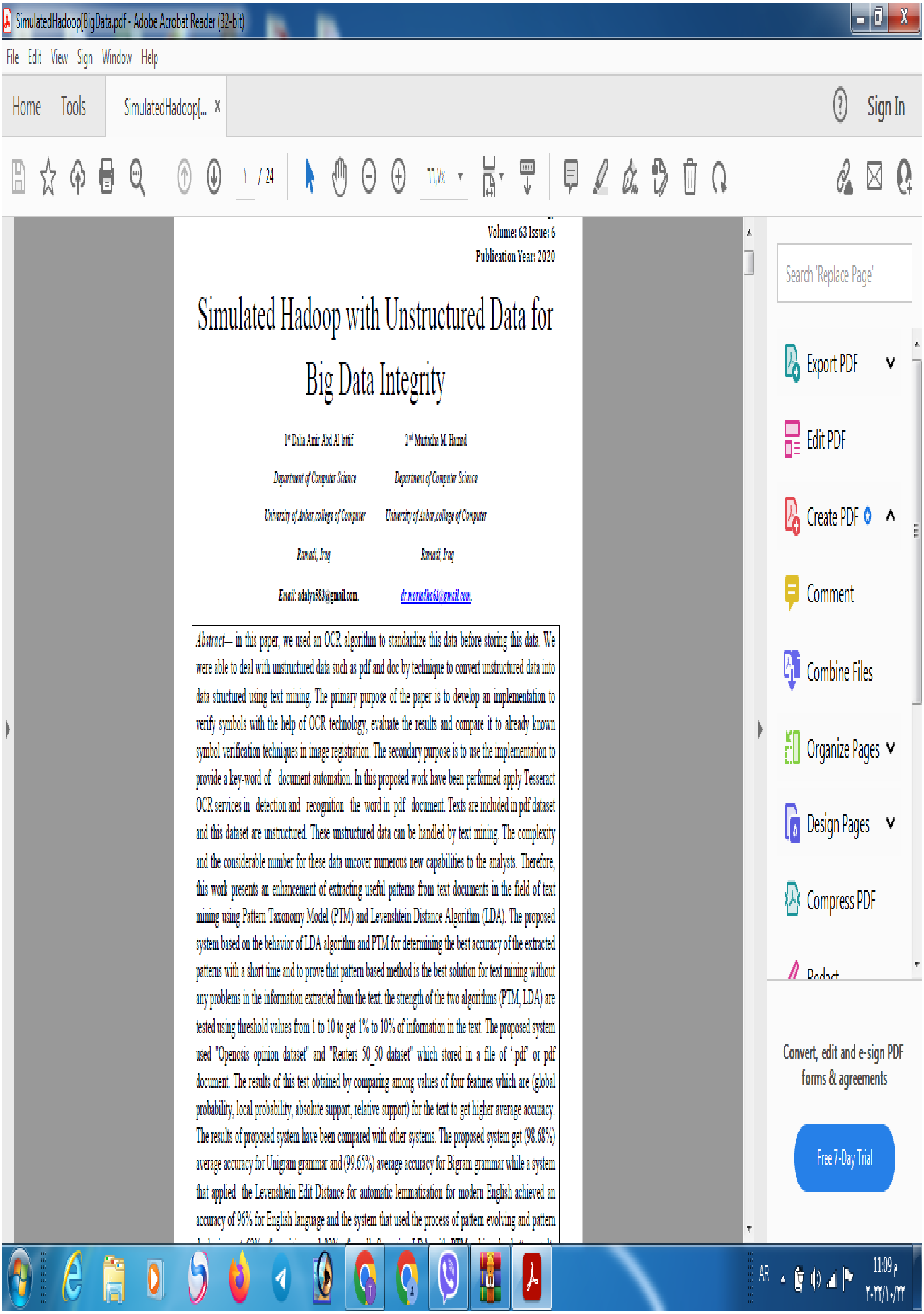
SimulatedHadoop[BigData.pdf - Adobe Acrobat Reader (32-bit)
File Edit View Sign Window Help
Home Tools SimulatedHadoop[...
Sign In
1 / 24
Search 'Replace Page'
Export PDF
Edit PDF
Create PDF
Comment
Combine Files
Organize Pages
Design Pages
Compress PDF
Redact
Convert, edit and e-sign PDF forms & agreements
Free 7-Day Trial
AR
11:09 م
٢٠٢٢/١/٢٢

Volume: 63 Issue: 6

Publication Year: 2020

Simulated Hadoop with Unstructured Data for Big Data Integrity

1st Dalia Amir Abd Al lattif
Department of Computer Science
University of Anbar,college of Computer
Ramadi, Iraq
Email: adalya583@gmail.com.

2nd Murtadha M. Hamad
Department of Computer Science
University of Anbar,college of Computer
Ramadi, Iraq
dr.mortadha61@gmail.com.

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 recognition the word in pdf document. Texts are included in pdf dataset and this dataset are unstructured. These unstructured data can be handled by text mining. The complexity and the considerable number for these data uncover numerous new capabilities to the analysts. Therefore, this work presents an enhancement of extracting useful patterns from text documents in the field of text mining using Pattern Taxonomy Model (PTM) and Levenshtein Distance Algorithm (LDA). The proposed system based on the behavior of LDA algorithm and PTM for determining the best accuracy of the extracted patterns with a short time and to prove that pattern based method is the best solution for text mining without any problems in the information extracted from the text. the strength of the two algorithms (PTM, LDA) are tested using threshold values from 1 to 10 to get 1% to 10% of information in the text. The proposed system used "Openosis opinion dataset" and "Reuters 50_50 dataset" which stored in a file of '.pdf' or pdf document. The results of this test obtained by comparing among values of four features which are (global probability, local probability, absolute support, relative support) for the text to get higher average accuracy. The results of proposed system have been compared with other systems. The proposed system get (98.68%) average accuracy for Unigram grammar and (99.65%) average accuracy for Bigram grammar while a system that applied the Levenshtein Edit Distance for automatic lemmatization for modern English achieved an accuracy of 96% for English language and the system that used the process of pattern evolving and pattern