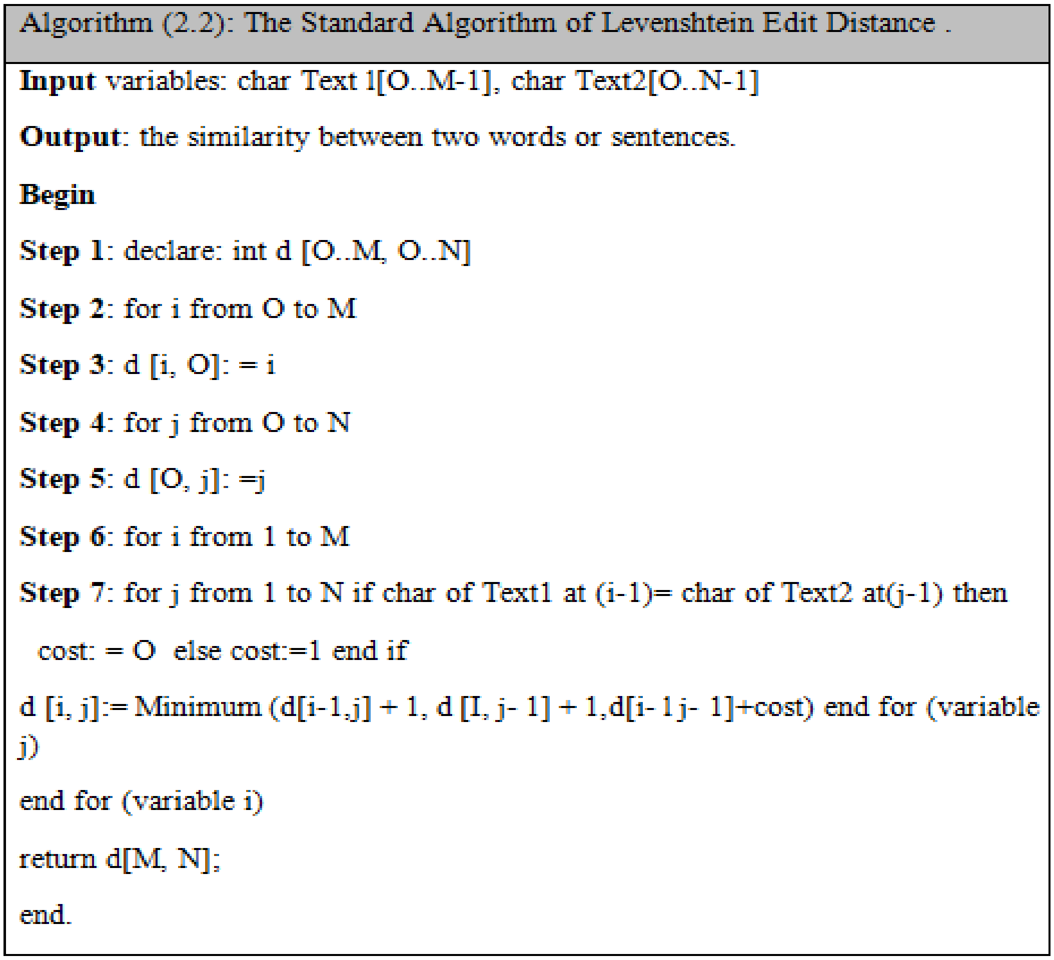
Figure (2.19) Binary Closing of an Image A by a Structuring Element B [32].

2.10 LSTM and OCR

The earliest design of OCR recognition engine was not has an algorithm of deep learning . Which causes losing of accuracy when it uses in the text detection area , thus it leads makes the OCR recognition rate to be not ideal. As mentioned earlier that the LSTM consider as an important method of machine learning methods leads to usability of LSTM with OCR to increasing the text detection and text lines accuracy [34].

2.11 Edit Distance Algorithm (Levenshtein Distance)

The Edit Distance Algorithm (also known as Levenshtein Distance) is dictionary-based algorithm, which detect the similarity strings and the most frequent terms in flexional suffixes model. Searching similar sequences of data is of great importance to many applications such as the gene similarity determination, speech recognition applications, database and/or Internet search engines, handwriting. Therefore, algorithms that can efficiently manipulate sequences of data (in terms of time and/or space) are highly desirable, even with modest approximation guarantees [33].Algorithm (2.2) shows the standard algorithm of Levenshtein edit distance.



```
Algorithm (2.2): The Standard Algorithm of Levenshtein Edit Distance .
Input variables: char Text 1[O..M-1], char Text2[O..N-1]
Output: the similarity between two words or sentences.
Begin
Step 1: declare: int d [O..M, O..N]
Step 2: for i from O to M
Step 3: d [i, O]: = i
Step 4: for j from O to N
Step 5: d [O, j]: =j
Step 6: for i from 1 to M
Step 7: for j from 1 to N if char of Text1 at (i-1)= char of Text2 at(j-1) then
  cost: = O  else cost:=1 end if
d [i, j]:= Minimum (d[i-1,j] + 1, d [I, j- 1] + 1,d[i- 1j- 1]+cost) end for (variable j)
end for (variable i)
return d[M, N];
end.
```

Where:

d -Levenshtein matrix of size N+1, M+1, formed for Text1 and Text2 terms.
M ,N - are the two terms lengths respectively.
d [i, j] – (i, j) -represents an element in Levenshtein matrix d. **min** – a function to compute the minimum of three variables. **cost -** a variable which obtains either 0 or 1 values. The distance *K* of Levenshtein is the minimum number of operations (deletion, insertion, and substitution) needed for changing the term into the another, as in the next equation:

$$K = d\,(M, N) \qquad (2.11)$$

Levenshtein Distance is a straightforward dynamic programming algorithm which addresses the issue of sequence matching primarily based on the apprehension of a primitive edit operation .It is clear that utilizing only the mentioned operations of primitive edit, it is permanently potential for transforming an initial string A into a target string B (The two strings have the same alphabet). The distance of Levenshtein for these two strings is the minimum number of single-character substitutions, deletions, and insertions needed for transforming A into B [31].Edit distance is the cost of unit operation which is needed for transforming a string to another one, where these two strings are becoming the same string. Unit operations can be categorized into four operations: deletion, insertion, transposition, and replacement. The deletions and insertions have the same costs, while the replacements have double of the insertion cost [33].

2.12 Similarity Measuring Among Terms

The measure of similarity P is the quotient of a number of Levenshtein operations (after calculation of LDA) by the number of all Levenshtein operations in pessimistic case. This means, before the calculations of LDA will be completed but with the maximum possible number of Levenshtein operations well known. The similarity measures *P* is calculated by the formula[31]:

$K \geq 0$, $M < 0$,$N < 0$, $P \epsilon (0,1)$

where:

K **max** – the length of the longest of analyzed two terms/text strings (i.e. pessimistic case

K **–** is equal to the length of the longest term. Table (2.2) illustrates an example of the Levenshtein distance and the measure of similarity between two sentences.

Table (2.1): An Example of The LED and The Similarity Measurements Between Two Sentences[30].

No	Sentence1	Sentence2	K	P
1	My car isn‟t working	My bicycle isn‟t working	1	0.7
2	What did you do	What have you done?	3	0.4
3	Tom is writing a letter	Tom is written letters	3	0.4

An N-gram is a sub-sequence of n-items in any given sequence. In models of computational linguistics, N-gram is utilized generally in predicting words (in word-level N-gram) or predicting characters (in character-level N-gram) for different applications. Most applications in NLP and IR include extracting sequences of successive words (generally referred to N-grams) from large text documents. Extracted N-grams from text data can be utilized for different purposes. In machine translation, for instance, N-gram has utilized for constructing a model of statistical language .The process of extracting N-grams from considerable text documents is a challenging issue especially if the word with longer sequences are considered. The number of distinct sequences increases as the data get bigger. With a large group of text documents, finding and quantifying the frequencies for every N- gram might need huge computational resources. Frequency of a term (TF) can be computed by using equation (2.13)[35]:

$$TF= Nk / N \qquad (2.13)$$

Where

Nk: is the ratio of the number of times a keyword K appears in a given document

N :is the total number of terms in that document.

N-gram have some features such as:

Absolute Support (Supa): absolute support can be computed by using equation(2.14):

supa(X)=|X| (2.14)

where

|X|: is the number of paragraphs in which the term appeared.

Relative Support (Supr): relative support is calculated by Implement the equation (**2.15**) :

Supr (x)= | x| / |PS(d) | (2.15)

Where:

|PS(d)|: is the number of paragraphs in the document.

Global Probability: it can be computed by using equation (2.16):

bi-grams: _T, TE, EX, XT, T_

tri-grams: _TE, TEX, EXT, XT_, T_ _

quad-grams: _TEX, TEXT, EXT_, XT_ _, T_ _ _

and in word level N-gram for example: bi-grams: San **Francisco**.

tri-grams: The Three Musketeers (is a 3-gram).

quad-grams: It stood up slowly (is a 4-gram) .

In this thesis a global probability feature was chosen because the designed statistical language model can consider as a probabilistic distribution over a sequence of words .

2.13 Skew Detection

Good recognition requires straight horizontal lines of words and symbols, increasing skew angle means an increasing recognition error rate. Thus, if a skew angle is detected it needs to be handled. The Hough Line Transform(HLT) and Minimum-Bounded Rectangle (MBR)algorithms are implemented to detect the eventual skew. Hough Line Transform (HLT); is applied on the image where natural edge lines can be found. Minimum-Bounded Rectangle (MBR); is used as a fail-safe option if HLT fails to find any lines[35].

Chapter Three

Proposed Method Design

Chapter Three

Proposed Method Design

3.1 Introduction

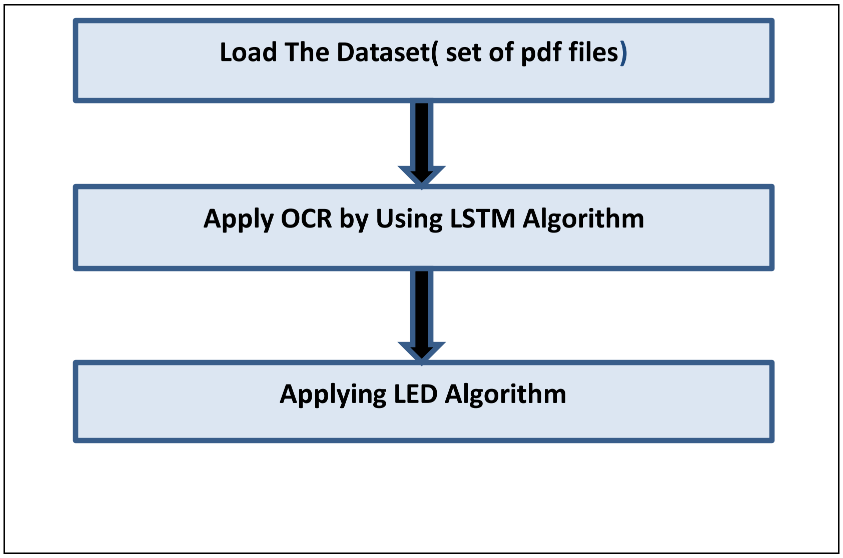
This chapter presents the design considerations, implementation requirements, and the steps taken for the establishment of an innovative solution for text mining by building a pattern discovery. The proposed system used various operations mainly relies on using deep learn and Hadoop techniques.

3.2General Steps of Proposed Model

The rapid growing of digital data in the form of text document has faced several problems such as (language type, errors, and noise) . Additionally, analyzing these data manually is a hard task. The documents include several important terms which pointing on the valuable information additionally is supporting with various words inside the document. It is an auto-process to identify a fixed amount of key phrase or word which better reflect the main document content. Generally, it means the process of discovering useful patterns, structures and other valuable information from unstructured natural language texts. The keywords detection refers to the process of documents summarization which assists other words inside the text for inaugurating the main content. The suggested system shows the major steps from collecting information to achieving results. Figure (3.1) demonstrates the suggested system's key stages. In this work several major steps will be illustrated as following:

- **Data Collection**: it will achieve by downloading a number of ‘pdf’ files from datasets real world documents for OCR testing large(N>10000) which contains a mix of document image pdf format in English from https://www.kagle.com/data/40647 .

