OBJECT CHARACTER RECOGNITION FOR AUTOMATIC LABELLING OF PHARMACEUTICAL PRODUCTS

MUHAMMAD HANAFI AKMAL BIN ABDUL RAHMAN

UNIVERSITI TEKNOLOGI MALAYSIA

OBJECT CHARACTER RECOGNITION FOR AUTOMATIC LABELLING OF PHARMACEUTICAL PRODUCTS

MUHAMMAD HANAFI AKMAL BIN ABDUL RAHMAN

A project report submitted in partial fulfilment of the
requirements for the award of the degree of
Master of Engineering (Computer And Microelectronic Systems)

School of Electrical Engineering
Faculty of Engineering
Universiti Teknologi Malaysia

JULY 2022

DEDICATION

This thesis is dedicated to my father, who taught me that the best kind of knowledge to have been that which is learned for its own sake. It is also dedicated to my mother, who taught me that even the largest task can be accomplished if it is done one step at a time.

ACKNOWLEDGEMENT

In preparing this thesis, I was in contact with many people, researchers, academicians, and practitioners. They have contributed towards my understanding and thoughts. In particular, I wish to express my sincere appreciation to my main thesis supervisor, Dr. Zaid Bin Omar, for encouragement, guidance, critics, and friendship. Without his continued support and interest, this thesis would not have been the same as presented here.

My fellow postgraduate student should also be recognised for their support. My sincere appreciation also extends to all my colleagues and others who have helped on various occasions. Their views and tips are useful indeed. Unfortunately, it is not possible to list all of them in this limited space. I am also grateful to all of my family members for their encouragement throughout my journey.

ABSTRACT

In the current modern era, storing data information from images or documents to a computer drive is in high demand as it can be utilized the information for various purposes, especially in the pharmaceutical industry. The current method of storing data information about pharmaceutical products is to manually key-in the information about the products to the computer system. Therefore, one simple method for storing information from documents on a computer system would be to scan the image or document and then save it as an image file. However, analysing this information from the image can be exceedingly difficult. There is a need for dependable manual labour to review the information on pharmaceutical products. For this reason, a method to automatically fetch and store the information from the image is required. Object Character Recognition (OCR) is a well-known method that can identify and process information from pixel-based images to text format. In this thesis, OCR is implemented to extract text characters from images for the labelling of pharmaceutical products. The challenges that are associated with this task include variances in illumination, rotation when acquiring the image, and the different fonts that are shown on the pharmaceutical product. Besides, there is too much information for the computer system to accurately retrieve from the images. In addition, Named Entity Recognition (NER) is implemented to identify the important information from the OCR process. The system successfully extracts all the important information for several pharmaceutical products and successfully converts them into a sample form. The results obtained by OCR show a 92.85% accuracy rate. Meanwhile, the results obtained by NER have a 100% accuracy rate for MAL numbers and a 90% accuracy rate for product names. Overall, it is hoped that this system may help to optimize the work in the pharmaceutical supply chain industry and contribute towards the national industry.

