

An OCR Post-correction Approach using Deep Learning for Processing Medical Reports

Srinidhi Karthikeyan^{1 2}, Alba G. Seco de Herrera¹, Faiyaz Doctor¹ and Asim Mirza²

¹School of Computer Science and Electronic Engineering, University of Essex, Colchester, CO4 3SQ UK

²Firza Group, London E10 5FA

According to a recent Deloitte study, the COVID-19 pandemic continues to place a huge strain on the global health care sector. Covid-19 has also catalysed digital transformation across the sector for improving operational efficiencies. As a result, the amount of digitally stored patient data such as discharge letters, scan images, test results or free text entries by doctors has grown significantly. In 2020, 2314 exabytes of medical data was generated globally. This medical data does not conform to a generic structure and is mostly in the form of unstructured digitally generated or scanned paper documents stored as part of a patient's medical reports. This unstructured data is digitised using Optical Character Recognition (OCR) process. A key challenge here is that the accuracy of the OCR process varies due to the inability of current OCR engines to correctly transcribe scanned or handwritten documents in which text may be skewed, obscured or illegible. This is compounded by the fact that processed text is comprised of specific medical terminologies that do not necessarily form part of general language lexicons. The proposed work uses a deep neural network based self-supervised pre-training technique: Robustly Optimized Bidirectional Encoder Representations from Transformers (RoBERTa) that can learn to predict hidden (masked) sections of text to fill in the gaps of non-transcribable parts of the documents being processed. Evaluating the proposed method on domain-specific datasets which include real medical documents, shows a significantly reduced word error rate demonstrating the effectiveness of the approach.

Index Terms—Optical Character Recognition (OCR), Natural Language Processing (NLP), Robustly Optimized Bidirectional Encoder Representations from Transformers (RoBERTa), Medical documents.

I. INTRODUCTION

From December 2019, the Covid-19 pandemic has demanded that countries around the world adopt digitalised healthcare solutions [1]. Governments of these countries have decided to use information technology to digitise the healthcare domain to reduce the challenges such as lack of communication, human errors or workflow [2]. In the healthcare sector, clinicians, General Practitioners (GPs), pharmacists and other healthcare workers read through medical documents (current treatments, test results, clinical notes, care plans and similar documents) and store them into patient records [3]. This can be a time consuming and error-prone process. In the United Kingdom (UK) alone there is a shortage of doctors for the general population and on average, a pharmacist spends 10 minutes reviewing a single document [4]. Pharmacists play a greater role in reducing the workload on doctors by handling many of the clinical tasks. According to Willis et al. [5], 44% percentage of the tasks performed by the GPs staff in 2020, can be automated using the digital technologies available. By automating this document management task, pharmacists can spend more time taking care of patient's needs. These documents have to be made easily accessible and available for the prescribers to use rather than them having to search through a large volume of documents which allows easy referencing of the medical history of patients.

In the UK, medical reports generated by the General Practices (GPs) are stored in the National Health Service (NHS) database. These documents are then allocated for companies

like Firza¹ where pharmacists read the reports and identify the crucial information available in the document (for example a new disease that is diagnosed) which are then added to the patient's database. Some previous work [6] have been carried out to automate this manual process to help improve efficiencies and reduce the cognitive burden.

Optical character recognition (OCR) can help achieve the online retrieval of the printed material such as medical documents, forms, or applications for retrieving valuable information that was available in the printed documents [7]. OCR technology has also improved over the years allowing us to digitise textual resources such as books, medical records, reports, documents, or newspapers. However, the accuracy of OCR can vary based on factors such as the amount of noise, quality of the original document or the font used [8]. Using the post-processing techniques, the quality of the OCR text can be improved. General post-processing methods include using domain-specific lexicons or dictionary-based methods [9]. This however limits their performance for the domains which require expensive resources, for example, in the medical domain building a vocabulary with all possible medical terms can be very expensive.

In particular, the UK NHS works with two types of PDF documents. The first type is the digitally created PDFs which consists of text and images that are created using dedicated software applications such as Microsoft® Word® or Excel®. These types of documents have text as well as metadata therefore it is easier for the OCR engines to access, recognise and edit the PDF documents. Current OCR engines can extract text from these types of documents with an excellent accuracy [10].

Manuscript received November 10, 2020. Corresponding author: S. Karthikeyan (email: srinidhi.karthikeyan@essex.ac.uk).

¹<https://firza.health/>

The second type of PDF document is the “image-only” which is created by converting images of scanned medical documents into PDFs locking the content of the image in a snapshot. These types of PDFs are more difficult for the OCR engines to process because they contain only images without any underlying text layer, making them unsearchable or uneditable. The NHS generally scan the physical documents and save them as scanned documents which fall under the second type of PDFs. Recent advancements in the field of Machine learning and deep learning has provided various techniques that can help us overcome these problems [8] which is explained briefly in the following section.