- **Text Detection**: OCR will be utilized using Tesseract. The Tesseract has implemented a Long Short Term Memory (LSTM) as describe in paragraph (3.3).
- **Keyword Detection(Hadoop)** : it finds the most key word in the file of PDF formal by using Levenshtein distance as describe in paragraph (3.4).


Load The Dataset(set of pdf files)
Apply OCR by Using LSTM Algorithm
Applying LED Algorithm

Figure(3.1) The Proposed System.

3.3 The Segmentation ,Detection ,and Recognition the Text by OCR Implementation In Tesseract

OCR implementation in Tesseract includes many steps as shown in figure(3.2) . To start the program the initial function needs be called. The call requires three parameters to proceed: what language to be recognized, if a camera is used as an image source and finally a calibration image to calculate eventual skew and interface theme.

OCR implementation in Tesseract include six steps as following :

Algorithm (3.1) OCR implementation in Tesseract Algorithm

```
Step 1:Initial Process Include:
1.Start.
2.Enable Tesseact.
3.Calculate Structure Element Size, if camera used then
4.Calculat skew angle
End if.
Else
If the theme bright then
Set inverse color flag and wait for step2.
If image require rotate go to step2 and rotate image and inverse colors
Else go to step2 and perform Canny Edge Detection.
End if .
End if.
Step2:Function Call:
1.Start the initial process .
 2.Apply OCR.
3.Find word.
4.Findtext then go to step 3.
```

Step3:Image Pre-Processing:

1.Canny Edge Detection.

2.Morphology Filtering.

3.Finding Contours.

4.Add Bounding Rectangle.

5.Remove big and small rectangles.

6.Add rectangle position to the result list.

7.Segment all positions from the result list then go to step 4.

Step 4: Recognition :

1.Recognize all segments.

2.Add recognition and confidence to the result list then go to step 5.

Step 5: Text Detection:

Adding bounding box around each text in image.

Step6:Return result.

End.

3.3.1 Initialization Process

The initialization process starts by enabling Tesseract, followed by calculations of the Structuring Element (SE). The SE is calculated from the image resolution and will be used later for morphological filtering. If the camera is used a skew then the angle of the image will be calculated. If the theme of the interface is bright then a flag will be set to ensure that the colors will be inverted later. The reason for the flag is because the image processing in the implementation requires white content on black background. Once the initialization process is completed the program is put to a halt and will wait for a new function call.

3.3.2 Function Calls

There are four functions calls available in the implementation depending on what you want to achieve .The first function is to start the initialization process. The second is to apply OCR to scan an image searching all words . The next function is to find word and is used to verify if one specific word can be found. The fourth function is to find text as equal to find word, but the only difference by searching for several words instead of one only.

3.3.3 Image Pre-Processing

Image Pre-Processing is an important step in the OCR workflow for to remove the noise from the image , remove the noise background from the image , and handle the different lightning condition in the image. Once the program receives a function call it proceeds to check whether the image needs to be rotated or not, this can be achieved checking the previously calculated skew angle. By detecting the image, it will be rotated, otherwise it will continue. Next step is to check if the image needs to be inversed. This is done by checking the inverse color flag that was calculated. If it is flagged to be inverted then it is put through a invert color function, otherwise it will continue.

Now the image has arrived into the image processing block where the objective is to extract blobs of information and transform them into segments. The first function is the Canny Edge Detector(CED), where the essential edges will be found and background noise is removed. Next up is the morphological closing filter where the SE will be used. The closing algorithm will connect characters and symbols into blobs. These blobs are then used to extract contours of every blob. This allows for the next function: bounding rectangle. The function will bound all contours in the image by bounding rectangles. Each rectangle holds the position where each segment is located. A filtering function removes all rectangles which are significantly large or small. Segments containing a word or a symbol is extracted from the image with the help of the remaining rectangles and sent into Tesseract for recognition.

3.3.4 Words Recognition

The recognition result will be returned along with the confidence from Tesseract and represent the grade of the recognition. The result will be inserted into a result list containing information about all the recognized words: what word was recognized, the recognition confidence and position in the image. The result list will then be used to deliver the final answer. After the result list had been built, the implementation returns to the waiting state where it will wait for a new function call to restart the process.

3.3.5 Text Detection

It is a techniques that used for detect the text in the image and create and bounding box around the portion of the image having text as shown in figure (3.2).

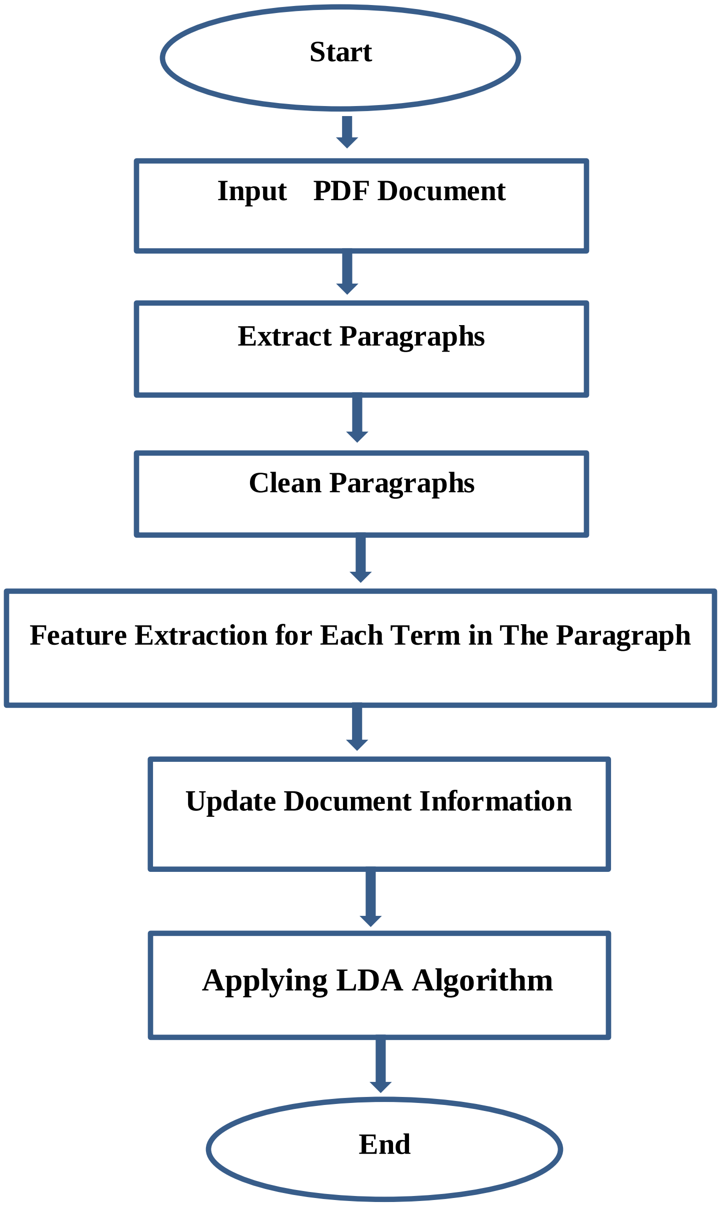
Figure (3.2) OCR Text Detection.

3.4The Keyword Detection

Generally, the proposed system includes a number of basic stages to perform all relevant process for effective frequent patterns extraction from unstructured data. The proposed system for unigram grammar will be explained in subsection (3.4.1) and bigram grammar will be explained in subsection (3.4.2).

3.4.1 Unigram Grammar:

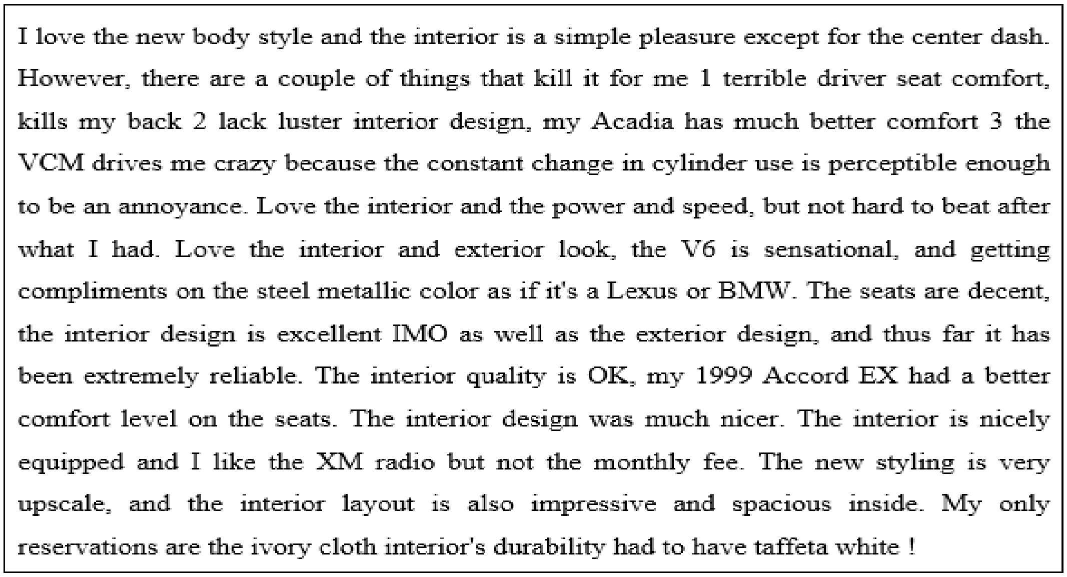
The proposed system of unigram grammar includes six steps including input text, extract paragraphs, clean paragraph, feature extraction, update document information, and apply Levenshtein algorithm. Figure (3.3) shows the proposed system flowchart of unigram grammar.



Figure(3.3) The Proposed System Flowchart of Unigram Grammar.

The description of each step at the figure (3.3) is given below:

1- **Input PDF Document Step**: this is the first step in the proposed system. The input is entered by the end-user. The raw substance for text mining is the text documents which are unstructured as these documents do not include pre-defined relations among phrases or words when storing them in the computer. Figure(3.2) gives an example for these documents.



I love the new body style and the interior is a simple pleasure except for the center dash. However, there are a couple of things that kill it for me 1 terrible driver seat comfort, kills my back 2 lack luster interior design, my Acadia has much better comfort 3 the VCM drives me crazy because the constant change in cylinder use is perceptible enough to be an annoyance. Love the interior and the power and speed, but not hard to beat after what I had. Love the interior and exterior look, the V6 is sensational, and getting compliments on the steel metallic color as if it's a Lexus or BMW. The seats are decent, the interior design is excellent IMO as well as the exterior design, and thus far it has been extremely reliable. The interior quality is OK, my 1999 Accord EX had a better comfort level on the seats. The interior design was much nicer. The interior is nicely equipped and I like the XM radio but not the monthly fee. The new styling is very upscale, and the interior layout is also impressive and spacious inside. My only reservations are the ivory cloth interior's durability had to have taffeta white !