ABSTRAK

Dalam era moden sekarang, penyimpanan maklumat data daripada imej atau dokumen ke pemacu komputer adalah permintaan yang tinggi kerana ia boleh digunakan untuk pelbagai tujuan terutamanya dalam industri farmaseutikal. Kaedah semasa menyimpan maklumat data tentang produk farmaseutikal adalah dengan memasukkan maklumat produk secara manual ke sistem komputer. Oleh itu, satu kaedah mudah untuk menyimpan maklumat daripada dokumen pada sistem komputer adalah dengan mengimbas imej atau dokumen. Kemudian, disimpan sebagai fail imej. Walau bagaimanapun, menganalisis maklumat ini daripada imej boleh menjadi sukar. Terdapat keperluan untuk buruh manual yang boleh dipercayai untuk menyemak maklumat mengenai produk farmaseutikal. Oleh kerana itu, kaedah untuk mengambil dan menyimpan maklumat daripada imej secara automatik diperlukan. Pengimbas Karakter Objek (OCR) ialah kaedah terkenal yang boleh mengenal pasti dan memproses maklumat daripada imej berasaskan piksel kepada format teks. Dalam tesis ini, OCR dilaksanakan untuk extrak aksara teks daripada imej untuk pelabelan produk farmaseutikal. Cabaran yang dikaitkan dengan tugas ini termasuk variasi dalam pencahayaan, imej diperoleh apabila putaran berlaku dan perbezaan fon yang ditunjukkan pada produk farmaseutikal. Selain itu, terdapat banyak maklumat untuk sistem komputer yang diambil dengan tepat daripada imej. Tambahan pula, Pengimbasan Entiti Nama (NER) dilaksanakan untuk mengenal pasti maklumat penting daripada proses OCR. Sistem ini, berjaya mengekstrak semua maklumat penting untuk beberapa produk farmaseutikal dan berjaya ditukarkan ke dalam bentuk sampel. Keputusan yang diperoleh daripada OCR menunjukkan kadar ketepatan sebanyak 92.85%. Sementara itu, keputusan yang diperolehi oleh NER mempunyai kadar ketepatan sebanyak 100% untuk nombor MAL dan kadar ketepatan sebanyak 90% untuk nama produk. Secara keseluruhan, harappanya sangat besar agar system ini dapat membantu mencapai tahap kerja yang optimum dalam industri rantaian bekalan farmaseutikal dan menyumbang kepada industri negara.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANPR	-	Automated Number Plate Recognition
CPU	-	Central Processing Unit
CRAFT	-	Character Region Awareness for Text detection
CTC	-	Connectionist Temporal Classification
CW	-	Complex Wavelet
DIA	-	Document Image Analysis
DNA	-	Deoxyribonucleic Acid
GPS	-	Global Positioning System
GUI	-	Graphical User Interface
HVS	-	Human Visual System
LSTM	-	Long Short-Term Memory
MOH	-	Ministry of Health
MSE	-	Mean Squared Error
NER	-	Named Entity Recognition
OCR	-	Object Character Recognition
ResNet	-	Residual Network
PLSB	-	Pharmaniaga Logistic Sdn. Bhd.
PSC	-	Pharmaceutical Supply Chain
RNN	-	Recurrent Neural Network
SSIM	-	Structural Similar Index Measure

LIST OF SYMBOLS

c	-	Contrast
l	-	Luminance
lev	-	Levenshtein Distance
s	-	Structure

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Background of Study

Pharmaniaga Logictics Sdn. Bhd. (PLSB) performs a critical role in the medicinal supply chain in Malaysia comprising storage, distribution, and delivery of medication and health products. The company oversees the annual tender and sampling procedure involving thousands of products submitted from suppliers for approval. In this process, essential information on each product such as name of medicine, dosage, and expiry date are all recorded for documentation. The task is currently performed by manual hands, which can be cumbersome, time-consuming, and prone to error. It is natural for us to want to build and develop devices which can recognise patterns. Machine pattern recognition would be immensely advantageous in a variety of applications, including automated optical character recognition, facial features recognition, iris identification, voice recognition, DNA biometrics identification, and many more.

The goal of object character recognition research is to develop a computer system which is capable of autonomously extracting and analysing text from images. There is a significant interest these days for collecting data on a computer storage device from available data in documentation or handwritten materials in attempt to re-use this data using computers. One straightforward method for transferring information from paper documents to a computer system would be to scan the pages and then save them as image files. However, it would be extremely challenging to extract text or other information from these image files for re-use this information. As a result, a method for automatically retrieving and storing information, particularly text, from image files is required. By all means, this is not a simple task. To accomplish successful automation, some main problems must be identified and addressed [1]. The font properties of characters in paper documents, as well as image quality, are just a

few of the latest issues. Characters may not be recognised correctly by the computer system because of these difficulties. As a result, character recognition techniques are necessary to carry out Document Image Analysis (DIA), which eliminates such obstacles and generates electronic format from paper-converted documents.

Alternatively, Object Character Recognition (OCR) is a technique of converting any type of text or text-containing document, such as handwriting, printed text, or scanned pictures, into an editable digital medium for further processing. Object character recognition system enables a system to automatically recognise text in such documents. In the real world, it is analogous to the union of the human intellect and vision. The human eye can identify, evaluate, and extract information from images. However, the human brain evaluates the text that the eye can be detects or extracts [2]. In fact, OCR technology has yet to improve further to match human capabilities. The quality of input documents has a direct impact on the accuracy and performance of OCR. The performance of the brain's process is closely related to the quality of the information read by eye when it comes to human text recognition. Several obstacles and challenges can arise in designing and implementing a computerised OCR system. Some numerals and letters vary only slightly enough for systems to accurately recognise and identify one from the others. Computers, for example, may find it difficult to differentiate between the integer "0" and the character "o," or between "8" and "B," especially if such characters are contained in a very dark and noisy environment. The main goals of OCR studies have been to recognise cursive characters and handwritten text for a variety of applications. There are several types of OCR software available today to handle the text recognition problem, including desktop OCR, server OCR, web OCR, and so on.