In this work, we proposed a customised OCR correction methodology where a post-processing component is added to the output of the OCR engine. This identifies the incorrect words in the OCR output, filters the entities and predicts the incorrect words to improve the accuracy of the OCR output text. The proposed method is considered a basic step towards building an overall medical document processing pipeline that can be used by healthcare companies that deal with medical documents like Firza. This post-processing approach will improve the accuracy of the extracted text as a first step towards automating and classification of medical documents. We propose a simple approach for OCR post-correction using Robustly Optimized Bidirectional Encoder Representations from Transformers (RoBERTa) [11] which improves the quality of the OCR output text. This approach does not require costly training data and uses a pre-trained model along with a spellchecker engine to improve the accuracy of the processed text. Hence the main contributions of this work are:

- Demonstrating the flexibility and robustness of using a deep learning based language model as part of an OCR post-correction document processing methodology.
- Evaluating the proposed approach on a publicly available dataset with varying differences in noise, font, quality, and alignment of the images.
- Evaluating the approach on the real-world medical documents where we are able to demonstrate a performance accuracy of 81% without training on domain-specific training.

The rest of the paper is organised as follows: a brief review of related work is given in Section II. Section III provides details of the proposed methodology. The datasets and evaluation metrics used are described in Section IV. Experimental results are presented in Section V. A discussion is presented in Section VI and Section VII presents conclusions with some final remarks.

II. RELATED WORK

The problem of spelling error detection and correction has been studied over years and the survey on those techniques before 1992 have been presented in [12].

Since that time, significant work has been done in this field not only in the English language but also in many other languages such as Arabic [13], French [14], Dutch [9] or German [15]. Magdy et al. [13] used a noisy channel model for the correction of errors in a scanned Arabic book and

Arabic newswire articles which reduced the word error rate approximately by 44% and 46% respectively. D’hondt et al. [14] used recurrent neural networks on French foetopathological reports which had 73% accuracy upon testing. De Does et al. experimented reducing the OCR errors in historical Dutch books [9] by adopting a lexicon-based method they were able to achieve precision 0.9028 and recall 0.9063. Furrer et al. [15] use not only a dictionary but also takes into account the information available in the text to correct the OCR errors in the German gothic scripts which helped in reducing the word error rate to 49.9%.

Most of the methods work by comparing the incorrect word and similar words in a vocabulary [16, 17], using N-grams [16] or using Hidden Markov Model [18] to find the correct candidate.

Bassil et al. [16] use N-grams for low-quality English and French documents where the error rate was reduced from 21.2% to 4.2% and 14.2% to 3.5% respectively. Borovikov et al. [18] propose using Hidden Markov Model to find the correct candidate for English documents for which the authors achieved 0.9444 recall and 0.9787 precision. Similarly for historical [19] or scanned documents that lack in quality of the images, use multiple OCR outputs or runs multiple times on the same document using the same engine [20, 21, 22, 23]. Thompson et al. used [19] rule-based as well as a medically tuned spell-checking strategy on historical medical documents from the British Medical Journal archive to improve the word-level accuracy up to 16%.

Kolak et al. [7] proposed a lexicon-free post-processing approach that used weighted finite state machines which were tested on books in different languages like Igbo which gave a 78% reduction in WER, for Cebuano language WER was reduced by 50.5% and in Arabic WER was reduced by 30%. Finite state machines are built from small datasets whereas [7] for a recognised character all OCR process hypothesis is combined [24]. But these approaches require a large amount of training data which is expensive and difficult to collect (e.g. medical data). To solve this Text Induced Corpus Cleanup (TICCL) system [25] has been used which requires no annotated training data while the statistical analysis is used to identify the high-frequency variants for the suggested Levenshtein distance. In addition to this, the authors have compared the performance metrics precision and recall based on different information available (with rank, without rank, with the lexicon, without the lexicon etc). However, these approaches require cleaning up of the corpus before using them which is difficult when the corpus is very large [25].

More recently, neural networks and deep learning approaches have been used to solve natural language problems. Post-correction methods have been particularly developed and applied such as auto-encoders [26] on Twitter and Wikipedia corpus which had promising improvement in the accuracy with appropriate settings like word lengths and type of the lexicon used to find the nearest match to the incorrect word or neural text embeddings [27]. Equally Long short term memory (LSTMs) [28] have been used for character-aligned strings or the Bidirectional Long Short Term Memory Networks (biLSTMs) [8] to produce a robust character-based language

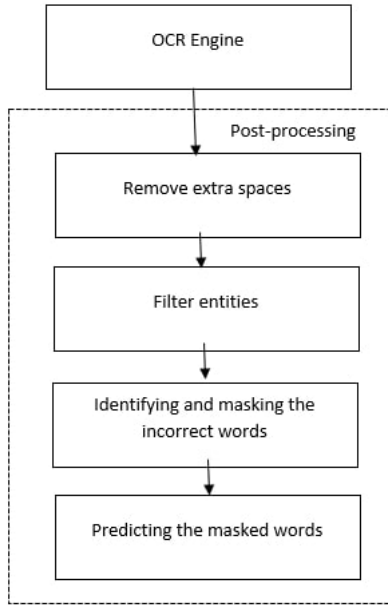


Fig. 1: Overview of the proposed framework for the OCR correction.

model which does not require annotated training data.