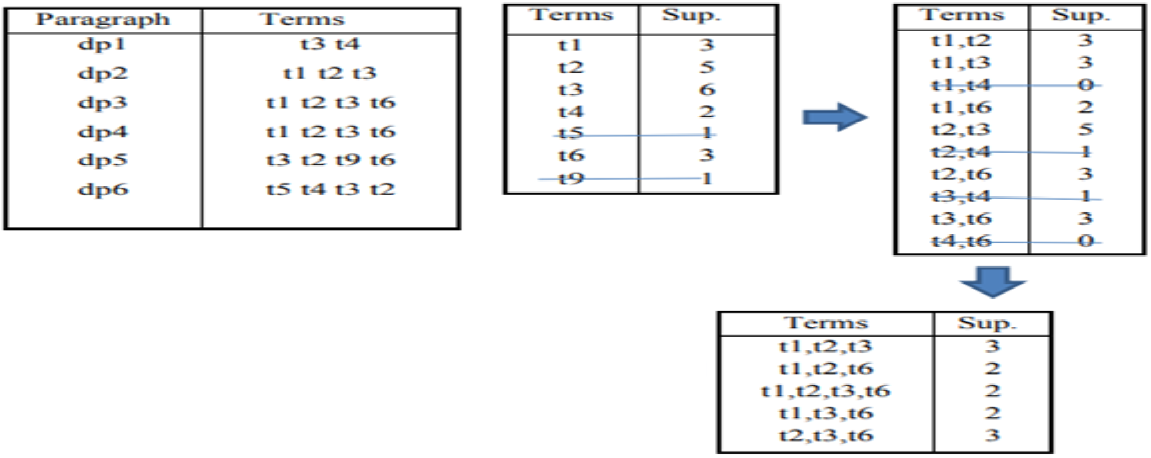
Figure (3.4) An Example of PDF Document.

2- Extract Paragraph Step: this is the second step in the proposed system, in order to efficiently utilize the discovered pattern, this step is related to Pattern Taxonomy Model (PTM). This model works on re- evaluating the patterns measures via deploying them into a common hypothesis space depending on their correlations in the taxonomies of the pattern. This results in high specificity patterns to the subject which can become adequate and reasonable important values leads to an important development in the system efficiency.

PTM method firstly works on scanning the uploaded document and converting the entire document into a set of paragraphs which are used as separate documents. Afterward, a process of extracting sets of terms from the obtained documents and terms form a specific pattern is achieved. A set of paragraphs are shown in Table (3.1), to the specified document "d", here

SP(d) = {dp1, …, dp6} the whole redundant words are removed. Considering that min-sup ≥ 2. The '5' frequent patterns .

Table (3.1): Paragraphs Set and Frequent Patterns.



Paragraph	Terms
dp1	t3 t4
dp2	t1 t2 t3
dp3	t1 t2 t3 t6
dp4	t1 t2 t3 t6
dp5	t3 t2 t9 t6
dp6	t5 t4 t3 t2

Terms	Sup.
t1	3
t2	5
t3	6
t4	2
~~t5~~	~~1~~
t6	3
~~t9~~	~~1~~

Terms	Sup.
t1,t2	3
t1,t3	3
~~t1,t4~~	~~0~~
t1,t6	2
t2,t3	5
~~t2,t4~~	~~1~~
t2,t6	3
~~t3,t4~~	~~1~~
t3,t6	3
~~t4,t6~~	~~0~~

Terms	Sup.
t1,t2,t3	3
t1,t2,t6	2
t1,t2,t3,t6	2
t1,t3,t6	2
t2,t3,t6	3

Where

Min-sup: it means that any frequency of a term less than two is canceled.

it means a term in the paragraph.

dpj: is the paragraph of a document " d".

-There are several terms used for Pattern Taxonomy Model in the proposed system such as:

- **Term Frequency (TF):** it is one of the main techniques for keyword extracting in which the word existence in the document is counting, for example, when TF of the word (Text) is equal to "8", this means that the term (Text) appeared "8" times in a document. Generally, if the TF is high which means that it appear more than two times ≥2, then the word is a significant one.
- **Term Supporting**: supposed a term set "X" in the "d" document,” X” is utilized for denoting the covering set of "X" to "d" document, that consists of all “db” paragraphs ∈ PS(d), X⊆ dp, this means;

X ={ dp/dp∈ ps(d), X⊆ dp}.

- **Threshold**: threshold represents a boundary between the important and non-important terms. The threshold is used to reducing the number of discovered patterns in a larger document. These discovered patterns of minimum relative support will maximize the training burden. In this thesis, threshold values between 1 to 10 are used in the tests of the proposed system, because the higher the value of the threshold as we get closer to the words of the document as a whole and this is not useful and confuse the information that extracted from the text.

Algorithm (3.2): The Proposed PTM Algorithm

```
Input: a set of documents, a threshold $T_H$
Output: a set of paragraphs
Begin
Step 1: For every term T into “d” document
Step2:Assign threshold $T_H$
Step 3: If SUPa or SUPr $\geq T_H$ Then
Step 4:Add T to key term class
Step 5: Measure the accuracy by applying LED algorithm
Step 6:Finish T
End
Else Ignore T then go to step 6
End.
```

3- Clean Paragraph Step:

This is the third step in the proposed system. The purpose of this step is to find the key terms in the text document. Each individual word considers a term. Clean paragraph consists of removing Stop-Word by comparing the terms that are extracted from a document with a list of common words “noise‟ words. This can speed up the process ignoring to run useless queries and any matches with the list are discarded. This is usually called a stop list (the words on this list are called stop-words).Articles, prepositions, and pronouns are the most popular used words in the text documents that provide no meaning and can be considered as stop words. These words are not needful in applications of text mining, therefore, these words will be eliminated. The following words represents an example of these stop- words are "a", "an", "the", "in", "and ", " but ", " near", "to", "it", "as", "able", "about", "above", "of", "allow", "allows", "alone", "am", "an"," and", "but"," clearly", "can", "consider", and ….etc.

4.Feature Extraction

Global probability: global probability is the probability of the term existence in the document. Global probability of a term (P) can be computed by using the equation (2.16).

5-Update Document Information This is the fifth step in the proposed system for the purpose of using the semantic information in the pattern taxonomy in order to improve the discovered pattern performance in text mining. The term with a higher value of TF would be no meaning when it has not cited via some significant parts of documents. The resulting terms is sorting by using Timsort algorithm.

Timsort is a hybrid stable sorting algorithm, constructed for doing a well-performing on many real-world data types. This algorithm obtains the data subsequences which are formerly ordered and utilizes that knowledge for sorting the residue more effectively. This is accomplished via merge an identified subsequence, named a run, with existing runs till certain criteria are done. The algorithm (3.2) describes Timsort algorithm and the algorithm (3.3) describes the applied Deploy Pattern Algorithm.

Algorithm (3.3) : Timsort Algorithm

```
Input : a stack Q of elements W,X,Y,Z.
Output : merging elements.
Begin
    Step 1 : procedure Timsort (s,n)
                R ← Run s
                Q ← θ
                While R‡ θ do
                    Remove the next run R from ℜ and push it on to θ
    Step 2 : Loop
                If /X/ < /Z/ then merge X and Y
                Else
                if /X/ ≤ /Y/ + /Z/ then merge Y and Z
                Else
                If /W/ ≤ /X/ + /Y/ then merge Y and Z
                Else
```

Step 3 :

```
Break out of the Loop
end if
end Loop
end while
While /Q/ ≥ 1 do
 merge Y and Z
  End while
  End procedure
```

W,X,Y,Z denote the top four elements of the stack (Q) , the test involving a stack member that does not exist evaluates as “False”, for example /X/ < /Z/ evaluates as false when /Q/ < 3 and X does not exist.

End.

3- Applying the Levenshtein Distance Algorithm:

This is the last step in the proposed system, algorithm (2.2) is applied for the resulting terms that are obtained from the previous step. LDA take the resulting terms with the title of the document and measure the similarity among them according to the equation (2.2). LDA was used to test the efficiency of PTM algorithm and get more accurate results for pattern discovery. Table (3.3) is an example to measure the similarity between two short texts using Levenshtein edit distance.

Table(3.2) Using LED Algorithm for Similarity Measurements Between Two Short Texts

Table	Text1	Text2	K	P
1	Boy	Boys	1	0.75
2	Baby	Babies	3	0.5
3	Tom is drawing a tree	Tom is draw trees	3	0.40

Where

K: is the number of the difference of characters between two words or sentences.

P: is the probability of similarity between two words or sentences.

From table (3.3), LDA calculates the probability through measuring the similarity between two words or two sentences by calculating the number of the letters of the word to the longest word between them.

Chapter Four

Experiments and Results

Chapter Four

Experiments and Results

4.1 Introduction

This chapter summarizes the implementation outcomes that were obtained by the developed system and described in chapter three. The experimental results and tests of the system phases will be explained. In other words, this chapter is to evaluate the performance of the proposed pattern discovery system. It will contain a detailed description of the steps involved in application implementation.

4.2 The Environment of Implementation

The implementation of the developed system is performed using python 3.8 and pycharm version 2019.1 programming language by a laptop computer with windows7 ultimate. The experiments were performed on an Intel (R) Core (TM) i5-4210U CPU @1.70 GHz 2.40 GHz, 64-bit Operating System, 4GB VGA (NVidia) and 8GB RAM. The detailed steps and implementation results will be explained for each step to accomplish the suggested system.

4.3The Dataset

In this thesis collection the image-pdf file by download a number of 'pdf' files from datasets real world documents for OCR testing large(N>10000) which contains a mix of document image pdf format in English from https://www.kagle.com/data/40647 as shown in figure (4.1).

Figure (4.1)The Dataset “pdf” Files.

4.4 Word Segmentation and Recognition by OCR Implementation in Tesseract

The OCR technique achieved by Tesseract, where the Tesseract has implemented a Long Short Term Memory(LSTM) as describe in following:

4.4.1 Initializing Process

To keep the implementation efficient with not recalculation the skew angle or the dynamic parameters for each new image the following assumptions were established throughout a test session:

- The image resolution will be the same.
- The theme will remain the same.

To add support for other types of interfaces, the initialization process can easily be restarted by calling the initialize function again. The first step is to start Tesseract OCR engine, it is done by executing and starting a new thread running Tesseract separate from the implementation thread. If the resolution width is greater than the height it is considered to be a wide image and the program will use a wide SE (SE(wide)). Otherwise the image is considered to be tall and a tall SE (SE(tall)) will therefore be used. The tall SE have a bigger vertical parameter in order to match the image. While the wide SE have a smaller vertical parameter.

