Journal of Big Data

RESEARCH Open Access

Advancing machine learning with OCR2SEQ: an innovative approach to multi-modal data augmentation

Michael Lowe[1], Joseph D. Prusa[1], Joffrey L. Leevy[1*] and Taghi M. Khoshgoftaar[1]

*Correspondence:
jleevy2017@fau.edu

[1] Florida Atlantic University, 777 Glades Road, Boca Raton, FL 33431, USA

Abstract

OCR2SEQ represents an innovative advancement in Optical Character Recognition (OCR) technology, leveraging a multi-modal generative augmentation strategy to overcome traditional limitations in OCR systems. This paper introduces OCR2SEQ's unique approach, tailored to enhance data quality for sequence-to-sequence models, especially in scenarios characterized by sparse character sets and specialized vocabularies. At the heart of OCR2SEQ lies a set of novel augmentation techniques designed to simulate realistic text extraction errors. These techniques are adept at generating diverse and challenging data scenarios, thereby substantially improving the training efficacy and accuracy of text-to-text transformers. The application of OCR2SEQ has shown notable improvements in data processing accuracy, particularly in sectors heavily dependent on OCR technologies such as healthcare and library sciences. This paper demonstrates the capability of OCR2SEQ to transform OCR systems by enriching them with augmented, domain-specific data, paving the way for more sophisticated and reliable machine learning interpretations. This advancement in OCR technology, as presented in the study, not only enhances the accuracy and reliability of data processing but also sets a new benchmark in the integration of augmented data for refining OCR capabilities.

Keywords: Large language models, Optical character recognition, Sequence-to-sequence models, Text-to-text transformers, Data augmentation, Noise correction

Introduction

Optical Character Recognition (OCR) is an important translation mechanism, transforming scanned documents into byte-encoded text that bridges the physical and digital worlds [1, 2]. Generally, OCR performs well with clean documents in standard use cases. However, its efficacy diminishes in specialized domains characterized by sparse character use and unique vocabulary. Such limitations are particularly evident with open-source OCR engines, necessitating post-processing corrections and typo adjustments. In the rapidly evolving field of machine learning, especially with the widespread adoption of large language models and transformer architectures, the dependability of OCR systems becomes increasingly vital. These systems act as

conduits to a wealth of historical data [3], essential for learning from vast data repositories, especially in data-rich sectors like healthcare and library sciences [4]. To enhance the utility of domain-specific data, cutting-edge methods such as prompted labeling [5] are employed to fine-tune large language models. These techniques compensate for data sparsity and computational resource limitations, adapting the models for specific domain applications.

Recent advancements in machine learning have incorporated large language models as integral components of post-processing pipelines [6] for error correction. Prominent among post-OCR correction techniques are model head learning tasks like fill-masking, text-generation, and text-to-text translation. These tasks vary in complexity based on the tokenization approach, or its absence, influencing their application [7, 8]. Fill-masking, for example, is particularly valued for its unsupervised nature and ability to predictively address tokens at the individual level. During training, if the model encounters an unfamiliar token, it employs this technique to suggest a probable correction. However, the effectiveness of fill-masking is constrained by factors such as the sequence's structural context, the quality of the training data, and language perplexity within that context [7, 9]. The reliance on language perplexity, optimized through a probability distribution rather than absolute metrics like Word Error Rate (WER) or Character Error Rate (CER), presents a limitation. It fails to directly measure how well the model understands or reproduces language as humans do, potentially overlooking errors that affect comprehension and communication quality. To overcome this challenge, more advanced approaches involve sequence de-noising using text-to-text transformers as well as more sophisticated methods like tokenless, byte-level learning tasks. All these techniques fundamentally hinge on the nuances of various tokenization schemes.

In addition to explicit training methodologies for enhancing post-OCR data correction, significant research efforts are being channeled into more subtle techniques aimed at improving OCR performance. These include data augmentation and conditional error reproduction [10]. Such tasks, while on the higher end of computational complexity, offer reliable means for enhancing performance, particularly in datasets with a higher incidence of noise. This often involves the manual annotation of error types common to specific document groups or estimating a probability distribution for manually entered data errors. Data augmentation strategies, typically governed by a set of rules, provide a systematic and repeatable framework for simulating erroneous data, which is crucial for training deep learning models. These methodologies not only enhance the robustness of the models but also contribute to more accurate and efficient data processing in scenarios where error prevalence is high [10–12].