In addition, to extract important information from the OCR process, in applications such as information extraction, question answering, and machine translation, Named Entity Recognition (NER) plays a significant role. Named entities, also known as NEs, are words or phrases that have been designated with a name or assigned a category in relation to a certain subject. They often hold important information inside a sentence, which most language processing algorithms use as important targets. Different Natural Language Processing (NLP) applications can

make effective use of accurate named entity recognition as a helpful source of information [3]. However, the challenge of recognising named entities may be broken down into two distinct parts which are identifying named entity borders and identifying named entity categories. The solutions to these issues are typically found simultaneously, however this is not always the case. There are inconsistencies in the language, which add to the difficulty of the process. This is similar to the majority of the issues that arise while processing language. For example, there is a degree of uncertainty regarding the identification of the designated entity "Leading," since it is possible for it to be misunderstood either as a gerund form of a verb or as a proper noun. The extensively lexicalized and domain-dependent nature of NER presents the majority of its issues. Names constitute a sizable portion of any language and are subject to ongoing change across a variety of contexts.

1.2 Problem Statement

The Sampling Division of PLSB is responsible for the sampling and documentation of over 10,000 pharmaceutical and medical products nationwide. Sampling is considered an important aspect of quality control in the pharmaceutical and healthcare industry, where manufacturers are obliged by the Ministry of Health to comply with a strict set of standards for their products. For example, the dosage for a box of medication may not be altered without prior approval from the Ministry. It is the mandate of PLSB to monitor all products, ensure their quality, and detect any changes through its sampling exercise. This leads to time consuming and needs extra manpower to do sampling and documentation of the pharmaceutical and medical products by PLSB.

In addition, since OCR study is a current and essential topic in general pattern recognition challenges, and substantial field research is required on a recurring basis to stay up with new discoveries. OCR algorithms require high quality or high-resolution pictures with certain basic structural qualities such as good distinguishing text and background for high quality and accuracy character recognition. The method used to capture an image is a critical and deciding factor in the reliability and performance of OCR, as it frequently has a considerable impact on image quality. OCR

utilizing pictures produced by scanners often yields great accuracy and performance. Camera photos, on the other hand, are frequently not as good as scanned images for OCR because of environmental or camera-related issues.

In a typical area, we can see a huge number of man-made items, such as paintings, buildings, and symbols, in camera-captured images. Text detection in the processed image is exceedingly tough due to the comparable structures and features of these items to text. To improve decipherability, the text is consistently laid out. The challenge with image complexity is that the surrounding environment makes differentiating non-text to text very challenging [1].

In addition, taking photographs in natural settings frequently results in imbalanced lighting and darkness. This is a difficulty for OCR since it can reduce the desirable image qualities, resulting in less accurate detection, segmentation, and identification outcomes [1]. This quality of uneven illumination separates a captured image from one generated by a camera. Because of the lack of such differences in lighting and darkness, scanned pictures are chosen over camera photographs due to its superior features and quality. Although employing an on-camera flash can solve difficulties with uneven lighting, it also creates new challenges.

1.3 Research Objectives

In order to achieve the goal, the objectives of this thesis are described below:

(a) To develop an Object Character Recognition technique that able to extract all the text information from the pharmaceutical products.

(b) To develop a Named Entity Recognition technique that able can identify the important information in OCR process.

(c) To implement the performance of the system in Graphical User Interface (GUI).

1.4 Scope of Study

This thesis concentrates on development of object character recognition system for automatic labelling of pharmaceutical products. Prior to that, the OCR will focus on the extracting the information of pharmaceutical products which come in 10 different types of pharmaceutical products. In addition, five important information need to identify from pharmaceutical products such as reference number, MAL number, product name, company name and active ingredient by using NER technique. This project collaborates with Pharmaniaga Logistic Sdn. Bhd (PLSB) and provide the image of pharmaceutical products and NPRA file as a database. For the purpose of this database is to apply in training the model for feature selection and parameter estimation by using Python program.