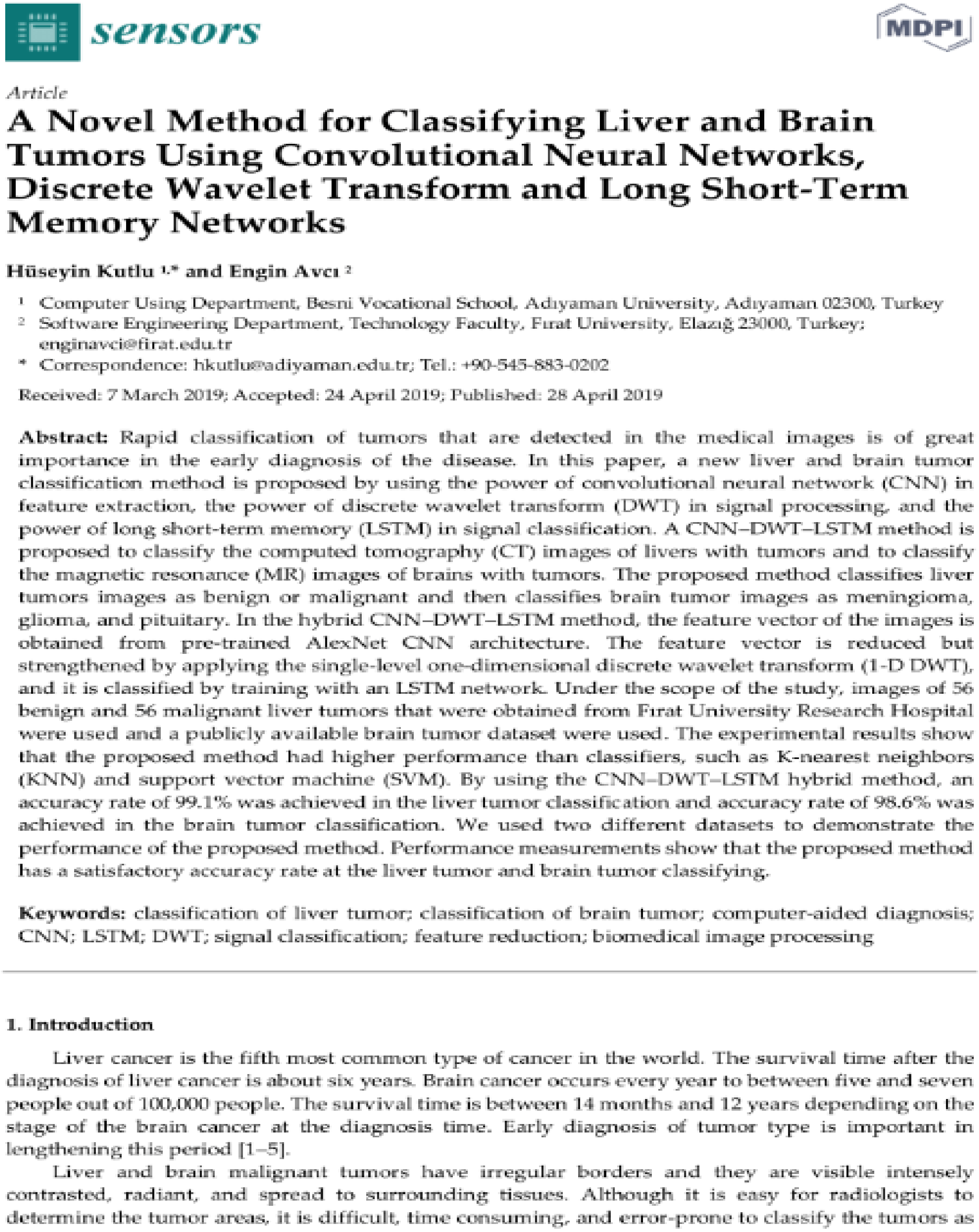
4.4.2 Function Calls

The function calls are important for system testing and also very helpful for experimenting with symbol recognition. OCR, find Word and find Text are three functions sharing the same logic for finding the result. However the last step of each function different results is returned depending on the function calls objective. Initialize is the first function to be used to start and initializing the system.

4.4.3 Segment Identification

Image processing is an essential to extract segments of words and symbols. The process starts by rotating an image but only if a skew angle were detected during the initialization. An image can be represented by a matrix to rotate it based on the given skew angle into a new perpendicular image. The rotated image is then inverted when the invert parameter flag is enabled from the initialization. The image is then converted into gray and morphological opening and closing is applied. This is to remove any static noise which can be found in bad quality images. The next step of the image processing is to convert it into binary, but only if the image is considered noisy, Otherwise it is put through the Canny Edge Detection. It is extremely helpful to enhance the characters and symbols and at the same time remove disjoint pixels. The image is then closed again, this is to connect letters and the small symbols into a blobs forming words or a symbols. The figure (4.4) showcasing the extraction of the blobs on a noisy image.

After the primary image processing is completed the blobs consisting of words are enhanced. Next step is to localize the blobs and extract their positions to later extract the segments. It is done by applying a contour on a closed binary image. This will help to finds all the blobs and encloses them by contours. bounding rectangle code is then used to bound each contour into the smallest possible rectangles. All rectangles which are extremely large or extremely small are considered to be insignificant, and will be removed for efficiency reasons. The remaining rectangles’ positions are inserted into a result as shown in table (4.1) and represent the positions of each segment which later is recognized by Tesseract figure (4.4) where the segment positions are extracted from an image without any noise. Figure (4.2) show a picture of one page from document scanned by OCR Tesseract, while figure (4.3) show the result after applying the steps of word segmentation by OCR implementation in Tesseract and rectangle code in python .



sensors

MDPI

Article

A Novel Method for Classifying Liver and Brain Tumors Using Convolutional Neural Networks, Discrete Wavelet Transform and Long Short-Term Memory Networks

Hüseyin Kutlu [1,]* and Engin Avcı [2]

[1] Computer Using Department, Besni Vocational School, Adıyaman University, Adıyaman 02300, Turkey
[2] Software Engineering Department, Technology Faculty, Fırat University, Elazığ 23000, Turkey; enginavci@firat.edu.tr
* Correspondence: hkutlu@adiyaman.edu.tr; Tel.: +90-545-883-0202

Received: 7 March 2019; Accepted: 24 April 2019; Published: 28 April 2019

Abstract: Rapid classification of tumors that are detected in the medical images is of great importance in the early diagnosis of the disease. In this paper, a new liver and brain tumor classification method is proposed by using the power of convolutional neural network (CNN) in feature extraction, the power of discrete wavelet transform (DWT) in signal processing, and the power of long short-term memory (LSTM) in signal classification. A CNN–DWT–LSTM method is proposed to classify the computed tomography (CT) images of livers with tumors and to classify the magnetic resonance (MR) images of brains with tumors. The proposed method classifies liver tumors images as benign or malignant and then classifies brain tumor images as meningioma, glioma, and pituitary. In the hybrid CNN–DWT–LSTM method, the feature vector of the images is obtained from pre-trained AlexNet CNN architecture. The feature vector is reduced but strengthened by applying the single-level one-dimensional discrete wavelet transform (1-D DWT), and it is classified by training with an LSTM network. Under the scope of the study, images of 56 benign and 56 malignant liver tumors that were obtained from Fırat University Research Hospital were used and a publicly available brain tumor dataset were used. The experimental results show that the proposed method had higher performance than classifiers, such as K-nearest neighbors (KNN) and support vector machine (SVM). By using the CNN–DWT–LSTM hybrid method, an accuracy rate of 99.1% was achieved in the liver tumor classification and accuracy rate of 98.6% was achieved in the brain tumor classification. We used two different datasets to demonstrate the performance of the proposed method. Performance measurements show that the proposed method has a satisfactory accuracy rate at the liver tumor and brain tumor classifying.

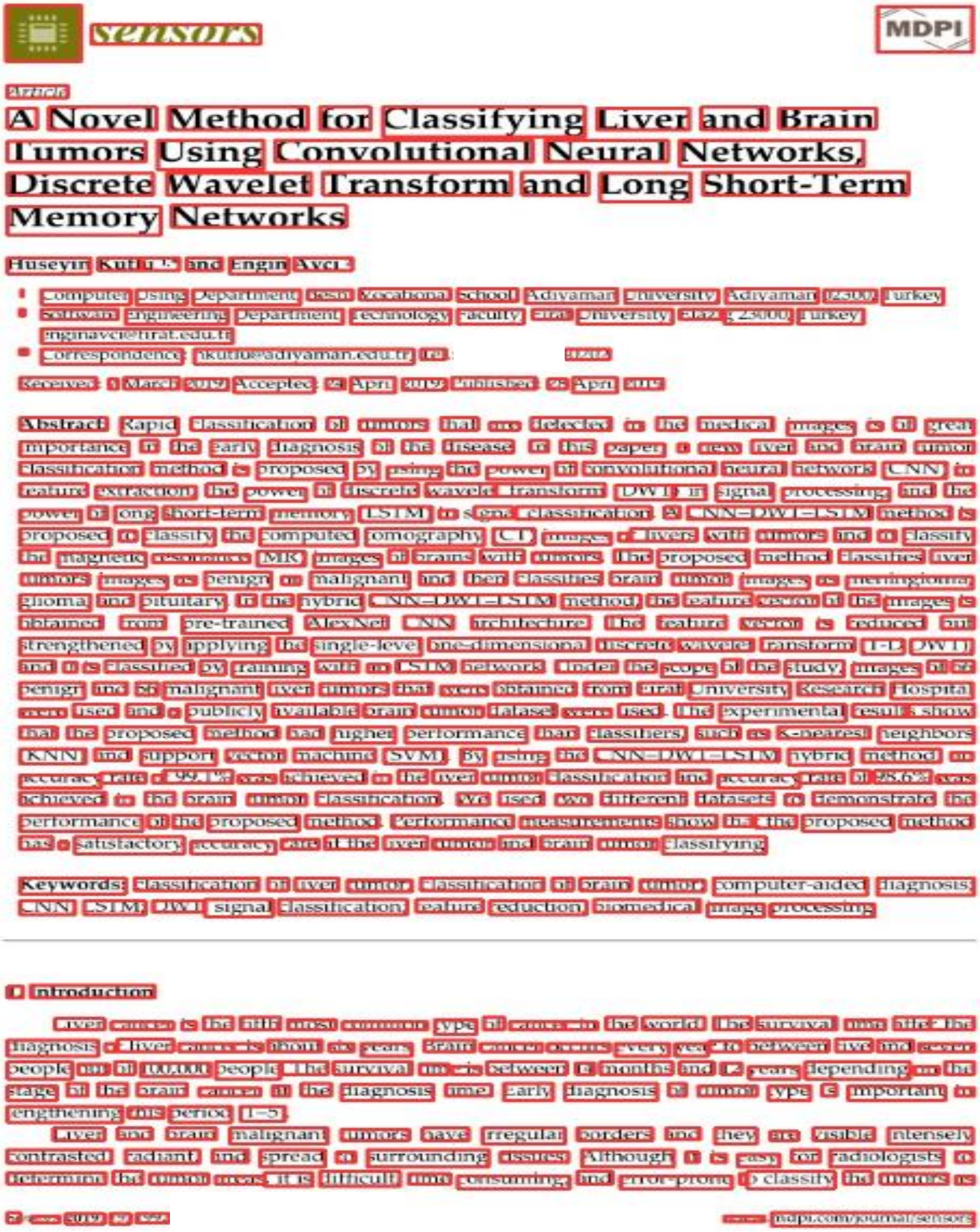
Keywords: classification of liver tumor; classification of brain tumor; computer-aided diagnosis; CNN; LSTM; DWT; signal classification; feature reduction; biomedical image processing

1. Introduction

Liver cancer is the fifth most common type of cancer in the world. The survival time after the diagnosis of liver cancer is about six years. Brain cancer occurs every year to between five and seven people out of 100,000 people. The survival time is between 14 months and 12 years depending on the stage of the brain cancer at the diagnosis time. Early diagnosis of tumor type is important in lengthening this period [1–5].

Liver and brain malignant tumors have irregular borders and they are visible intensely contrasted, radiant, and spread to surrounding tissues. Although it is easy for radiologists to determine the tumor areas, it is difficult, time consuming, and error-prone to classify the tumors as

Figure(4.2) A Picture of One Page from Document Scanned by OCR Tesseract.



sensors

MDPI

Article

A Novel Method for Classifying Liver and Brain Tumors Using Convolutional Neural Networks, Discrete Wavelet Transform and Long Short-Term Memory Networks

Huseyin Kutlu 1,* and Engin Avci 2

1 Computer Using Department, Besni Vocational School, Adiyaman University, Adiyaman 02300, Turkey
2 Software Engineering Department, Technology Faculty, Firat University, Elazig 23000, Turkey; enginavci@firat.edu.tr
* Correspondence: hkutlu@adiyaman.edu.tr; Tel.: 3212

Received: 1 March 2019; Accepted: 24 April 2019; Published: 26 April 2019

Abstract: Rapid classification of tumors that are detected in the medical images is of great importance in the early diagnosis of the disease. In this paper, a new liver and brain tumor classification method is proposed by using the power of convolutional neural network (CNN) in feature extraction, the power of discrete wavelet transform (DWT) in signal processing, and the power of long short-term memory (LSTM) in signal classification. A CNN–DWT–LSTM method is proposed to classify the computed tomography (CT) images of livers with tumors and to classify the magnetic resonance (MR) images of brains with tumors. The proposed method classifies liver tumors images as benign or malignant and then classifies brain tumor images as meningioma, glioma, and pituitary. In the hybrid CNN–DWT–LSTM method, the feature vector of the images is obtained from pre-trained AlexNet CNN architecture. The feature vector is reduced but strengthened by applying the single-level one-dimensional discrete wavelet transform (1-D DWT), and it is classified by training with an LSTM network. Under the scope of the study, images of 56 benign and 56 malignant liver tumors that were obtained from Firat University Research Hospital were used and a publicly available brain tumor dataset were used. The experimental results show that the proposed method had higher performance than classifiers, such as K-nearest neighbors (KNN) and support vector machine (SVM). By using the CNN–DWT–LSTM hybrid method, an accuracy rate of 99.1% was achieved in the liver tumor classification and accuracy rate of 98.6% was achieved in the brain tumor classification. We used two different datasets to demonstrate the performance of the proposed method. Performance measurements show that the proposed method has a satisfactory accuracy rate at the liver tumor and brain tumor classifying.