III. METHODOLOGY

This section presents a detailed description of the proposed corrective OCR post-processing methodology. Figure 1 presents the steps to implement the proposed approach. First, the OCR output is post-processed by removing the extra spaces. Second, the entities are filtered and third, the incorrect words are identified and masked. Finally, the masked words are predicted and replaced.

A. OCR engine

In this work, the text from the image is extracted using the tesseract OCR² engine which is open source and simple to use OCR engine. Tesseract was developed by Hewlett Packard Laboratories Bristol and at the lab in Greeley Colorado between 1985 and 1994 which was made open source in 2005 [29] under Google open source projects. Tesseract has Unicode (UTF-8) support and can recognise more than 100 languages.

B. Post-processing

As you can see in Figure 1 the post-processing has four steps: remove extra spaces; filter entities; identifying and masking the incorrect words, and Predicting the masked words. Each of the steps is described below.

1) Remove extra spaces

Whitespaces and extra lines will not add any information to our process so they are removed from the OCR output.

2) Filter entities

Domain-specific entities in the text such as names, address, drugs, or scientific terms are filtered. These words are not part of the English vocabulary of the spell checkers used in the next step and will be identified as incorrect words by them. To prevent this, we filter these entities before using the spell checkers. ScispaCy³ is a Python package containing spaCy models⁴ (open-source software library for advanced natural language processing) for processing biomedical, scientific, or clinical text. Entity linker function of Scispacy can link to entities in the Unified Medical Language System (UMLS) [30], Medical Subject Headings [31], RxNorm [32], Gene Ontology [33], Human Phenotype Ontology databases [34] to identify the scientific and medical entities. This function identifies entities by overlapping the provided string with the knowledge base using an approximate nearest neighbours search (Spacy⁵).

3) Identifying and masking the incorrect words

Once the domain-specific entities are filtered, the incorrect predictions made by the OCR engine for the English language are only identified. The incorrect words are identified using two spell checkers: pyspellchecker library⁶ and Hunspell⁷. These spell checkers will identify the misspelt words and also the words that are not available in the corpus. The final list of incorrect words is a combination of the words from the two spell checkers used. The list is created by the union of both lists produced from pyspellchecker and Hunspell without any duplicates.

4) Predicting the masked words

The selected incorrect words are corrected based on a RoBERTa word prediction model. For word prediction task, the happy transformer API⁸ is used which is available for XLNET [35], Bidirectional Encoder Representations from Transformers (BERT) [36] and RoBERTa language models. This API allows various complex Natural Language Processing (NLP) tasks like predicting masked words, predicting the probability of a sentence being followed by another sentence and other similar applications. Masked word prediction is a task wherein a sentence with [MASK] placeholder is given to a language model and it predicts the word that is supposed to be there in that [MASK] placeholder. For implementing the masked word prediction task RoBERTa language model is used. According to Liu et al. [11] Robustly Optimised Bidirectional Encoder Representations from Transformers (RoBERTa) is specifically trained for predicting masked words by modifying the key parameters of BERT, this includes removing BERT's next-sentence pretraining objective, and training with much larger mini-batches and learning rates. This allows the RoBERTa to yield better performance than BERT [11] for the masked word prediction task. RoBERTa is built on BERT's

³<https://allenai.github.io/scispacy/>

⁴<https://spacy.io/>

⁵<https://spacy.io/usage/linguistic-features>

⁶<https://pypi.org/project/pyspellchecker/>

⁷<http://hunspell.github.io/>

⁸<https://pypi.org/project/happytransformer/>

²<https://opensource.google/projects/tesseract>

language masking strategy where the system intentionally hides specific words and learns to predict the hidden words in unannotated data. For example, if the OCR output text is “wher even the tamest will always show the greatest fear of a little boy” then the word “wher” will be identified as the misspelt word. Then this word is masked “[MASK] even the tamest will always show the greatest fear of a little boy” before being fed into the RoBERTa model. Two levels of prediction are considered here. First, the RoBERTa model will predict the word without any options. Secondly, a suggested words list is used as options for RoBERTa to predict the most suited word for that masked position. In the previous example the suggested list of possible words by the SpellChecker library is [“her”, “whet”, “wer”, “sher”, “whey”, “where”, “whew”, “when”, “whee”]. These lists of words are fed as the options to the model. From this, the model returns the same set of words and their probability values for them to be the masked word in the given input sentence. These two predictions are then considered together and the words with maximum probability values are chosen and the similarity between the incorrect word and the predicted words are estimated. The word with maximum similarity is replaced in the actual text. For calculating the similarity between the words, the SequenceMatcher function from the difflib library⁹ is used. This uses the basics of “gestalt pattern matching” by Ratcliff and Obershelpin 1980s which finds the longest contiguous matching subsequence. The reason for using the two levels of implementation is because the predictions without options will be broad. For example, if the text “On examination the scar at the left upper arm has healed well. Skin check did not reveal any suspicious skin [MASK].” is the input for the language model, then the possible predictions are “lesions”, “cells”, “mass”, “mark”, “evidence” and other similar words. However while using the options such as [“lesson”, “lessons”, “lesions”, “lesion”] we can narrow down the set of predicted words to the words that have a similar structure (as in the number of words, spelling or length of the word) for this example the word “lesions” will be the correct prediction.