1.5 Significance of Study

The thesis presents a significant shift for Pharmaniaga into the Industrial Revolution 4.0 and serves as a potent collaboration between the industry, academia, and society. As potentially the bulk of the recognition task on product information will be executed by machine-based algorithms, Pharmaniaga stands to benefit from faster and less labor-intensive product sampling procedures, which in turn may save considerable operational costs.

1.6 Outline of Thesis

This thesis is divided into five chapters, the first of which is an introduction to the thesis. The first chapter describes the study's background, problem statements, objectives, scope of study, and significance. This sub-topic will give the reader a basic overview of the object character recognition system. In Chapter 2, present the object character recognition system literature review in terms of the theoretical background for this thesis. Furthermore, Chapter 3 represents the approach of this thesis and gives a guideline framework with specific algorithms, methods, or techniques used in the

thesis' implementation. This study continues with Chapter 4, which describes the topic's results and discussions. Finally, Chapter 5 provides summaries of the conclusions in accordance with the objectives of the research, as well as some recommendations for further work on the thesis.

REFERENCES

[1] Q. Ye, D. Doermann D, "Text detection and recognition in imagery: A survey". IEEE transactions on pattern analysis and machine intelligence, vol. 37, no. 7, pp. 1480-500, July 2015.

[2] C. Patel, A. Patel, D. Patel, "Optical character recognition by open source OCR tool tesseract: A case study". International Journal of Computer Applications, vol. 55, no. 10, pp. 50-56, January 2012.

[3] B. Mohit, "Named Entity Recognition," in Natural Language Processing of Semitic Languages, Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 221–245.

[4] G. Mehralian, F. Zarenezhad, and A. Rajabzadeh Ghatari, "Developing a model for an agile supply chain in pharmaceutical industry," International Journal of Pharmaceutical Healthcare Marketing, vol. 9, no. 1, pp. 74–91, 2015.

[5] A. Moosivand, A. Rajabzadeh Ghatari, and H. R. Rasekh, "Supply chain challenges in pharmaceutical manufacturing companies: Using qualitative system dynamics methodology," Iranian Journal of Pharmaceutical Research, vol. 18, no. 2, pp. 1103–1116, 2019.

[6] N. Baharuddin, "About the Drug Control Authority (DCA)," Gov.my, 11-Oct-2021. [Online]. Available: https://www.npra.gov.my/index.php/en/about/drug-control-authority-dca/about-the-dca. [Accessed: 15-April-2022].

[7] O. Aigbogun, Z. Ghazali, and R. Razali, "A Framework to Enhance Supply Chain Resilience The Case of Malaysian Pharmaceutical Industry," Global Business and Management Research: An International Journal, vol. 6, no. 3, pp. 219-228, January 2014.

[8] "Pharmaniaga," Pharmaniaga.com. [Online]. Available: https://pharmaniaga.com/services/logistics-distribution/malaysia-logistics-distribution/. [Accessed: 15-April-2022].

[9] J. M. Sharp, Z. Irani, and S. Desai, “Working towards agile manufacturing in the UK industry,” International Journal of Production Economics, vol. 62, no. 1–2, pp. 155–169, 1999.

[10] P. M. Swafford, S. Ghosh, and N. N. Murthy, “A framework for assessing value chain agility,” International Journal of Operation and Production Management, vol. 26, no. 2, pp. 118–140, 2006.

[11] A. Chaudhuri, K. Mandaviya, P. Badelia, and S. K. Ghosh, “Soft computing techniques for optical character recognition systems,” in Optical Character Recognition Systems for Different Languages with Soft Computing, Cham, Springer International Publishing, vol. 352, pp. 43–83, 2017.

[12] A. Mutholib, T. S. Gunawan, and M. Kartiwi, “Design and implementation of automatic number plate recognition on android platform,” in 2012 International Conference on Computer and Communication Engineering (ICCCE), pp. 540-543, July 2012.

[13] B. Verma, P. Gader, and W. Chen, “Fusion of multiple handwritten word recognition techniques,” Pattern Recognition. Lett., vol. 22, no. 9, pp. 991–998, 2001.

[14] I. Shamsher, Z. Ahmad, J. K.Orakzai, A. Adnan, “OCR for printed urdu script using feed forward neural network.” In Proceedings of World Academy of Science, Engineering and Technology, vol. 23, pp. 172-175, Aug 2007.