Keywords: classification of liver tumor; classification of brain tumor; computer-aided diagnosis; CNN; LSTM; DWT; signal classification; feature reduction; biomedical image processing

1. Introduction

Liver cancer is the fifth most common type of cancer in the world. The survival time after the diagnosis of liver cancer is about six years. Brain cancer occurs every year in between five and seven people out of 100,000 people. The survival time is between 6 months and 12 years depending on the stage of the brain cancer at the diagnosis time. Early diagnosis of tumor type is important in lengthening this period [1–5].

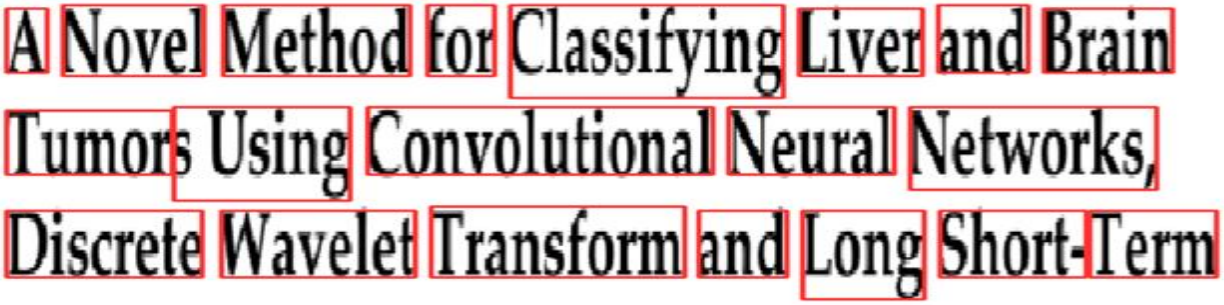
Liver and brain malignant tumors have irregular borders and they are visible intensely contrasted, radiant, and spread to surrounding tissues. Although it is easy for radiologists to determine the tumor areas, it is difficult, time consuming, and error-prone to classify the tumors as

Sensors 2019, 19, 1992 www.mdpi.com/journal/sensors

(4.3) The Result After Applying The Steps of Word Segmentation by OCR Implementation in Tesseract .

4.4.4 Recognition

By allowing Tesseract to recognize the whole image in a single process, the recognition results will be poor. This is because of the different text and symbol sizes causing problems for Tesseract with structuring the information. To address this problem, the isolation of each respective segment is essential to provide a more accurate recognition. This is also practical because it can pinpoint the exact location of the segment which is required for localization during system testing. Each segment is extracted from the original image into small images containing only the information of the segment as shown in figure(4.4). The segments are then fed into the Tesseract engine one by one. Tesseract receives them and starts to recognize them, and as a result Tesseract will return an answer of the recognition along with a confidence of the analysis. This will be added to the already existing result list where segment positions were inserted.



A Novel Method for Classifying Liver and Brain
Tumors Using Convolutional Neural Networks,
Discrete Wavelet Transform and Long Short-Term

Figure (4.4) Words Recognition Result by Tesseract.

The rectangles' positions are inserted into a result as shown in table (4.1) after applying the word recognition by Tesseract OCR as shown in figure (4.4).

Table(4.1) Words Recognition by OCR Tesseract.

Level	Left	Top	Width	Height	Text
1.	9	34	23	76	A
2.	41	30	80	84	Novel
3.	132	30	106	84	Method
4.	248	30	38	84	for
5.	295	30	155	110	Classifying
6.	459	34	70	80	Liver
7.	538	30	51	80	and
8.	598	34	73	76	Brain
9.	9	155	94	76	Tumor
10.	104	151	100	111	Using
11.	214	151	196	80	Convolutional
12.	419	151	94	80	Neural
13.	522	151	140	98	Networks
14.	9	274	111	79	Discrete
15.	131	274	110	79	Wavelet
16.	250	269	144	84	Transform
17.	402	274	50	79	and
18.	461	274	69	105	Long
19.	539	274	83	79	Shot

4.5 The Keyword Detection

4.5.1 The Word Detection

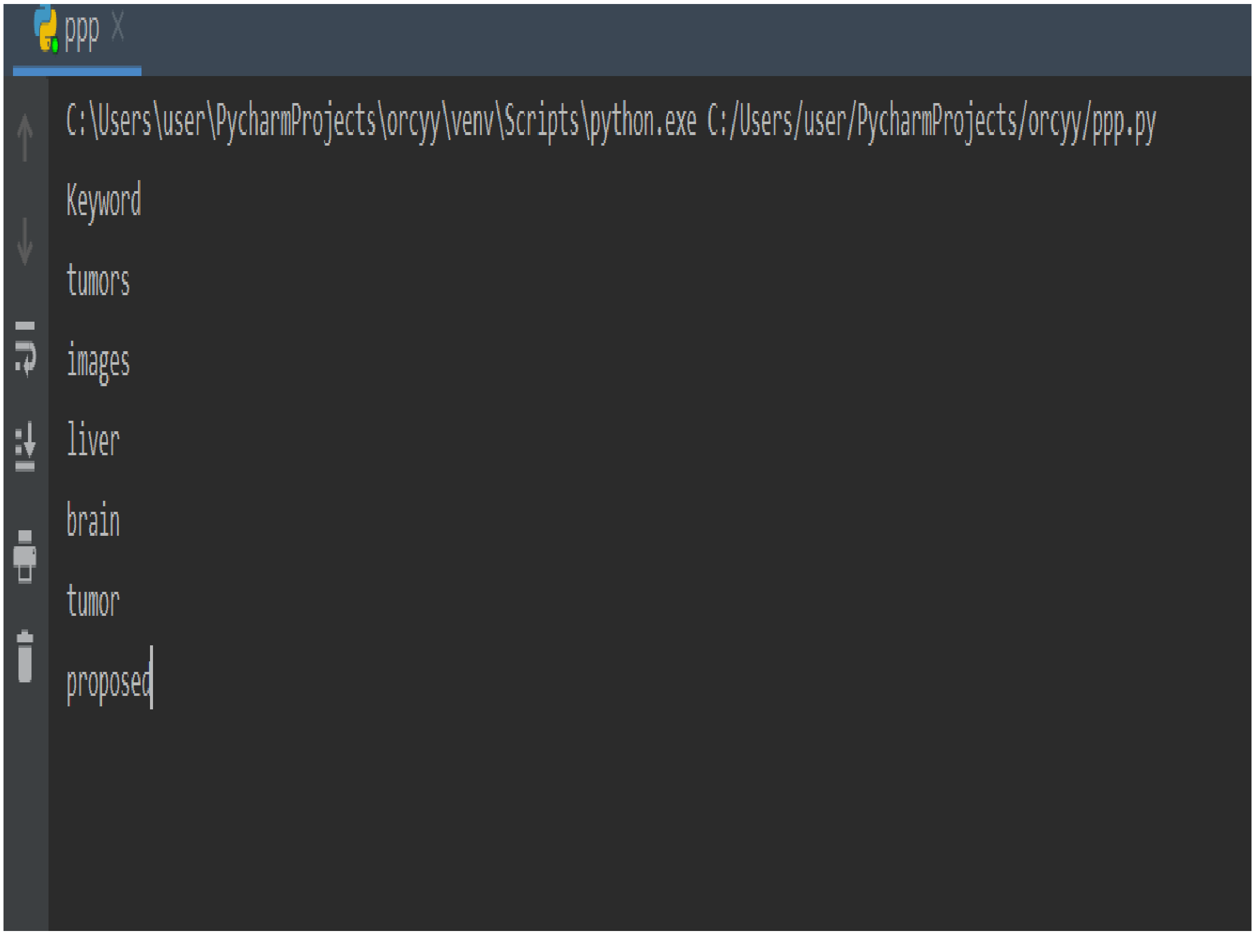
In this step, after loading the text, PTM algorithm is performed on the text as implemented in the flowchart of figure (3.5) and algorithm (3.2). The value of threshold is between the range[1-10] to get from 1% to 10% the information in the text. The value of threshold will be compared with the value of global probability, absolute support, and the relative support. After the comparison, stop-words will be eliminated .The result of this step is that each word in each paragraph is considered as a term. Table(4.3). states the division of the text into paragraphs(all result shown in Appendix A). The table (4.2) shows the extraction of 426 word of document, these term obtain from Tesseract OCR.

Table(4.2)The Words Detection In Document by OCR.

level	width	height	text	width	height	Text
1.	19	18	A	44	9	tumors
2.	70	20	Novel	24	11	That
3.	91	20	Method	19	7	Are
4.	33	20	for	52	11	detected
5.	132	26	Classifying	11	7	In
6.	60	19	Liver	19	11	The
7.	43	20	and	49	11	medical
. .						
207	68	12	contrasted	38	9	tumor
212	79	15	surrounding	73	11	consuming
213.	44	9	tissues	60	15	Although

4.5.2 Applying LDA Algorithm

By applying the similarity equation which is described previously in the equation (2.12) on the resulting terms from the previous step, we can get the keyword of each document by implement equation (2.11). After extracting the word form document and recognition, apply the Levenshten Distance Algorithm for calculate the similarity from word and calculate the more frequent word , where this word represent the key word as shown in figure(4.5) which shows the key word of the whole document that used and figure(4.2) shows picture of one page from this document that was scanned by OCR Tesseract , the five keyword extract from more than500 word .



```
C:\Users\user\PycharmProjects\orcyy\venv\Scripts\python.exe C:/Users/user/PycharmProjects/orcyy/ppp.py
Keyword
tumors
images
liver
brain
tumor
proposed
```

Figure (4.5) The Keywords That are Founded In The Document.

4.5.3 Feature Extraction step

In this step, the global probability feature were calculated for the key words of whole document by implementing equation (2.16).

Table (4.3) The Calculation of The Global Probability for Each Key Word in The Document.

keyword	Global Probability
tumors	0.028
images	0.036
liver	0.013
brain	0.039

The table (4.3) shows the value of feature keyword (tumors, images, liver, brain) where the keyword of "tumors" show the more word has frequently repeated in document.

4.6 Results and Discussion

This subsection shows the results of finding the average accuracy for the global probability and the time that required for calculation it in the proposed system. This is done based on the value of the threshold from one to ten .Where the accuracy is calculated by taking a picture of one page from the file , so that the number of repeated words calculated manually to determine how many it was, then applying the process of calculating accuracy programmatically by making the threshold greater than one and see how many words it will match, then increases the threshold value by one and so on calculate the percentage of match between what was calculated manually and what was calculated programmatically.Table (4.5)describes in details the average accuracy and time of processing for the document.

Table(4.4) The Average Accuracy and Time of The Proposed System of Unigram Grammar for The Global Probability.

Threshold	The Average Accuracy of Global Probability	Elapse Time In Sec
1	94.02%	5.84
2	78.01%	5.74
3	55.19%	6.20
4	39.5%	4.96
5	31.61%	5.33
6	26.36%	4.94
7	23.95%	5.04
8	19.74%	4.82
9	18.02%	4.81
10	16.72%	4.81

From table (4.5), the average accuracy of global probability for threshold values from 1 to 10 are decreasing from 94.02% to 16.72% because the probability of the term appearing in the document decreases as the threshold value increases, The table shows that time is convergent and decreases slightly with increased threshold.