IV. EVALUATION

In this section, we describe the evaluation methodology used in this work as well as the two datasets used.

A. Datasets

The main goal of this work is to improve the OCR process for the NHS medical reports in the UK. However, an extensive NHS reports dataset is not publicly accessible due to security and privacy reasons. Hence, in order to check the robustness of the proposed model, it is also evaluated on the publicly available Mining Biodiversity (MiBio) dataset [37].

1) UK NHS reports dataset

The NHS reports dataset consists of scanned documents that include clinical letters, reports and discharge summaries which are part of the NHS clinical routine. The documents are provided by Firza¹⁰ who is responsible for processing patient

documents provided by the NHS. This dataset cannot be released publicly because of the sensitive patient information. In this work, 100 NHS medical reports have been used for testing out the proposed approach. The dataset consists of different types of medical reports/letters such as communication letters, did not attend appointment letters or discharge summaries. The dataset contains only scanned documents in PDF format (image-only type PDF). The ground truths for these documents were manually checked and corrected by professional pharmacists from Firza.

2) Mining Biodiversity (MiBio) dataset

Mining Biodiversity (MiBio) dataset [37] is chosen in this work due to its similarity to medical documents in terms of the type of document (scanned paper documents). MiBio dataset consists of the scanned pages of the book named “Birds of Great Britain and Ireland (Volume II)” [38]. This is made publicly available by the Biodiversity Heritage Library (BHL) for Europe. Out of 460 pages, 211 pages comprise of text and other pages have both text and images. Pages like index, appendix and only pictures being ignored. Therefore, only those 211 pages are scanned and used for this work. The published dataset consists of these pages as well as the ground truths which are manually corrected by Mei et al. [37]. The ground truths do not contain the titles and footnotes even though they are available in the images. Therefore, the titles and footnotes are removed in this work.

An easy method to remove the titles and footnotes from the documents is to crop them out from the images. Depending on how much the image was cropped, the clarity of the image will increase or decrease. Therefore, cropping the image could change the sizes and clarity of the resulting images which will affect the fairness of the evaluation of the model. Therefore, the heading and footnotes are blacked out and the original size of the image is maintained in this work. This Wikipedia¹¹ article lists and explain the different types of birds in Great Britain. This article contains the links for the types of birds mentioned, these URLs are extracted and the contents from the 704 URLs are scraped out and used for training and testing of the post-processing approach. Out of the 704 URLs, 500 documents are used for training and 204 documents are used for testing.

B. Measuring OCR quality

Carrasco et al. [39] present the following metrics that we used in this work. *Word error rate (WER)* which is calculated as

$$WER = (i_w + s_w + d_w)/n_w$$

where n_w is the number of words in the ground truth text, i_w is the number of words inserted, s_w is the number of words substituted, d_w is the number of words deleted to get the original ground truth values.

The *Character Error Rate (CER)* is defined in a similar way as,

$$CER = (i + s + d)/n \quad (1)$$

⁹<https://docs.python.org/3/library/difflib.html>

¹⁰<https://firza.health/>

¹¹https://en.wikipedia.org/wiki/List_of_birds_of_Great_Britain

where n is the total number of characters, i is the minimal number of character insertions, s is substitutions and d deletions required to transform the reference text into the OCR output.

According to Carrasco et al. [39], there are three types of OCR errors:

- Misspelled characters (substitutions);
- Lost or missing text (deletions);
- Spurious symbols (insertions).

Therefore, the average insertion, average substitution and average deletion character errors are also calculated. This is calculated to analyse which type of error is prominent in each type of dataset.

a) Misspelled characters (substitutions) error: happen when the OCR output text have characters that are misspelled. Using Equation 1, the misspelled character can be calculated as,

$$s = (CER * n) - (i + d)$$

Figure 2 presents an example of misspelt character errors found in the MiBio dataset after processing it with the OCR engine. It is read as “the general form of their heads they somewhat remind one of Starlings, not be confounded with the so-called “Orioles” of the New World”. In this example, the letter “O” is misinterpreted as “0”.

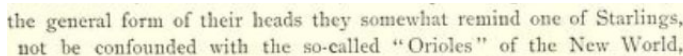


Fig. 2: Example of misspelled error observed in the MiBio dataset.

An example of Misspelled characters errors is also found for the NHS reports documents after processing it with the OCR engine. Figure 3, shows an example. This image is read as “Purther to your referral from your GP undei the two week wait pathway we are writing to inform. It is clearly seen that letter “F” is misinterpreted as “P” and letter “r” is misinterpreted as “i”.