[15] A. Rehman and T. Saba, “Performance analysis of character segmentation approach for cursive script recognition on benchmark database,” Digit. Signal Process., vol. 21, no. 3, pp. 486–490, 2011.

[16] X. Zhai, F. Bensaali, and R. Sotudeh, “OCR-based neural network for ANPR,” in 2012 IEEE International Conference on Imaging Systems and Techniques Proceedings, pp. 393-397, 2012.

[17] Yetirajam M, Nayak MR, Chattopadhyay S. “Recognition and classification of broken characters using feed forward neural network to enhance an OCR solution.” International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), vol. 1, no. 8, pp. 11-15, Oct 2012.

[18] Y. Kang, Z. Cai, C.-W. Tan, Q. Huang, and H. Liu, “Natural language processing (NLP) in management research: A literature review,” J. manag. anal., vol. 7, no. 2, pp. 139–172, 2020.

[19] D. Lawrie, J. Mayfield, D. Etter, “Building OCR/NER Test Collections,” Proceeding of the 12th Conference on Language Resources and Evaluation (LREC 2020), pp. 4639-4646, May 2020.

[20] D. Nadeau and S. Sekine, “A survey of named entity recognition and classification,” in Benjamins Current Topics, Amsterdam: John Benjamins Publishing Company, 2009, pp. 3–28, August 2007.

[21] R. Salah, M. Mukred, L. Qadri binti Zakaria, R. Ahmed, and H. Sari, “A new rule-based approach for Classical Arabic in natural language processing,” J. Math., vol. 2022, pp. 1–20, 2022.

[22] K. Shaalan and H. Raza, “NERA: Named entity recognition for Arabic,” Journal of American Society Information Science and Technology, vol. 60, no. 8, pp. 1652–1663, April 2009.

[23] S. K. Saha, S. Narayan, S. Sarkar, and P. Mitra, “A Composite Kernel For Named Entity Recognition,” Pattern Recognition Letters, vol. 31, no. 12, pp. 1591–1597, September 2010.

[24] N. Freire, J. Borbinha, and P. Calado, “An approach for named entity recognition in poorly structured data,” in Lecture Notes in Computer Science, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 718–732, 2012.

[25] R. Alfred, L. C. Leong, C. K. On, and P. Anthony, “Malay named entity recognition based on rule-based approach,” International Journal of Machine Learning and Computing, vol. 4, no. 3, pp. 300–306, June 2014.

[26] D. Küçük, “Automatic compilation of language resources for named entity recognition in Turkish by utilizing Wikipedia article titles,” Comput. Stand. Interfaces, vol. 41, pp. 1–9, 2015.

[27] O.-E. Ganea and T. Hofmann, “Deep Joint Entity Disambiguation With Local Neural Attention,” in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, July 2017.

[28] Y. Yang, O. Irsoy, and K. S. Rahman, “Collective entity disambiguation with structured gradient tree boosting,” in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol 1, April 2018.

[29] I. Yamada and H. Shindo, “Pre-training of Deep Contextualized Embeddings of Words and Entities for Named Entity Disambiguation,” arXiv preprint arXiv:1909.00426, pp. 76, September 2019.

[30] M. Cheriet, N. Kharma, L. Cheng-Lin, and S. Y. Ching, “Character recognition systems: A guide for students and practitioners.” John Wiley & Sons, 2007.

[31] A. Chaudhuri, K. Mandaviya, P. Badelia, and S. K. Ghosh, “Optical Recognition Systems,” in Optical Character Recognition Systems for Different Languages with Soft Computing, Cham, Springer International Publishing, pp. 9–41, 2017.

[32] N. V. Rao, A. S. C. S. Sastry, A. S. N. Chakravarthy, and P. Kalyanchakravarthi. "Optical Character Recognition Technique Algorithms." Journal of Theoretical and Applied Information Technology, vol. 83, no. 2, pp.275-282, 2016.

[33] K. Hamad and M. Kaya, “A detailed analysis of optical character recognition technology,” International Journal of Applied Mathematics, Electronics and Computers, vol. 4, no.1, pp. 244–249, 2016.

[34] N. Arica, F. T. Y. Vural, “An Overview of Character Recognition focused on Offline Handwriting,” IEEE Transactions on Systems, Man and Cybernetics – Part C: Applications and Reviews, vol. 31, no. 2, pp. 216–233, May 2001.