Chapter Five

Conclusions and Future Works

Chapter Five

Conclusions and Future Works

5.1 Conclusions

In this section will illustrated as following conclusions of proposed work:

1. This work study was been proved that there was a way to create a structure for unstructured image-based files.
2. Although that OCR technology is useful for reading the whole pages of unstructured elements of any type of documents, such as scanned paper documents, PDF files , Printed or written texts , and images captured by digital camera by analysis and translate the character image into character codes such as ASCII codes that be editable and searchable data, but it has some problems such as lack the ability to detect simple elements in the page and it also lacks in structuring the data that it read because it prints it in one long string.
3. The system exploits the OCR capabilities as well as improves it by using AI in order to give it the ability to detect any word inside the page by adding a bounding box around it automatically, and gives it the ability to convert unstructured element to structured one to response the queries that are represented by determining the key words in the whole document by providing a technique for discovering patterns by using (LDA) with (PTM).
4. The system can be modified and programmed according to quires or user's requirements, and hence improve BI.
5. The system contributes to improving archiving process by discovering the repeated words from many documents which will speeding up the organizing files process, information search process as well as

ensuring text integrity by making it free from any noise or loss. Hence, saving effort.

6. The system can be used in enterprises, business environment, and museum to save old files.
7. Each language has its own properties that differ from other languages, so this system needs modifications and additions when dealing with a language other than English.

5.2 Recommendations for Future Work

For future development and from our viewpoint which was its basis built on the results of the present work, we introduce the following recommendations

1. This would provide the possibility to allow one tester to run tests in several languages. However, if a translation package were available the tester could build and validate tests in their preferred language, translated from the foreign language.
2. The proposed system can be personalized, so that data or results obtained from the text document is concerned with the profile of the user. Here, to search for information, the profile of the user will be generated to the system.

References

References

[1] Dobre,C,. Xhafa,F,(2014).”**Intelligent services for big data science .Future generation computer Systems”**.

[2] Gantz,J, .Reinsei,D.(2012).”**The Digital Universe in 2020;Bigger digital shadows, and biggest growth in far East”.** Corporation. Online Available at http://www.emc.com/collateral/analyst-reports/ide-the-digital-universe-in-2020.pdf.

[3] Kamal,M.,Irani,Z.Weerakkody,V.(2017). **“Critical analysis of Big Data challenges and analytical methods**". University London ,Brunel Business School, UBB 3PH,United Kingdom.

[4] Ahmed,F.,Mohammed., Vikas T., Santosh S .(2016**).” Review of Big Data Environment and Its Related Technologies**”.

[5]Han,H.,Yonggag.,Tat,S.,Chu.,Xuelong.(2014)**“Toward Scalable Systems for Big Data Analytics”**.

[6] Patrick,M., Ilias O. Pappas1.,John Krogstie1., Michail Giannakos “**Big data analytics capabilities: a systematic literature review and research agenda “**.

[7] Fatimah,S., Iskandar, I ., Marzanah,A.(2018).“**An Ontological Approach to Knowledge Transformation in Malay Unstructured Documents”.**

[8] Fonferko ,S., Beata., L, Arron S., Roberts,A.,Akbari, A., Thompson, S., Ford,David V., Lyons, Ronan A., Rees, Mark I. & Pickrell, William Owen. (2019). "**Using natural language processing to extract structured epilepsy data from unstructured".**

[9]Krishnan ,A., Abilash,A.(2019) **"Large Scale Product Categorization using Structured and Unstructured Attributes".**

[10] Abdul Robby, G **,.** (2019)**. "Implementation of Optical Character Recognition using Tesseract with the Javanese Script Target in Android Application"**.

[11] Olalekan,J., ONI. (2020**)."Computational modelling of an optical character recognition system for Yorùbá printed text images".**

[12] Mayur, B., Bora .(2020). **" Handwritten Character Recognition from Images using CNN-ECOC"**.

[13] Bitty,B.,Detelina,M.(2018)**."Unstructured data in marketing".**

[14] R, Sethy .,M, Panda.(2015) **."Big data analysis using hadoop: a survey".**

[15]S, V., Phaneendra. , E, M., Reddy.(2013) **."Big Data-solutions for RDBMS problems-A survey".**

[16] I. Taleb, M. A. Serhani, R. Dssouli .(2018**)."Big data quality assessment model for unstructured data".**

[17] Imane, L.,Said ,EL H., Ghizlane.(2016**)." Managing big data integrity**".

References

[18] A. J. I. J. o. I. R. I. A. E. Nath.(2015) **."Big Data Security Issues and Challenge".**

[19] H. Jagadish et al.(2014**)."Big data and its technical challenges".**

[20] J. J. h. w. m. c. I. M. R. T. a. I. B. d. T. n. f. f. i. Manyika.(2011) **."Big data".**

[21] A. L'heureux, K. Grolinger, H. F. Elyamany, and M. A. J. I. A. Capretz, (2017**)."Machine learning with big data: Challenges and approaches".**

[22]Y. Bengio, A. Courville, P. J. I. t. o. p. a. (2013**)."Representation learning: A review and new perspectives"** .

[23]M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. J. J. o. B. D. Muharemagic.(2015). **"Deep learning applications and challenges in big data analytics"** .

[24] Rumelhart, D. E., Hinton, G. E., Williams, R. J. (1986**)." Learning representations by backpropagating errors**".

[25]Sak, H., Senior, A., & Beaufays, F. (2014**)." Long short-term memory recurrent neural network architectures for large scale acoustic modeling**".

[26] J. Serra. (1982).**"Image analysis and mathematical morphology".**

[27] A. Niewiarowski.(2016) .**''Short Text Similarity Algorithm Based on The Edit Distance and Thesaurus**''.

[28]X. Chen and Q. Huang.(2013) .**"The data protection of MapReduce using homomorphic encryption".** in 2013 IEEE 4th International Conference on Software Engineering and Service Science, 2013, pp. 419-421: IEEE.

[29] S-S. Kang, (2015).**''Word Similarity Calculation by Using the Edit Distance Metrics with Consonant Normalization''**.

[30] P. A. Utama. , B. Distiawan,(2015) **'' Spark-Gram: Mining Frequent Ngrams using Parallel Processing in Spark''**.

[31] Ray Smith.(2007).”An Overview of the tesseract OCR Engine”.

[32]Vector Ohlsson .(2016) .**”Optical Character and Symbol Recognition using Tesseract”.**

[33] Bo Wang.,Yi-wei Ma.,Hong-Tao Hu(2020**).”Hybrid model for Chinese character recognition based on Tesseract”.**

[34] William, B. ,Cavnar ., John M. Trenkle.(2004). **''N-Gram-Based Text Categorization''**.

[35] J, Houvardas ., E. Stamatatos.(2006). **'' N-gram Feature Selection for Authorship Identification ''.**

Appendix A

Appendix A

Word Detection: The result of this step is that each word in each paragraph is considered as a term as shown in table 1

Table 1: the recognition of word in document (as shown 4.4) in OCR

level	width	height	text	width	height	text
1.	19	18	A	44	9	tumors
2.	70	20	Novel	24	11	that
3.	91	20	Method	19	7	are
4.	33	20	for	52	11	detected
5.	132	26	Classifying	11	7	in
6.	60	19	Liver	19	11	the
7.	43	20	and	49	11	medical
8.	63	18	Brain	44	11	images
9.	90	18	Tumors	9	7	is
10.	69	26	Using	13	11	of
11.	168	19	Convolutional	31	13	great
12.	80	19	Neural	71	13	importance
13.	120	23	Networks,	11	10	in
14.	95	19	Discrete	18	11	the
15.	94	19	Wavelet	30	14	early
16.	123	20	Transform	60	14	diagnosis
17.	43	20	and	12	11	of
18.	59	25	Long	19	11	the
19.	134	20	Short-Term	45	11	disease.
20.	100	25	Memory	12	10	In
21.	114	19	Networks	22	11	this
22.	57	11	Abstract:	38	11	paper,
23.	37	15	Rapid	6	7	a
24.	80	11	classification	26	8	new
25.	12	11	of	28	11	liver
26.	33	10	brain	37	15	signal
27.	38	10	tumor	70	11	processing,
28.	80	11	classification	22	11	and
29.	48	11	method	19	11	the
30.	9	7	is	40	11	power
31.	58	15	proposed	12	11	of
32.	15	15	by	27	14	long
33.	35	11	using	65	11	short-term

34.	19	11	the	53	11	memory
35.	40	11	power	47	13	(LSTM)
36.	13	11	of	11	7	in
37.	86	11	convolutional	25	10	signal
38.	41	11	neural	88	11	classification.
39.	53	11	network	10	10	A
40.	41	14	(CNN)	120	11	CNN-DWT-L.STM
41.	11	7	in	47	11	method
42.	44	11	feature	9	8	is
43.	65	11	extraction,	59	15	proposed
44.	19	11	the	11	9	to
45.	40	11	power	46	15	classify
46.	12	11	of	19	11	the
47.	49	11	discrete	63	15	computed
48.	45	11	wavelet	77	15	tomography
49.	71	11	transform	27	14	(CT)
50.	37	13	(DWT)	44	10	images
51.	7	7	of	64	15	malignant
52.	43	11	livers	22	11	and
53.	27	11	with	27	11	then
54.	44	9	tumors	56	11	classifies
55.	23	11	and	32	11	brain
56.	11	9	to	38	9	tumor
57.	46	14	classify	44	11	images
58.	19	11	the	13	7	as
59.	57	13	magnetic	82	11	meningioma,
60.	63	7	resonance	46	14	alioma,
61.	31	13	(MB)	23	10	and
62.	44	11	images	55	14	pituitary.
63.	12	11	of	12	10	In
64.	39	11	brains	19	10	the
65.	28	11	with	42	14	hybrid
66.	44	9	tumors.	120	10	CNN-DWT-LSTM
67.	24	11	The	51	12	method,
68.	59	15	proposed	19	10	the
69.	48	11	method	44	11	feature
70.	55	11	classifies	39	9	vector

71.	28	11	liver	12	11	of
72.	44	9	tumors	19	10	the
73.	44	11	images	44	11	images
74.	13	7	as	9	7	is
75.	43	15	benign	54	11	obtained
76.	12	7	or	9	7	is
77.	29	11	from	57	10	classified
78.	71	15	pre-trained	16	12	by
79.	53	11	AlexNet	49	13	training
80.	32	11	CNN.	27	10	with
81.	77	11	Architecture.	14	7	an
82.	24	11	The	38	10	LSTM
83.	44	11	feature	53	10	network.
84.	39	9	vector	40	10	Under
85.	9	7	is	19	10	the
86.	50	11	reduced	35	11	scope
87.	21	11	but	12	11	of
88.	82	15	strengthened	19	10	the
89.	15	15	by	38	14	study,
90.	56	15	applying	44	11	images
91.	19	11	the	12	10	of
92.	71	15	single-level	14	10	56
93.	49	10	discrete	42	14	beign
94.	49	10	wavelet	22	10	and
95.	27	13	(1-D	64	14	malignant
96.	40	13	DWT),	28	11	liver
97.	23	10	and	44	10	tumors
98.	8	9	it	24	10	that
99.	30	8	were	23	10	had
100.	53	11	obtained	41	14	higher
101.	29	11	from	79	14	performance
102.	29	10	Firat	27	10	than
103.	66	14	University	64	12	classifiers,
104.	56	11	Research	28	10	such
105.	53	14	Hospital	13	7	as
106.	30	7	were	60	10	K-nearest
107.	29	11	used	62	14	neighbors
108.	23	11	and	43	13	(ENN)
109.	6	7	a	23	10	and