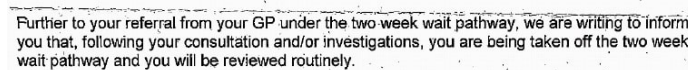


Fig. 3: Example of misspelled error observed in the NHS dataset.

b) Lost or missing text (deletions) error: is when the OCR output text misses a character. Using Equation 1, the lost character can be calculated as,

$$d = (CER * n) - (i + s)$$

For example in Figure 4, is read as “She takes muslin and Clopidogrel amongst other medication and is allergic to Penicillin”. Here “i” in the text is disappeared and is misunderstood as “m”. In this example the letter “i” is lost or missing.

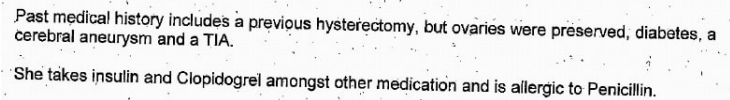


Fig. 4: Example of missing text error observed in the NHS dataset.

TABLE I: Average WER and average CER of NHS reports dataset.

Method	Average WER	Average CER
Before post-processing	15.06	13.31
After post-processing	13.67	12.48

TABLE II: Average insertions, average substitutions and average deletion errors for NHS reports dataset

Types of Error	Average Insertions	Average Substitutions	Average Deletions
Before post-processing	0.878	0.7114	0.8453
After post-processing	0.7956	0.6432	0.7154

c) Spurious symbols (insertions) errors: happen if the OCR output text contains new characters inserted that are not available in the document. Using Equation 1, the insertion character error can be calculated as,

$$i = (CER * n) - (s + d)$$

For example 5, is read as “Based on her cornbination of symptoms I think a colonoscopy is sensible and I have arranged it.” The letter “m” is read as a combination of “r” and “n”. An additional letter is inserted to the text.

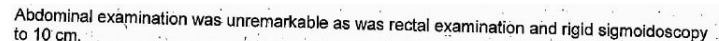


Fig. 5: Example of insertion symbols error observed in the NHS reports dataset.

V. RESULTS

In order to measure the effectiveness of the proposed approach, we test them on the two datasets namely: the MiBio dataset and NHS report dataset.

A. NHS report dataset

Table I presents the results on the NHS reports dataset before and after the post-processing. After applying the post-processing steps the average WER and average CER values are reduced by 1.39 and 0.83 respectively.

Table II presents the average insertion, average substitution and average deletion errors. It is observed that 0.08, 0.06 and 0.13 of the average insertion, substitution and deletion errors have been reduced respectively.

Figure 6 shows a line chart of the WER values of the NHS reports dataset before and after post-processing which shows a reduction in the WER values after post-processing.

TABLE III: Average WER and Average CER of MiBio dataset.

Method	Average WER	Average CER
Before post-processing	24.325	19.24
After post-processing	18.616	17.17

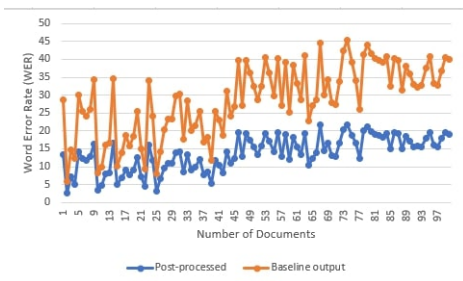


Fig. 6: Graph representing the WER values for NHS reports dataset before and after RoBERTa correction

B. MiBio dataset

Similar to the NHS dataset evaluation, Table III presents the results on the MiBio dataset reports dataset before and after the post-processing. We can see that the average WER and CER values have been reduced by 5.709 and 2.07 respectively.

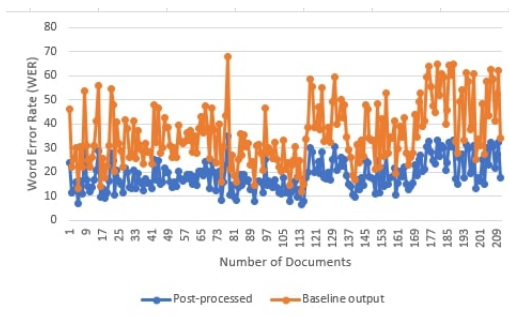


Fig. 7: Graph representing the WER values for MiBio dataset before and after RoBERTa correction

Figure 7 shows a line chart of the WER values of the MiBio dataset before and after post-processing which shows a reduction in the WER values after post-processing.

In addition to WER and CER, Table IV presents the average insertions, substitutions and deletion errors. It is observed that 0.117,0.167,0.147 of the average insertions, substitutions and deletions errors have been reduced respectively.

TABLE IV: Average insertions, average substitutions and average deletions Errors for MiBio Dataset.

