Optical Character Recognition and Named Entity Recognition for Highly Confidential Documents

Alaa Najmi
Dept. of Computer Science, Arab East Colleges,
Riyadh, Saudi Arabia

Mohamed A. El-Dosuky
Dept. of Computer Science, Arab East Colleges,
Riyadh, Saudi Arabia; and
Faculty of Computers and Information Sciences,
Mansoura University, Egypt

ABSTRACT
Optical character recognition (OCR) is a crucial technique for extracting textual data from various sources, reducing human labor, and enhancing accessibility. Named Entity Recognition (NER) organizes and categorizes data, while Regular expression (Regex) patterning facilitates data extraction from OCR-read text. This technology reduces human labor for extracting large amounts of confidential and sensitive data, improving accessibility and preservation, especially in confidential and sensitive situations. The study utilizes the Tesseract OCR tool and the Marefa-NER NER Model, combining Artificial Neural Networks (ANN), Support Vector Machines (SVM), and Natural Language Processing (NLP) techniques. The technologies have been successfully integrated into websites, and have proven their effectiveness in accurately identifying textual content and categorizing it using OCR, NER, and Regex patterns. The combination of OCR, NER, and Regex pattern matching has proven to be a successful and efficient method for extracting textual information from various sources, reducing human effort and improving accessibility, particularly in cases of confidentiality and sensitivity.

General Terms
OCR, NER, Regex, Confidential documents, Sensitive data.

Keywords
OCR, NER, Regex, Confidential documents, Sensitive data.

1. INTRODUCTION
Character recognition has a long history, predating the development of computers. The first optical character recognition (OCR) systems used mechanical devices to recognize characters, but their processing rates and accuracy were limited. In 1951, M. Sheppard developed the reading and robotic GISMO [1]. To improve OCR recognition, standards for three fonts were implemented, and in 1970, OCRA and OCRB were developed by the American National Standards Institute (ANSI) and Enterprise Manager Configuration Assistant (EMCA). These systems showed notable levels of recognition accuracy. Over the last three decades, significant progress has been made in document image analysis (DIA), including the creation of multilingual, handwritten, and omni-font OCR systems [2]. The Message Understanding Conference (MUC) proposed the notion of Named Entity Recognition (NER) in its sixth iteration [3]. Researchers have continued to categorize alignment into more distinct classifications, such as geographic locations, individuals, and proper nouns [4]. Individuals may be categorized into several groups, such as politicians, entertainers, and groupings [5].

An ANN is a computational model that mimics the human brain's ability to acquire knowledge and adapt to changing environments. The human brain can process and interpret ambiguous information, enabling independent assessments. For example, humans can decipher handwriting, recognize geometric shapes, and distinguish between familiar individuals even in suboptimal images [6].

SVMs are widely used machine learning models for binary classification tasks. They use a scoring system to produce deterministic classification rules, which can be converted into probabilistic rules using SVM libraries. However, setting regularization parameters requires significant processing resources and can result in underutilization of information [7].

Natural Language refers to natural languages used by humans, excluding artificial or constructed ones like computer languages. NLP includes computational approaches for processing natural language using computers. Spoken Natural Language Processing is introduced to include speech and other aspects of NLP, as it is often used without excluding speech [8]. NLP comprises several subfields, including voice synthesis, speech recognition, Natural Language Understanding, and Machine Translation [9].

OCR is a process that converts textual images into machine-readable formats. It can extract and reuse data from various sources, like scanned papers, pictures, and PDF files. OCR software identifies and categorizes individual alphabetic letters in images, organizing them into intelligible words and generating coherent sentences. This process aids in recovering and modifying original content, eliminating the need for manual data input by human operators, and reducing human involvement in data entry tasks. This technology enables the reduction of human involvement in data entry tasks.

NER is the process of naming and categorizing significant data elements in written text. Entities are linguistic units, such as single words or sequences, that consistently denote or refer to a specific object or concept, assigned to a distinct category. Named Entity (NE) refers to proper nouns, including individuals, locations, and organizations. For instance, Mr. Mohammed Okfie, CEO of Digital Nudhj, a firm with its headquarters located in Riyadh, and Alaa Najmi, a software development professional, completed a bachelor's degree at Jazan University. NER is a computational task that detects, extracts, and classifies named entities in unstructured textual data, such as newspaper articles and databases.

Regular expression (Regex) patterns are used to make text patterns that make it easier to get useful information from OCR-extracted text and then organize it. This encompasses a range of data, such as national identification numbers, mobile phone numbers, emails, and several other identifiable patterns.