110.	51	15	publicly	49	13	support
111.	56	11	available	38	10	vector
112.	33	11	brain	53	11	machine
113.	39	9	tumor	40	13	(SVM).
114.	44	11	dataset	16	14	By
115.	31	7	were	35	11	using
116.	28	11	used.	19	10	the
117.	24	11	The	120	11	CNN-DWT-LSTM
118.	82	15	experimental	41	14	hybrid
119.	35	11	results	51	12	method,
120.	43	11	show	14	7	an
121.	24	10	that	53	11	accuracy
122.	19	10	the	29	9	rate
123.	58	14	proposed	7	7	of
124.	48	10	method	11	9	to
125.	45	11	99.1%	79	11	demonstrate
126.	24	7	was	18	11	the
127.	55	11	achieved	80	14	performance
128.	11	7	in	12	11	of
129.	19	11	the	19	11	the
130.	28	11	liver	59	14	proposed
131.	38	9	tumor	47	11	method
132.	80	11	classification	79	11	Performance
133.	22	11	and	91	9	measurements
134.	53	11	accuracy	33	11	show
135.	29	9	rate	18	10	that
136.	12	11	of	27	11	the
137.	37	11	98.6%	59	14	proposed
138.	24	7	was	47	11	method
139.	56	11	achieved	21	11	has
140.	11	7	in	5	7	a
141.	19	11	the	71	15	satisfactory
142.	32	11	brain	55	11	accuracy
143.	38	9	tumor	24	9	rate
144.	80	11	classification	34	11	the
145.	20	10	We	28	11	liver
146.	29	11	used	39	9	tumor
147.	23	9	two	23	11	and

148.	55	11	different	33	11	brain
149.	50	11	datasets	38	9	tumor
150.	67	15	classifying;	31	9	most
151.	69	14	Keywords:	55	7	common
152.	13	11	of	27	13	type
153.	28	11	liver	13	11	of
154.	42	10	tumor;	35	7	cancer
155.	80	11	classification	21	7	in
156.	13	11	of	19	11	the
157.	33	11	brain	40	11	world
158.	42	10	tumor,	24	11	The
159.	63	15	diagnosis	52	11	survival
160.	36	12	CNN	26	9	time
161.	40	12	LSTM	23	11	after
162.	33	10	DWT	29	11	the
163.	44	15	signal	59	14	diagnosis
164.	84	13	classification	7	7	of
165.	43	11	feature	36	11	liver
166.	63	12	reduction;	35	7	cancer
167.	69	11	biomedical	19	7	is
168.	37	11	image	35	10	about
169.	68	11	processing	16	7	six
170.	81	11	Introduction	33	11	years.
171.	32	10	Liver	33	10	Brain
172.	40	7	cancer	40	7	cancer
173.	9	7	is	40	7	occurs
174.	19	11	the	35	11	every
175.	26	11	fifth	21	11	year
176.	21	9	to	32	11	brain
177.	53	10	between	40	7	cancer
178.	23	11	five	11	9	at
179.	23	10	and	19	11	the
180.	36	7	seven	60	14	diagnosis
181.	42	15	people	30	10	time
182.	20	9	out	32	15	Early
183.	13	11	of	60	15	diagnosis
184.	45	12	100,000	11	11	of
185.	42	15	people	38	9	tumor
186.	30	11	The	27	13	type

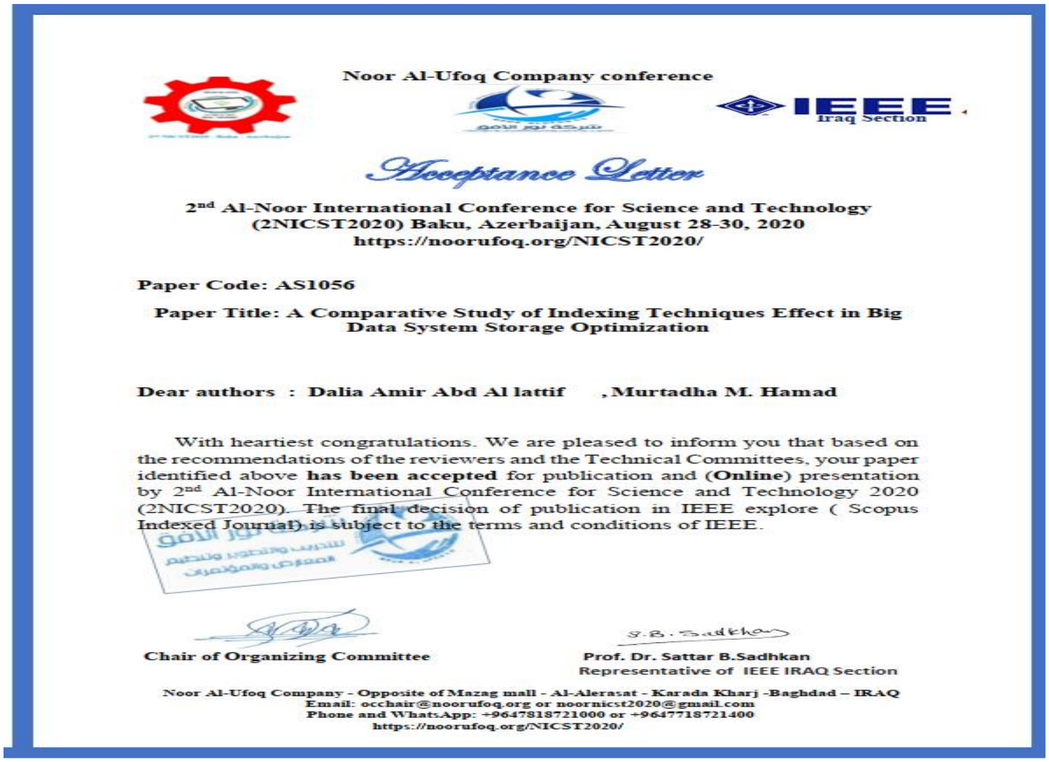
187.	51	11	survival	9	10	is
188.	19	9	time	63	13	important
189.	19	7	is	12	7	in
190.	52	11	between	23	9	this
191.	12	10	14	40	14	period
192.	22	11	and	33	11	Liver
193.	13	10	12	22	10	and
194.	33	11	years	32	10	brain
195.	68	15	depending	64	14	malignant
196.	16	7	on	44	10	tumors
197.	19	11	the	30	11	have
198.	32	13	stage	54	14	irregular
199.	13	11	of	8	9	it
200.	19	11	the	9	7	is
201.	49	11	borders	27	11	easy
202.	22	10	and	18	11	for
203.	27	14	they	74	15	radiologists
204.	18	8	are	11	9	to
205.	40	11	visible	64	10	determine
206.	55	14	intensely	19	11	the
207.	68	12	contrasted	38	9	tumor
208.	48	12	radiant,	32	7	areas
209.	23	11	and	31	12	it is
210.	42	15	spread	52	13	difficult,
211.	12	9	to	27	9	time
212.	79	15	surrounding	73	11	consuming
213.	44	9	tissues	60	15	Although

Appendix B

Appendix B

The published papers:

1. A Comparative Study of Indexing Techniques Effect in Big Data System Storage Optimization



Noor Al-Ufoq Company conference

IEEE Iraq Section

Acceptance Letter

2nd Al-Noor International Conference for Science and Technology
(2NICST2020) Baku, Azerbaijan, August 28-30, 2020
https://noorufoq.org/NICST2020/

Paper Code: AS1056

Paper Title: A Comparative Study of Indexing Techniques Effect in Big Data System Storage Optimization

Dear authors : Dalia Amir Abd Al lattif ,Murtadha M. Hamad

With heartiest congratulations. We are pleased to inform you that based on the recommendations of the reviewers and the Technical Committees, your paper identified above **has been accepted** for publication and (**Online**) presentation by 2nd Al-Noor International Conference for Science and Technology 2020 (2NICST2020). The final decision of publication in IEEE explore (Scopus Indexed Journal) is subject to the terms and conditions of IEEE.

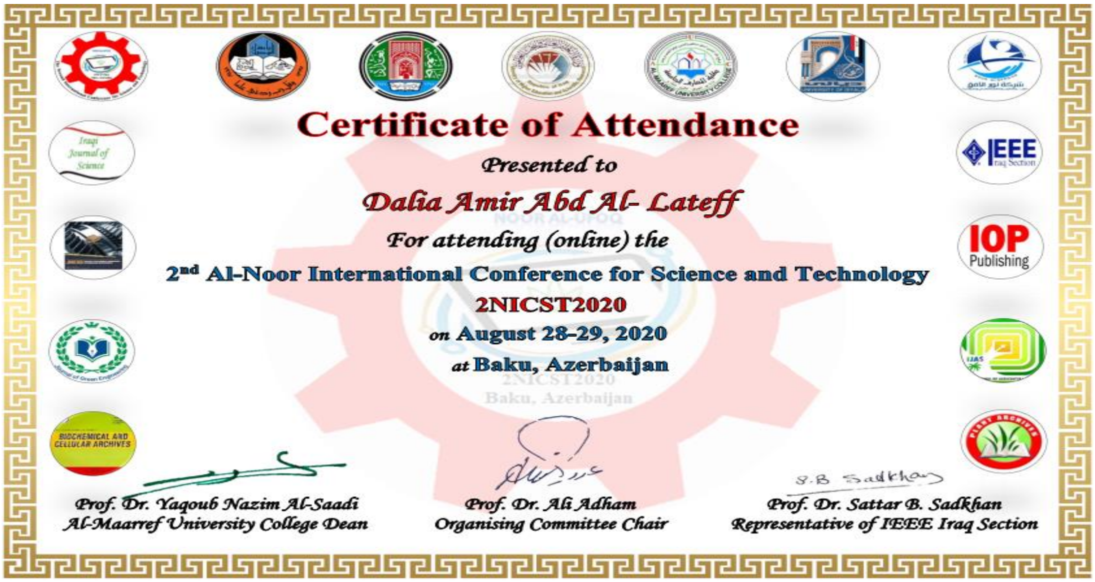
للتدريب والتطوير وتنظيم
المعارض والمؤتمرات

S.B. Sadkhan

Chair of Organizing Committee

Prof. Dr. Sattar B.Sadhkan
Representative of IEEE IRAQ Section

Noor Al-Ufoq Company - Opposite of Mazag mall - Al-Alerasat - Karada Kharj -Baghdad – IRAQ
Email: occhair@noorufoq.org or noornicst2020@gmail.com
Phone and WhatsApp: +9647818721000 or +9647718721400
https://noorufoq.org/NICST2020/



Certificate of Attendance

Presented to

Dalia Amir Abd Al- Lateff

For attending (online) the

2nd Al-Noor International Conference for Science and Technology

2NICST2020

on **August 28-29, 2020**

at **Baku, Azerbaijan**

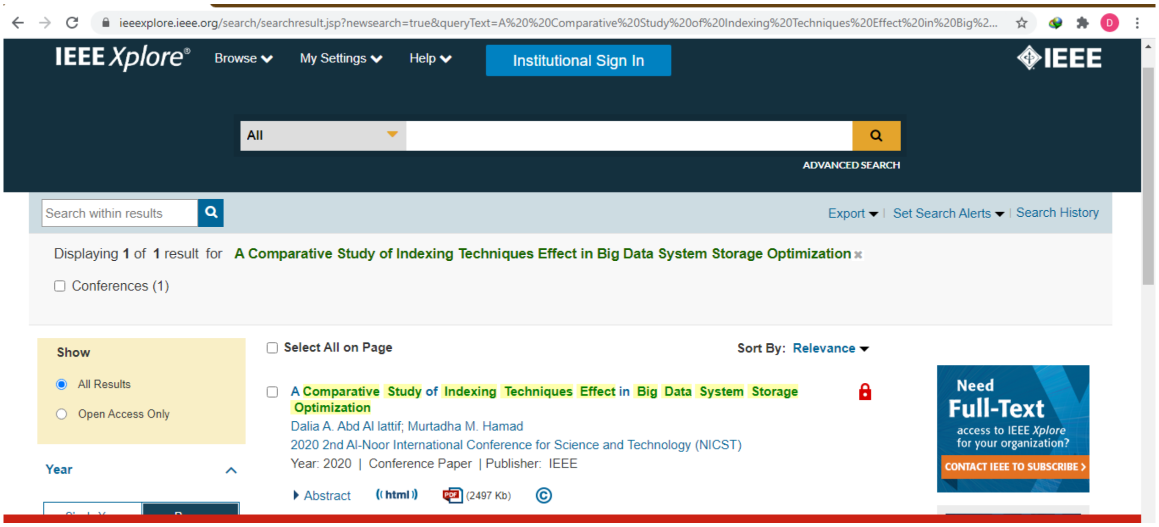
S.B Sadkhan

Prof. Dr. Yaqoub Nazim Al-Saadi
Al-Maarref University College Dean

Prof. Dr. Ali Adham
Organising Committee Chair

Prof. Dr. Sattar B. Sadkhan
Representative of IEEE Iraq Section

Appendix B


ieeexplore.ieee.org/search/searchresult.jsp?newsearch=true&queryText=A%20%20Comparative%20Study%20of%20Indexing%20Techniques%20Effect%20in%20Big%2...
IEEE Xplore®
Browse
My Settings
Help
Institutional Sign In
IEEE
All
ADVANCED SEARCH
Search within results
Export
Set Search Alerts
Search History