Types of Error	Average Insertions	Average Substitutions	Average Deletions
Before post-processing	0.93	0.812	0.858
After post-processing	0.813	0.645	0.711

VI. DISCUSSION

From Table III we can see that after using the post-processing technique on the MiBio dataset the average WER and average CER are reduced by 5.709 and 2.07. Similarly

from Table I, the post-processing technique on the NHS dataset has helped us reduce the average WER and average CER by 1.39 and 0.83 respectively. This shows that without even training the language model with domain-specific data there is a considerable reduction in both the average WER and average CER. In addition to the average error rates (average of the entire MiBio as well as the NHS dataset), the individual WER values of both datasets also show a reduction in the WER which is shown in Figure 6 and Figure 7. To analyse the type of error that occurs the most in both the datasets the average insertions, average substitutions and average deletion errors are calculated for the entire dataset. From Table IV, it is observed that for the MiBio dataset, the insertions type of error is notably high, possibly because the quality of the scanned images and font that is used in the book is difficult for the OCR engine to recognise. From Table II, it is noticeable for the NHS reports dataset that the average substitutions error is less than the other two types of error. One possible reason for this could be the font used for the text is less confusing for the OCR engine, which means there are reduced misinterpretations of the characters. Also, to check the effect of fine-tuning, the approach is tested on the MiBio dataset without training the model with the Wikipedia articles. Without fine-tuning the average WER and average CER are 18.926 and 17.46 respectively which is 0.31 and 0.29 more than the fine-tuned results. These results provide a piece of evidence that training on domain-specific data can improve the performance of the model which might be the case for the NHS dataset as well. This will be tested in the future. This work showed evidence that word prediction can be used for OCR post-correction. In comparison to the prior research, the approach proposed by Bassil et al. [16] of using the Google Web 1T 5-Gram Data Set for error correction is tested on two documents of two different languages namely French and English. Using the proposed method, the authors were able to reduce the error rate from 21.2% to 4.2% in the English document. Our approach uses the pre-trained model which is not as computationally intense as the approach proposed by Bassil et al. [16] given that our approach is not accessing a tremendous database.

Thompson et al. [19] propose the use of an approach that combines multiple techniques like pre-processing the OCR output using a spellchecker which is also fine-tuned for medical terms after which the pre-processed text is corrected selectively. This method was able to reduce the error rate by 16%. Our approach is similar to the one proposed by Thompson et al. [19] where the incorrect word identification steps include the usage of the spell checker. However [19], use the Hunspell dictionary which is modified to add medical terms to the vocabulary. The additional use of domain-specific dictionaries is expected to be implemented in our future works. The work presented by D'hondt et al. [8] proposes a model that does not require annotated corpus data for training. This proposed method is tested in the French foetopathological reports dataset which outperforms the baseline method by 14.3%. This is similar to our approach, as our model training with domain-specific data (MiBio dataset) and does not require any annotation for the training dataset.

However, even in the absence of additional training, we

can see that our approach gives a significant reduction in the word and character error rates. The results from [8] also shows that training a similar neural network approach shows improved performance in the post-correction method. Training our proposed model with domain-specific medical data will therefore further improve its performance where this will be investigated in future works.

VII. CONCLUSIONS

A novel approach for post-correction of the OCR generated output is presented in this paper. The proposed method uses RoBERTa language model for post-processing of the OCR output text. This method is tested on the UK NHS medical reports dataset. Additionally, to test the robustness of the approach it was tested on a publicly available dataset, the MiBio dataset. The experiments carried out on both datasets shows a good result of reductions in the average WER and CER. The results on both MiBio and NHS report datasets showed that this approach could be applicable to domain-specific applications where documents have similar characteristics.

This work is part of a document processing pipeline that aims to automate the extraction of medical information from NHS patient reports, classify different types of medical documents, determine, and recommend crucial amendments that should be recorded in live patient records. The proposed post-processing approach has been developed to be used as the first phase of this pipeline. In the future, appending the medical terminologies to the spell checking vocabulary, work on improving the quality of the images and training the model with the domain-specific dataset are expected to be implemented to reduce the error rates even more.

ACKNOWLEDGEMENT

This study has been funded through an Innovate UK Knowledge Transfer Partnership between Firza and the School of Computer Science & Electronic Engineering, University of Essex, Partnership No: 12086.