Confidential documents and papers are considered highly confidential because they contain a wide range of confidential information about security issues and people's identities. To enhance its confidentiality, an OCR text scanning service is used to convert scanned documents into searchable text. Arabic is a major focus due to its large vocabulary, ensuring that all relevant information is shared during research projects. Security data acquisition is sensitive, but NER technologies and Regex patterns can help categorize spoken and written information for better prediction. This process creates interactive security cards, providing a comprehensive framework for data extraction and use. Administrators can establish connections across diverse data sources, enabling informed decision-making. The development of connections between information sources, especially when they originate from different sources, is crucial for accurate classification.

Texts written in Arabic often have complicated morphological structures. This is because Arabic is an inflectional language, which means it has a lot of different morphological types. A range of agglutination strategies may be used to generate diverse lexical forms. The issue pertaining to morphology has been effectively tackled in several NLP tasks and applications. Some of these are machine translation [10], extracting noun compounds [11], clearing up word meanings [12], figuring out how semantically related words are [13], and mapping lexical sources [14]. The trial's results indicated that the use of stemming might potentially improve the performance and capabilities of traditional NLP methods. The benefits of nationality include a combination of lexical and contextual attributes, which may be classified into several groups. For example, (وصل ولي العهد السعودي الأمير محمد بن سلمان آل سعود إلى الرياض) the arrival of Crown Prince Mohammed bin Salman Al Saud in Riyadh, The advantages of having a certain nationality One approach to determining the inclusion of a word on the Nationality List is by using a binary feature. To ascertain if it is included in the Nationality List or not [15]. The use of Arabic, Hindi, and other similar characters, along with font styles like italics and others that have overlapping properties, makes it very hard for OCR systems to correctly identify and separate individual characters during the segmentation process [16]. The complexity of OCR arises from the diverse array of languages, font kinds, writing styles, and precise linguistic laws involved in the process. Consequently, several techniques derived from multiple disciplines within computer science, including image processing, pattern recognition, and NLP, have been used [17]. Auxiliary Vowels: The Arabic language has a variety of diacritical marks that serve as indicators for vowels, altering the semantic interpretation of individual words. Consequently, modifying the diacritics associated with words might lead to the acquisition of entirely different meanings. As an example, the word "Noor-" might be associated with a feminine name, the verb "enlightenNooar," or the genuine name "Noor-light" [18]. Capitalization in Arabic orthography is challenging due to the absence of capital letters for proper names, unlike in English, which uses Latin script. This ambiguity between proper nouns and common nouns or descriptive terms in Arabic contexts makes it difficult to identify named non-NEs as individual words or word combinations. A system relying solely on noun dictionaries would lack certainty [11].

The challenge lies in handling large amounts of information in files, which requires methodologies for extracting, manipulating, and structuring important material. This task is often assigned to personnel within a facility, which may be vulnerable to unauthorized access. OCR technology helps in extracting text while restricting access to authorized personnel, enhancing document security, and maintaining information confidentiality. The integration of OCR and NER technology has significantly impacted various aspects of life, work, and occupational duties, enabling efficient data collection and faster search procedures.

2. METHODOLOGY

The system has a hybrid architecture, including a PHP-based backend with Python-based modules responsible for OCR, NER, and Regex pattern analysis. Microsoft SQL Server functions as a Relational Database Management System (RDBMS) that facilitates the tasks of data administration, storage, and retrieval. The essential elements include a user interface developed in PHP, an OCR engine, a NER model, and bespoke Regex pattern matching algorithms. The proposed project encompasses an enhanced file-based software program designed for Information Extraction (IE).

This application is specifically tailored to extract various types of information, including but not limited to national identity numbers, mobile phone numbers, email addresses, individuals, entities, business establishments, and government organizations. The tool categorizes textual information into organized and structured data, then integrates it into a Database (DB) and facilitates content retrieval via search functionality. Additionally, this software facilitates the conversion of Word documents, pictures, Excel spreadsheets, and PowerPoint presentations into PDF files independently from their respective source files.