Displaying 1 of 1 result for A Comparative Study of Indexing Techniques Effect in Big Data System Storage Optimization

□ Conferences (1)

Show

● All Results

○ Open Access Only

Year

□ Select All on Page

Sort By: Relevance

□ A Comparative Study of Indexing Techniques Effect in Big Data System Storage Optimization

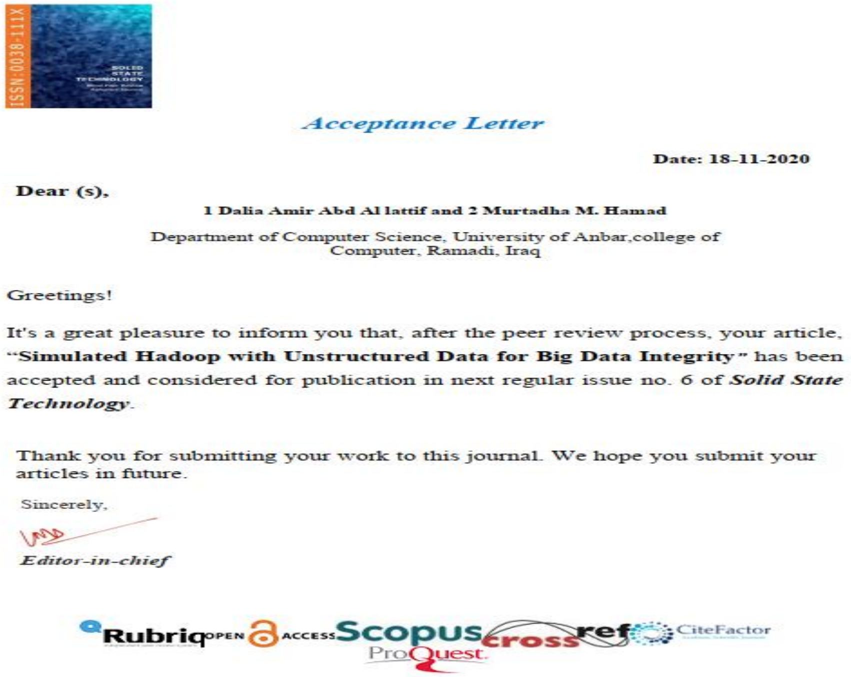
Dalia A. Abd Al lattif; Murtadha M. Hamad

2020 2nd Al-Noor International Conference for Science and Technology (NICST)

Year: 2020 | Conference Paper | Publisher: IEEE

Abstract (html) (2497 Kb)

2.Simulated Hadoop with unstructured data for Big Data Integrity


ISSN:0038-111X
SOLID STATE TECHNOLOGY

Acceptance Letter

Date: 18-11-2020

Dear (s),

1 Dalia Amir Abd Al lattif and 2 Murtadha M. Hamad

Department of Computer Science, University of Anbar,college of Computer, Ramadi, Iraq

Greetings!

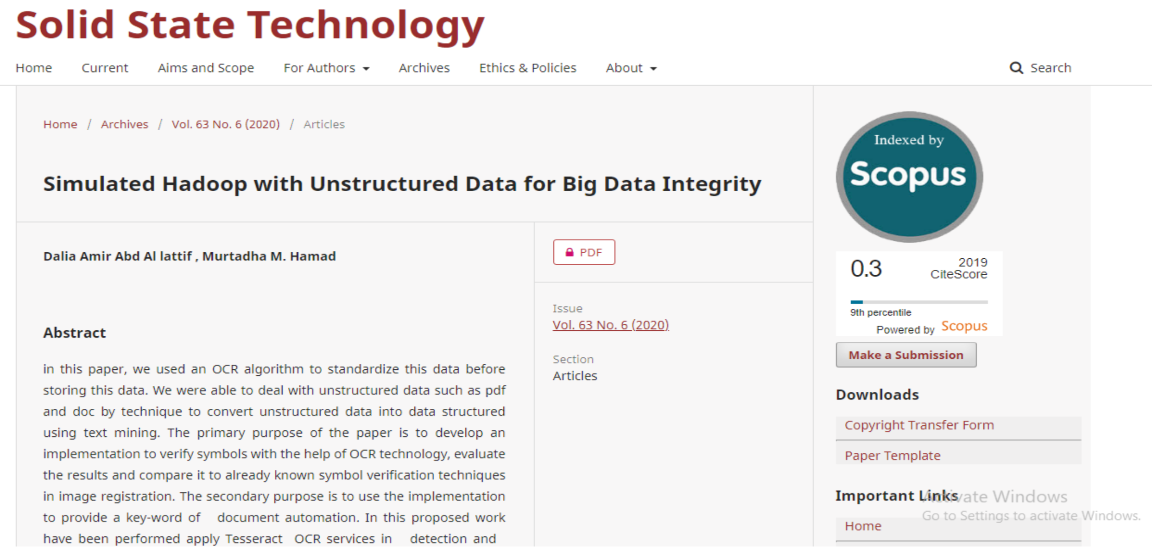
It's a great pleasure to inform you that, after the peer review process, your article, "**Simulated Hadoop with Unstructured Data for Big Data Integrity**" has been accepted and considered for publication in next regular issue no. 6 of ***Solid State Technology***.

Thank you for submitting your work to this journal. We hope you submit your articles in future.

Sincerely,

Editor-in-chief

Appendix B



Solid State Technology

Simulated Hadoop with Unstructured Data for Big Data Integrity

Dalia Amir Abd Al lattif , Murtadha M. Hamad

Issue
Vol. 63 No. 6 (2020)

Section
Articles

0.3
2019
CiteScore
9th percentile
Powered by Scopus

Downloads

Important Links

Abstract

in this paper, we used an OCR algorithm to standardize this data before storing this data. We were able to deal with unstructured data such as pdf and doc by technique to convert unstructured data into data structured using text mining. The primary purpose of the paper is to develop an implementation to verify symbols with the help of OCR technology, evaluate the results and compare it to already known symbol verification techniques in image registration. The secondary purpose is to use the implementation to provide a key-word of document automation. In this proposed work have been performed apply Tesseract OCR services in detection and

الخلاصة

أدت الزيادة المستمرة في البيانات باستخدام أنظمة وتطبيقات مختلفة عبر الإنترنت إلى مشكلة أساسية تتعلق بكيفية إدارة ومعالجة الحجم الضخم من البيانات. هناك العديد من الطرق لتخزينها مثل البيانات المهيكلة وشبه المهيكلة وغير المهيكلة. ومع ذلك ، فإن النقطة الأكثر أهمية هي طريقة خزن البيانات غير المنظمة لأنها تمثل معظم البيانات عبر إدارة الإنترنت باستخدام الأساليب التقليدية غير مناسبة بسبب توفر البيانات الكبيرة والمعقدة. من هنا كان (هدوب)الحل مناسب للزيادة المستمرة في أحجام البيانات وكذلك التعامل معها وتحليلها كما هي من اي مصدر او سرعة او حجم او كمية كانت.

في هذه الأطروحة ، تم اقتراح نظامًا لتحليل البيانات الضخمة من الإنترنت. هذا النظام لديه القدرة على تحديد الكلمات المتكررة (الكلمات الرئيسية) في عدد كبير من ملفات الصور التي تم مسحها ضوئيًا بواسطة(جهاز التعرف الضوئي على المحارف أو ما يسمى OCR) لتسريع البحث عن المعلومات داخل البيانات الورقية وبالتالي توفير الوقت والجهد. يمكن استخدام هذا النظام سواء في المؤسسات وبيئة الأعمال والطب والتعليم والمتاحف والتعداد وما إلى ذلك. يدعم النظام اتخاذ القرار من خلال توفير كيان مهم للرد على الاستفسارات القائمة على الكلمات الرئيسية .استخدمنا في هذا النظام تقنيات وخوارزميات تتعامل مع البيانات الضخمة بما في ذلك جمع البيانات والمعالجة المسبقة للبيانات وتوحيد البيانات. توفر لنا وسائل التواصل الاجتماعي والبيانات الحقيقية كمية هائلة من البيانات بأشكال مختلفة مثل والصور والصوت والفيديو وما إلى ذلك. ركزت هذه الأطروحة على الملفات في شكل صور. تحتوي هذه الملفات المستندة إلى الصور على نصوص تتضمن بيانات قيمة مع معلومات مفيدة للغاية. قد يكون هذا مفتاحًا جيدًا لوصف محتوى الصورة. تعتبر هذه الملفات ذات أهمية كبيرة للشركات ، وخاصة ذكاء الأعمال ، من خلال تحليل واستكشاف البيانات للوصول إلى معلومات قيمة لفهم بيئة العمل والقدرة على التنبؤ بالحالة الجديدة وبالتالي القدرة على التطوير والمنافسة.

الهدف من المشروع هو عمل (brain) لجهاز (OCR) باستخدام تقنيات الذكاء الاصطناعي(AI) وتقنيات التنقيب من خلال الاستفادة من قابليته على عمل مسح ضوئي للنصوص والصور الورقية وقراءتها وتحليلها وتحويلها الى كودات ASCII وبنفس الوقت حل مشكلاته المتمثلة في عدم قدرته على تحديد الكلمات كلمة كلمة وانما قراءة نصوص كاملة بالإضافة الى عدم قدرته على تحويل البيانات الغير مهيكلة الى مهيكلة ، وبالتالي تطوير قدراته لتسهيل الاستفادة منه في بيئة الاعمال(ذكاء الاعمال) وتطوير عملية الارشفة. يتم ذلك من خلال جعل الجهاز قابل على تحديد الكلمات المطلوبة من خلال اضافة مستطيلات حول كل كلمة واعطاءه القابلية على تحويل البيانات الغير مهيكلة الى مهيكلة باستخدام خوارزمية LSTM.

تم تنفيذ هذا العمل باستخدام OCR Tesseract باستخدام مجموعة بيانات تتضمن 10000 صورة على شكل ملفات ممسوحة ضوئيا. يقدم هذا العمل تحسينًا لاستخراج الأنماط المفيدة من المستندات النصية في حقل التنقيب عن النص باستخدام نموذج تصنيف الأنماط باستخدام(PTM) وخوارزمية المسافة ليفينشتين (LDA). اظهرت النتائج ان هذه الطريقة القائمة على النمط قد حققت افضل دقة للأنماط المستخرجة في وقت قصير حيث تم اختبار الخوارزميتين باستخدام قيم عتبة من 1% الى 10% والمقارنة على سمة (global probability) و حقق النظام نتائج واعدة باستخراج الكلمات المتكررة من الاف الصور بالإضافة الى ضمان سلامتها من الخطأ أو التشويش والفقدان .

جمهورية العراق
وزارة التعليم العالي والبحث العلمي
كلية علوم الحاسوب وتكنولوجيا المعلومات
قسم علوم الحاسبات

تطوير الماسح الضوئي لضمان سلامة البيانات الكبيرة غير المهيكلة

رسالة مقدمة الى
مجلس كلية علوم الحاسوب وتكنلوجيا المعلومات
- قسم علوم الحاسبات/ كليةعلوم الحاسوب وتكنولوجيا المعلومات/ جامعة الانبار

وهي جزء من متطلبات نيل درجة الماجستير في علوم الحاسبات

قُدمت من قبل

داليا عامر عبد اللطيف

بأشراف

أ.د. مرتضى محمد حمد

1442 AH　　2021 A.C