REFERENCES

- [1] E. Mahase, "Covid-19: Mental health consequences of pandemic need urgent research, paper advises," 2020.
- [2] A. Laur, "Fear of e-health records implementation?" *Medico-Legal Journal*, vol. 83, no. 1, pp. 34–39, 2015.
- [3] M. Honeyman, P. Dunn, and H. McKenna, "A digital nhs," *An introduction to the digital agenda and plans for implementation*, 2016.
- [4] M. Taylor, "Why is there a shortage of doctors in the UK?" *The Bulletin of the Royal College of Surgeons of England*, vol. 102, no. 3, pp. 78–81, 2020.
- [5] M. Willis, P. Duckworth, A. Coulter, E. T. Meyer, and M. Osborne, "Qualitative and quantitative approach to assess of the potential for automating administrative tasks in general practice," *BMJ Open*, vol. 10, no. 6, p. e032412, 2020.
- [6] M.-O. Wright, A. Fisher, M. John, K. Reynolds, L. R. Peterson, and A. Robicsek, "The electronic medical record as a tool for infection surveillance: successful automation of device-days," *American journal of infection control*, vol. 37, no. 5, pp. 364–370, 2009.
- [7] O. Kolak and P. Resnik, "Ocr post-processing for low density languages," in *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 2005, pp. 867–874.
- [8] E. D'hondt, C. Grouin, and B. Grau, "Generating a training corpus for ocr post-correction using encoder-decoder model," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2017, pp. 1006–1014.
- [9] J. de Does and K. Depuydt, "Lexicon-supported ocr of eighteenth century dutch books: a case study," in *Document Recognition and Retrieval XX*, vol. 8658. International Society for Optics and Photonics, 2013, p. 86580L.
- [10] S. Vijayarani and A. Sakila, "Performance comparison of ocr tools," *International Journal of UbiComp (IJU)*, vol. 6, no. 3, pp. 19–30, 2015.
- [11] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.
- [12] K. Kukich, "Techniques for automatically correcting words in text," *Acm Computing Surveys (CSUR)*, vol. 24, no. 4, pp. 377–439, 1992.
- [13] W. Magdy and K. Darwish, "Word-based correction for retrieval of arabic ocr degraded documents," in *International Symposium on String Processing and Information Retrieval*. Springer, 2006, pp. 205–216.
- [14] E. D'hondt, C. Grouin, and B. Grau, "Low-resource ocr error detection and correction in french clinical texts," in *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis*, 2016, pp. 61–68.
- [15] L. Furrer and M. Volk, "Reducing OCR errors in gothic-script documents," 2011.
- [16] Y. Bassil and M. Alwani, "OCR context-sensitive error correction based on google web 1t 5-gram data set," *arXiv preprint arXiv:1204.0188*, 2012.
- [17] J. Evershed and K. Fitch, "Correcting noisy ocr: Context beats confusion," in *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, 2014, pp. 45–51.
- [18] E. Borovikov, I. Zavorin, and M. Turner, "A filter based post-ocr accuracy boost system," in *Proceedings of the 1st ACM workshop on Hardcopy document processing*, 2004, pp. 23–28.
- [19] P. Thompson, J. McNaught, and S. Ananiadou, "Customised OCR correction for historical medical text," in *2015 Digital Heritage*, vol. 1. IEEE, 2015, pp. 35–42.
- [20] M. Volk, L. Furrer, and R. Sennrich, "Strategies for reducing and correcting ocr errors," in *Language technology for cultural heritage*. Springer, 2011, pp. 3–22.
- [21] W. B. Lund, E. K. Ringger, and D. D. Walker, "How well does multiple ocr error correction generalize?" in *Document Recognition and Retrieval XXI*, vol. 9021.

- International Society for Optics and Photonics, 2014, p. 90210A.
- [22] F. Boschetti, M. Romanello, A. Babeu, D. Bamman, and G. Crane, "Improving ocr accuracy for classical critical editions," in *International Conference on Theory and Practice of Digital Libraries*. Springer, 2009, pp. 156–167.
- [23] D. Lopresti and J. Zhou, "Using consensus sequence voting to correct ocr errors," *Computer Vision and Image Understanding*, vol. 67, no. 1, pp. 39–47, 1997.
- [24] R. Llobet, J.-R. Cerdan-Navarro, J.-C. Perez-Cortes, and J. Arlandis, "Ocr post-processing using weighted finite-state transducers," in *2010 20th International Conference on Pattern Recognition*. IEEE, 2010, pp. 2021–2024.
- [25] M. Reynaert, "Non-interactive ocr post-correction for giga-scale digitization projects," in *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, 2008, pp. 617–630.
- [26] S. Raaijmakers, "A deep graphical model for spelling correction," in *BNAIC 2013: Proceedings of the 25th Benelux Conference on Artificial Intelligence, Delft, The Netherlands, November 7-8, 2013*. Delft University of Technology (TU Delft); under the auspices of the Benelux . . . , 2013.
- [27] G. Chrupała, "Normalizing tweets with edit scripts and recurrent neural embeddings," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2014, pp. 680–686.
- [28] M. Al Azawi, "Statistical language modeling for historical documents using weighted finite-state transducers and long short-term memory," 2015.
- [29] A. Kay, "Tesseract: an open-source optical character recognition engine," *Linux Journal*, vol. 2007, no. 159, p. 2, 2007.
- [30] O. Bodenreider, "The unified medical language system (umls): integrating biomedical terminology," *Nucleic acids research*, vol. 32, no. suppl_1, pp. D267–D270, 2004.
- [31] F. B. Rogers, "Medical subject headings," *Bulletin of the Medical Library Association*, vol. 51, no. 1, pp. 114–116, 1963.
- [32] S. J. Nelson, K. Zeng, J. Kilbourne, T. Powell, and R. Moore, "Normalized names for clinical drugs: Rxnorm at 6 years," *Journal of the American Medical Informatics Association*, vol. 18, no. 4, pp. 441–448, 2011.
- [33] G. O. Consortium *et al.*, "Creating the gene ontology resource: design and implementation," *Genome research*, vol. 11, no. 8, pp. 1425–1433, 2001.
- [34] S. Köhler, N. A. Vasilevsky, M. Engelstad, E. Foster, J. McMurry, S. Aymé, G. Baynam, S. M. Bello, C. F. Boerkoel, K. M. Boycott *et al.*, "The human phenotype ontology in 2017," *Nucleic acids research*, vol. 45, no. D1, pp. D865–D876, 2017.
- [35] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *arXiv preprint arXiv:1906.08237*, 2019.
- [36] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [37] J. Mei, A. Islam, A. Moh'd, Y. Wu, and E. E. Milios, "MiBio: A dataset for OCR post-processing evaluation," *Data in brief*, vol. 21, pp. 251–255, 2018.
- [38] J. T. R. Sharrock, *The atlas of breeding birds in Britain and Ireland*. A&C Black, 2010.
- [39] R. C. Carrasco, "An open-source OCR evaluation tool," in *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, 2014, pp. 179–184.