The functionality of the system is unaffected by the size of the file or the magnitude of the information encountered. The software process suggested is shown in Figure 1. The depicted Add Subjects Workflow, as seen in Figure 2, involves users accessing a web-based platform, selecting subjects from pre-existing lists, reviewing all available subjects, inputting a novel subject name, and uploading or scanning pertinent documents. The act of eliciting a response in the form of a "thank you" or an "error" message The process of OCR is shown in Figure 3 via an activity diagram. This diagram showcases the verification of file types that occurs throughout the user upload or scanning phase. The system verifies the adherence of the file to the PDF format, examines files such as Word, Excel, Picture, or PowerPoint, and stores the resultant data in a database table. The utilization-case diagram illustrating the NER procedure, as shown in Figure 4 below, After OCR, the NER method is used to consistently arrange and classify the results into structured outcomes. These outcomes are then recorded in a database table. The use of Regex patterns occurs subsequent to the completion of the OCR process. This enables the system to effectively index, categorize, and store the obtained results in a table inside the database. This use-case of Regex patterns is shown in Figure 5. Figure 6 is a sequence diagram illustrating the process of doing a site search. The procedure entails accessing the search page, inputting the topic name or OCR text, and then activating the search button.

.

Fig 1: Proposed software workflow

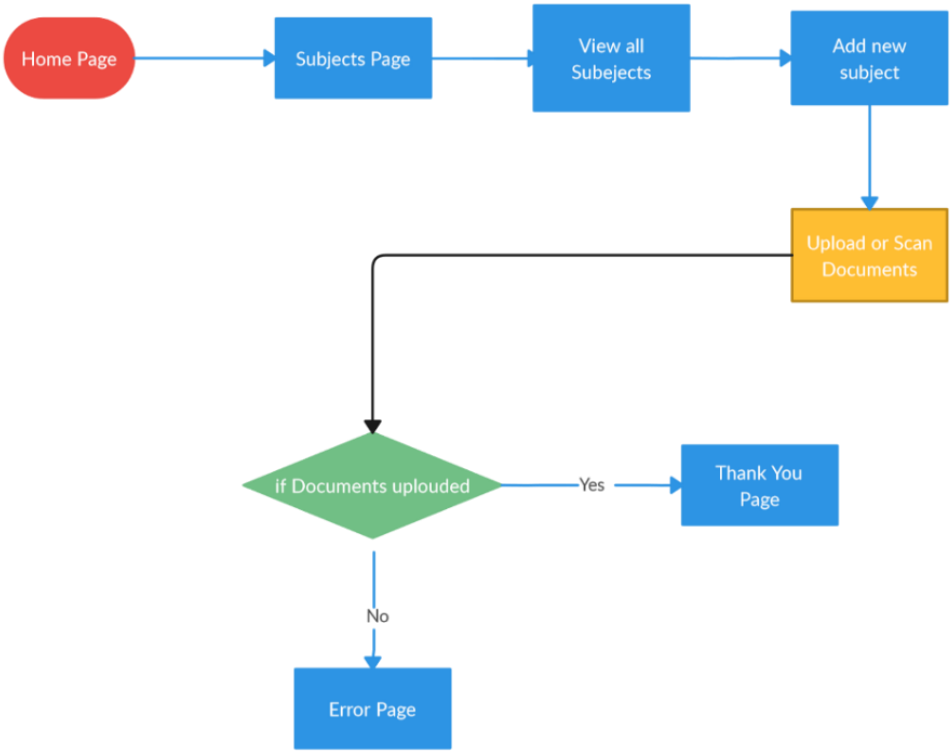

Fig 2: Add Subjects Workflow

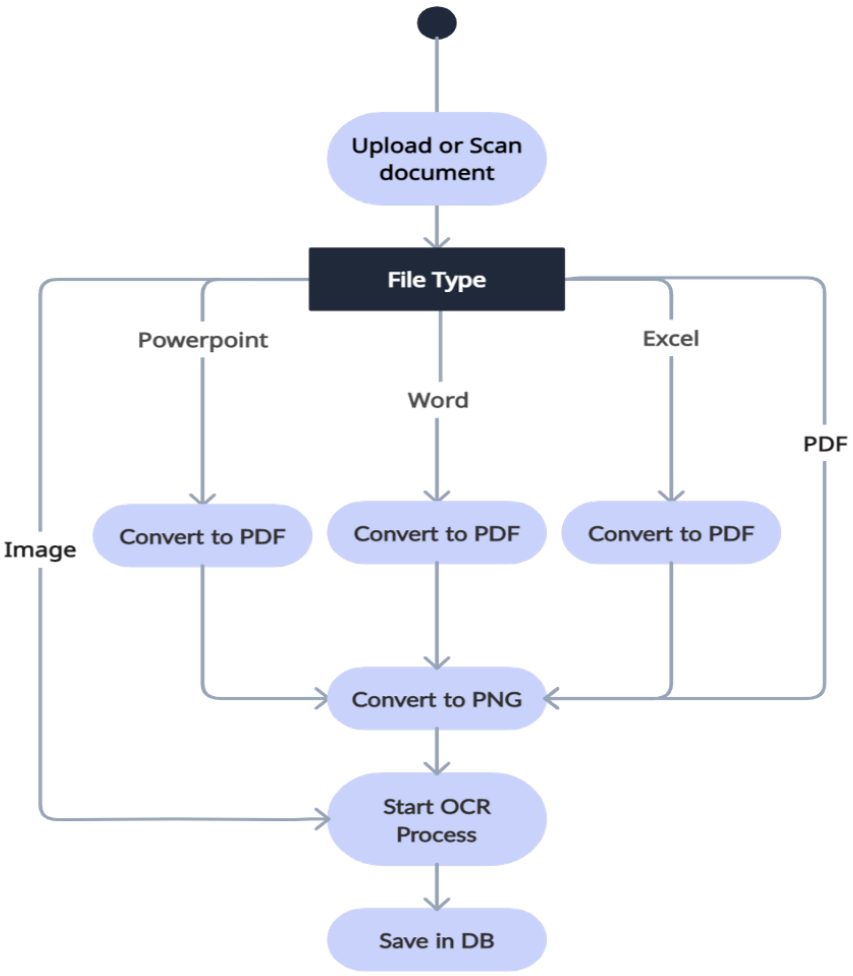

Fig 3: OCR process

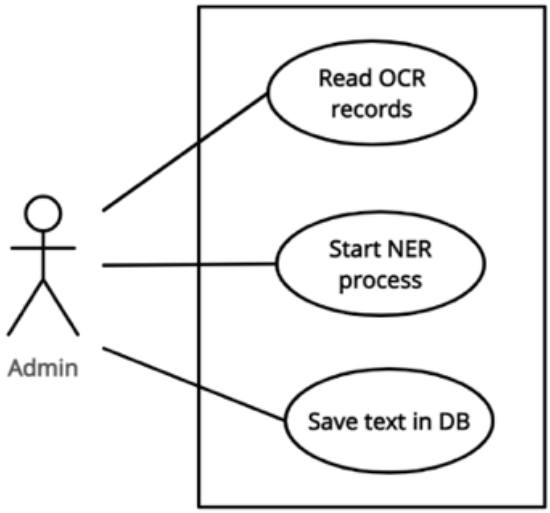

Fig 4: NER process

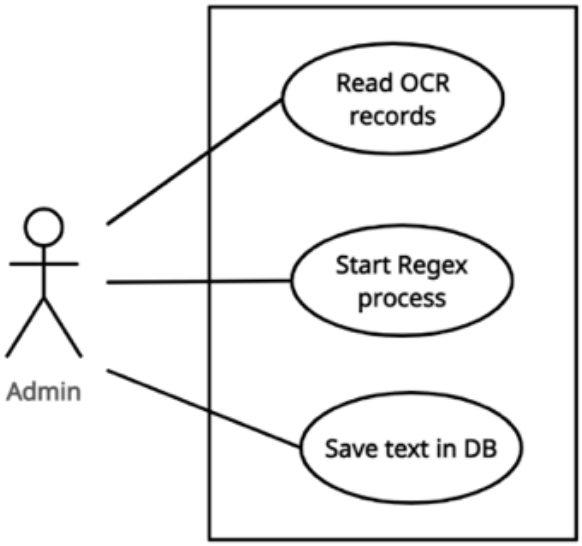

Fig 5: Regex patterns process