[35] P. Shah, S. Karamchandani, T. Nadkar, N. Gulechha, K. Koli, and K. Lad, “OCR-based chassis-number recognition using artificial neural networks,” in 2009 IEEE International Conference on Vehicular Electronics and Safety (ICVES), pp. 31-34, 2009.

[36] E. Oztop, A. Y. Mulayim, V. Atalay, and F. Yarman-Vural, “Repulsive attractive network for baseline extraction on document images,” in 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 75, no. 1, pp. 1-10, 2002.

[37] K. Karthick, K. B. Ravindrakumar, R. Francis, and S. Ilankannan, “Steps involved in text recognition and recent research in OCR; a study.” International Journal of Recent Technology and Engineering, vol. 8, no. 1, pp. 2277-3878, May 2019.

[38] O. Matei, P. C. Pop, and H. Vălean, “Optical character recognition in real environments using neural networks and k-nearest neighbor,” Applied Intelligence, vol. 39, no. 4, pp. 739–748, 2013.

[39] S. Singh, “Optical character recognition techniques: a survey.” Journal of Emerging Trends in Computing and Information Sciences, vol. 4, no. 6, pp. 545-550, 2013.

[40] J. Pradeep, E. Srinivasan, and S. Himavathi, “Diagonal based feature extraction for handwritten character recognition system using neural network,” in 2011 3rd International Conference on Electronics Computer Technology, vol. 4, pp. 364-368, 2011.

[41] Jaided AI, “EasyOCR: Ready-to-use OCR with 80+ supported languages and all popular writing scripts including Latin, Chinese, Arabic, Devanagari, Cyrillic and etc.” Available: https://github.com/JaidedAI/EasyOCR.

[42] “PyTorch,” Pytorch.org. [Online]. Available: https://pytorch.org/get-started/locally/. [Accessed: 9-Febuary-2022].

[43] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, “Character region awareness for text detection,” in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9365-9374, 2019.

[44] Y. Tai, J. Yang, and X. Liu, “Image super-resolution via deep recursive residual network,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3147-3155, 2017.

[45] S. Hochreiter, and J. Schmidhuber, “Long short-term memory.” Neural computation, vol. 9, no. 8, pp. 1735-1780, 1997.

[46] M. Liwicki, A. Graves, S. Fernàndez, H. Bunke, and J. Schmidhuber, “A novel approach to on-line handwriting recognition based on bidirectional long short-term memory networks.” In Proceedings of the 9th International Conference on Document Analysis and Recognition, ICDAR, 2007.

[47] J. Gu, K. Cho, and V. O. K. Li, “Trainable greedy decoding for neural machine translation,” in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Feb 2017.

[48] J. J. Webster and C. Kit, “Tokenization as the initial phase in NLP,” in Proceedings of the 14th conference on Computational linguistics, vol. 4, pp. 1106-1110, August 1992.

[49] “spaCy 101: Everything you need to know · spaCy Usage Documentation,” spaCy 101: Everything you need to know. [Online]. Available: https://spacy.io/usage/spacy-101. [Accessed: 21-June-2022].

[50] L. Yujian and L. Bo, “A normalized Levenshtein distance metric,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 6, pp. 1091–1095, June 2007.

[51] S. Sen, A. Ekbal, and P. Bhattacharyya, “Parallel corpus filtering based on fuzzy string matching,” in Proceedings of the Fourth Conference on Machine Translation, vol. 3, pp. 289-293, August 2019.

[52] G. A. Rao, G. Srinivas, K.V. Rao, and P. V. G. D. P. Reddy, “A Partial Ratio And Ratio Based Fuzzy-Wuzzy Procedure For Characteristic Mining Of Mathematical Formulas From Documents,” ICTACT Journal On Soft Computing, vol. 8, no. 4, pp. 1728-1732, July 2018.

[53] Jyotsna, S. Chauhan, E. Sharma, and A. Doegar, “Binarization techniques for degraded document images - A review,” in 2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), pp. 163-166, Sep 2016.

[54] A. M. Hambal, Z. Pei, and F. L. Ishabailu, “Image Noise Reduction and Filtering Techniques” International Journal of Science and Research (IJSR), vol. 6, no. 3, pp. 2033-2038, March 2015.

[55] S. B. Tambe, D. Kulhare, M. D. Nirmal and G. Prajapati, “Image processing (IP) through erosion and dilation methods” International Journal of Emerging Technology and Advanced Engineering, vol. 3, no. 7, pp. 285-289, July 2013.