Srinidhi Karthikeyan was born in India in 1997. She completed her bachelor's degree in Computer Science and Engineering from Anna University, Chennai in 2019 and in 2020 she completed her master's degree in Data Science from the University of Essex, Colchester, UK. Her main area of interest is Natural Language processing and computer vision.



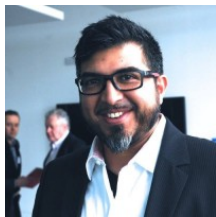
Alba G. Seco de Herrera was born in Madrid, Spain, in 1984. She received a Diploma in Mathematics from the Complutense University in Madrid, Spain, in 2008 and a Master's in Science in Telemedicine and Bioengineering from the Technical University of Madrid, Spain, in 2009. She received her doctorate degree in Computer Science from University of Geneva, Switzerland, in 2015. In 2015, she joined the National Library of Medicine - National Institutes of Health (USA) as a postdoctoral fellow. Since 2017, she has been a Lecturer at the

School of Computer Science and Electronic Engineering, University of Essex, Colchester, UK. Her research interest lie primarily in the area of Computer Vision for Biomedical Imaging with a special focus on image retrieval and evaluation. She also cast the problem of image understanding as a cross-modality matching scenario in which visual content and textual information need to be combined.



Faiyaz Doctor (M'08-SM'21) received the B.Sc. degree in computer science and artificial intelligence from University of Birmingham in 1998. M.Sc. degree in artificial intelligent agents in 2001 and PhD. degree in computer science in 2006 from the University of Essex. He is currently a Senior Lecturer and head of the Intelligent Connected Societies Group at the School of Computer Science and Electronic Engineering, University of Essex, U. K and a Director at Interactive Coventry Ltd. His research interests are in computational intelligence emphasising on fuzzy

systems, deep learning, evolutionary algorithms, explainable machine learning and ambient intelligence. Dr Doctor has published over 75 papers in peer-reviewed international journals, conferences, and workshops. He has led and co-led several projects funded through Innovate UK, Harvard University, Newton Fund and the Royal Academy of Engineers. He is associate editor of the IEEE Transactions on Fuzzy Systems, fellow of the Higher Education Academy and member of the IEEE Computational Intelligence Society's Emergent Technologies Technical Committee.



Asim Mirza is the founder and chief executive officer of Firza Group. Running businesses and organisations in Primary Care and Community Pharmacy, Asim is recognised as a leader in innovation and service development in Healthcare. With a diverse background in Digital Health through to Multi-Site management, Asim is known for creating quality systems and processes from scratch to enable healthcare businesses and organisations to run smoothly and efficiently. He has created innovative technological solutions for many GP Practices, Pharmacies

and Dispensaries to increase productivity. Asim has helped organisations work more closely with carehomes to significantly improve service to this sector. He has a good understanding of how primary care organisations can improve cost effectiveness without shortcuts on the patient's experience. Asim has worked for both large national healthcare chains and independent multiples in leading management roles for a number of years. Asim Mirza is the founder and chief executive officer of Firza Group. Running businesses and organisations in Primary Care and Community Pharmacy, Asim is recognised as a leader in innovation and service development in Healthcare. With a diverse background in Digital Health through to Multi-Site management, Asim is known for creating quality systems and processes from scratch to enable healthcare businesses and organisations to run smoothly and efficiently. He has created innovative technological solutions for many GP Practices, Pharmacies and Dispensaries to increase productivity. Asim has helped organisations work more closely with carehomes to significantly improve service to this sector. He has a good understanding of how primary care organisations can improve cost effectiveness without shortcuts on the patient's experience. Asim has worked for both large national healthcare chains and independent multiples in leading management roles for a number of years.