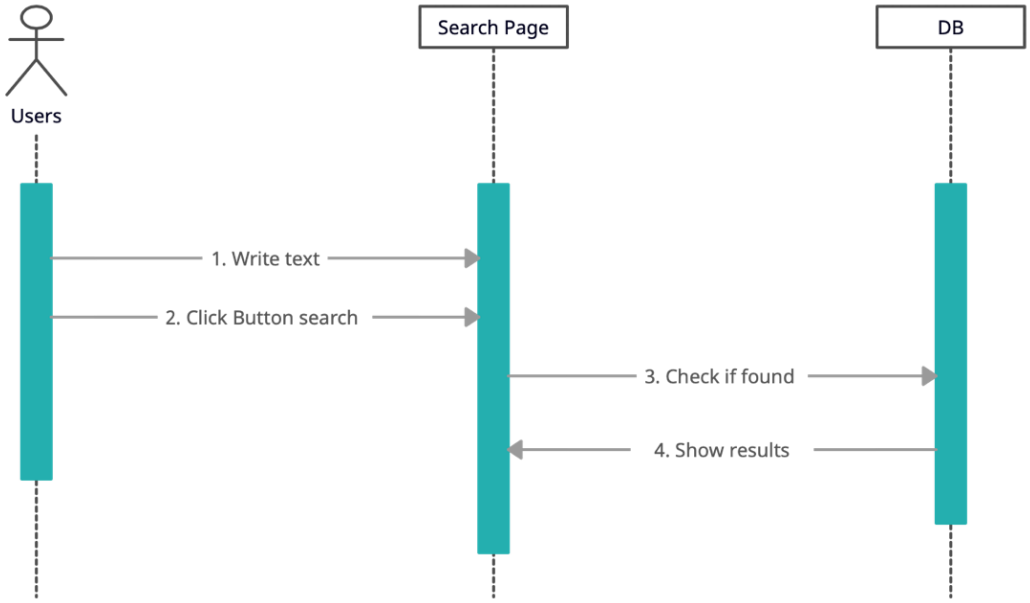

Fig 6: Sequence diagram showing how to search

Table 1. Measure of Marefa Arabic NER Model

Type	F1-score	Precision	Recall	Support
Artwork	0.653552	0.678005	0.630802	474
Event	0.686695	0.733945	0.645161	744
Job	0.837656	0.79912	0.880097	2477
Location	0.891537	0.896926	0.886212	4939
Nationality	0.871246	0.843153	0.901277	2350
Organization	0.781317	0.773328	0.789474	2299
Person	0.93298	0.931479	0.934487	4335
Product	0.625483	0.553531	0.718935	338
Time	0.873003	0.876087	0.869941	1853

Table 2. Features of Tesseract engine

Feature	Tesseract
Accuracy	98.86%
Average Time	1 second (color images) 0.82 Seconds (gray scale images)
Languages Supports	+100
Operating System	Window, MAC, Linux
Open Source	Yes
Online	No
License	Apache
Cost	Free

The present research uses the OCR engine, namely Tesseract, in conjunction with the NER Model, known as Marefa-NER. The Marefa-NER model incorporates ANN, SVM, and NLP techniques. The primary purpose of this task is to discern and classify pre-established items inside written information into prearranged groupings. It is worth mentioning that the form has the capability to support languages other than Arabic. The measurement test set for the Marefa Arabic NER Model [19] is shown in Table 1 in great detail. It is made up of 1959 sentences. Table 2 presents a complete summary of the primary characteristics linked to the Tesseract engine.

3. RESULTS AND DISCUSSION

The OCR, NER, and Regex patterns website project has been meticulously developed and tested, ensuring the platform's readiness for production. Key components include the file upload system, OCR module, and DB integration, which have been integrated seamlessly, resulting in improved data consistency and accessibility. The platform has also been enhanced with user-friendly functionalities like document scanning, subject management, and document retrieval, enhancing the overall User Experience (UX) and accessibility. These tests ensure the system works correctly, retrieves and displays data accurately, can handle user interactions, meets performance standards, and protects against security threats. The website is now ready for deployment, providing users with a reliable, streamlined, and safeguarded platform to fulfill their OCR and regex pattern requirements. The reliability and readiness for real-world application are emphasized by the combination of thorough testing and strong implementation.

The experiment evaluates the duration required to convert a singular picture to OCR at several dpi settings, with the objective of quantifying the mean time necessary for effective document processing. The shown data represents the time required for the conversion process of a typical PDF document to an OCR of 5 pages. It is shown how long it takes to convert Word documents to PDF format and then perform OCR, with an average page count of 5. The duration required for the conversion process of PPT to PDF to OCR is calculated, specifically focusing on the conversion of 5 slides. It is shown that the duration required for the conversion process of Excel to PDF to OCR, specifically for a spreadsheet including 100 rows and 8 columns is fast.

4. CONCLUSION

The process of extracting text from documents in a facility is complicated, mostly because the documents contain sensitive information and strict security measures must be put in place. The use of OCR technology facilitates the extraction of textual information while concurrently limiting access only to those with proper authorization. The establishment of limited entry for a specific group is crucial for maintaining the authenticity of data, reducing the potential for illegal dissemination, bolstering the protection of documents, and guaranteeing confidentiality.

This study introduces a web-based application that utilizes OCR, NER, and Regex pattern technology to extract and categorize data from texts written in both Arabic and English. The OCR engine is responsible for the conversion of Arabic text into a format that can be easily interpreted by machines. On the other hand, the NER modules and Regex pattern are used to identify and categorize significant components within the text. The project underwent testing on a diverse range of research articles written in both Arabic and English. The findings obtained from these tests demonstrated the project's efficacy in extracting and categorizing fundamental information. This application exhibits the capacity to enhance the efficiency of IE procedures.

5. REFERENCES

[1] Satti, Danish Altaf. "Offline Urdu Nastaliq OCR for printed text using analytical approach." MS thesis report (2013): 141.

[2] Al-Badr, Badr, and Sabri A. Mahmoud. "Survey and bibliography of Arabic optical text recognition." Signal processing 41, no. 1 (1995): 49-77.

[3] Grishman, Ralph, and Beth M. Sundheim. "Message understanding conference-6: A brief history." In COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics. 1996.

[4] Lee S, Lee G, 2005, Proceedings of the International Joint Conference on Natural Language Processing, October 11-13, 2005: Heuristic Methods for Reducing Errors of Geographic Named Entities Learned by 68 Volume 6; Issue 5 Bootstrapping. Springer Verlag, Jeju Island, Korea, 658-669.

[5] Liu, Xing, Huiqin Chen, and Wangui Xia. "Overview of named entity recognition." Journal of Contemporary Educational Research 6, no. 5 (2022): 65-68.

[6] Kukreja, Harsh, N. Bharath, C. S. Siddesh, and S. Kuldeep. "An introduction to artificial neural network." Int J Adv Res Innov Ideas Educ 1 (2016): 27-30.

[7] Benítez-Peña, Sandra, Rafael Blanquero, Emilio Carrizosa, and Pepa Ramírez-Cobo. "Cost-sensitive probabilistic predictions for support vector machines." European Journal of Operational Research (2023).

[8] Hannan, Shaikh Abdul, Jameel Ahmed, Naveed Ahmed, and Rizwan Alam Thakur. "Data Mining and Natural Language Processing Methods for Extracting Opinions from Customer Reviews." International Journal of Computational Intelligence and Information Security: 52-58.

[9] Sætre, Rune. "GeneTUC: Natural Language Understanding in Medical Text." (2006).

[10] Zollmann, Andreas, Ashish Venugopal, and Stephan Vogel. "Bridging the inflection morphology gap for Arabic statistical machine translation." In Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers, pp. 201-204. 2006.

[11] Saif, Abdulgabbar Mohammed, and Mohd Juzaiddin Ab Aziz. "An automatic noun compound extraction from Arabic corpus." In 2011 International Conference on Semantic Technology and Information Retrieval, pp. 224-230. IEEE, 2011.

[12] Zouaghi, Anis, Laroussi Merhbene, and Mounir Zrigui. "Combination of information retrieval methods with LESK algorithm for Arabic word sense disambiguation." Artificial Intelligence Review 38, no. 4 (2012): 257-269.

[13] A. Saif, M. J. Ab Aziz, and N. Omar, "Evaluating knowledge-based semantic measures on Arabic," International Journal on Communications Antenna and Propagation, vol. 4, pp. 180-194, 2014.

[14] Saif, Abdulgabbar, Mohd Juzaiddin Ab Aziz, and Nazlia Omar. "Mapping Arabic WordNet synsets to Wikipedia articles using monolingual and bilingual features." Natural Language Engineering 23, no. 1 (2017): 53-91.

[15] Alshaikhdeeb, Basel, and Kamsuriah Ahmad. "Biomedical named entity recognition: a review." International Journal on Advanced Science, Engineering and Information Technology 6, no. 6 (2016): 889-895.

[16] Awel, Muna Ahmed, and Ali Imam Abidi. "Review on optical character recognition." International Research Journal of Engineering and Technology (IRJET) 6, no. 6 (2019): 3666-3669.

[17] Islam, Noman, Zeeshan Islam, and Nazia Noor. "A survey on optical character recognition system." arXiv preprint arXiv:1710.05703 (2017).

[18] Salah, Ramzi Esmail, and L. Qadri binti Zakaria. "A comparative review of machine learning for Arabic named entity recognition." International Journal on Advanced Science, Engineering and Information Technology 7, no. 2 (2017): 511-518.

[19] Marefa Arabic Named Entity Recognition Model (huggingface.co/marefa-nlp/marefa-ner), Last access 2023/02/08.