The healthcare sector, particularly in the realm of precision health, presents a prime opportunity for enhancing data quality in large language models, where OCR plays a pivotal role [13]. The emergence of patient-centered, data-driven ecosystems, such as learning health systems and value-based care models, underscores the value of leveraging historical data for process and quality improvement. Despite ongoing digitization efforts in healthcare services, these data-driven approaches often face challenges, including the absence of general support mechanisms like open-source datasets [14], specialized language models, and extensive historical data. In the healthcare industry, most data sources are proprietary and heavily regulated, necessitating significant manual

effort for data annotation and cleaning, which is essential for their use in domain-specific clinical language systems [15].

Contrary to common belief, not all datasets and their generating systems are created equal. This discrepancy has spurred research into fine-tuning and other domain-specific data adaptation strategies to enhance performance. In the U.S. healthcare system, for example, over 85% of all electronic health records consist of unstructured free text. This challenge of data quality is also prevalent in information management and library systems, particularly concerning historical news articles and literature. Despite advancements in digital record processing, a substantial portion of historical data remains in image form, requiring OCR for text conversion. Our focus is on generating realistic noise patterns that mirror the most common errors observed in OCR processes, such as WER, CER, and Vocabulary Disjoint Rate. By doing so, we aim to utilize not only the translation errors from probabilistic character conversion but also the inherent cosmetic qualities of documents. This approach seeks to create more robust and lifelike noise scenarios encountered during the OCR process, thereby enhancing the overall quality and reliability of data extraction.

With the sophisticated use of large language models in data translation and cleansing across diverse architectures, a critical question arises: can we harness the capabilities of autoencoders to decipher the intricacies that lead OCR engines to specific errors, and use this insight to transform error-laden text into more accurate renditions? Our exploration begins with identifying what constitutes 'quality noise' for training these systems, thereby generating datasets that are optimally suited for domain adaptation, even in cases where such datasets are not readily available. Our focus is on creating generative noise that accurately replicates the most common errors observed in OCR processes, specifically targeting metrics like CER, WER, and Vocabulary Disjoint Rate. By doing this, our goal is to utilize augmented mock samples to instruct text-to-text transformers on generalizing the types of errors produced by an OCR engine for a particular domain. This involves modifying images and their structures to induce a broad spectrum of OCR errors, thus enabling the automated handling of prevalent error types. Such an approach not only enhances the accuracy of text extraction but also contributes significantly to the efficiency of OCR systems in various application domains.

OCR2Seq as a pre-training framework

We introduce Optical Character Recognition to Sequence (OCR2Seq) as a novel pre-training framework designed to achieve several key objectives:

1. Develop a resilient and noise-tolerant pre-training mechanism, encapsulated in a universally applicable autoencoder. This element of OCR2Seq is crafted to withstand and adapt to a variety of noise levels, thereby enhancing its robustness and utility in diverse data environments.
2. Establish an adaptive design pattern dedicated to the generation of simulated physical documents. These mock documents are characterized by their heterogeneous structures and adjustable tolerance to cross-modal noise, allowing for extensive experimentation and application across different domains and document types.

3 Conduct an in-depth exploratory analysis focused on evaluating the quality of specific augmentation types utilized on these mock documents. This aspect of OCR2Seq is critical in understanding the effectiveness of different augmentation strategies and their role in improving the accuracy and reliability of OCR processes.

Through these initiatives, OCR2Seq positions itself as a comprehensive and versatile framework, pivotal in advancing the capabilities of OCR systems and enhancing the quality of data used in machine learning models, particularly in the context of text extraction and processing.

Contribution

This study involves the development and utilization of two modified datasets derived from existing large datasets, as well as the implementation of a novel training framework. These components are central to our research, each serving a unique purpose in exploring and enhancing OCR technology.

1 MIMIC-III: Mock Docs—Invoice and Patient Notes This component is a new dataset that facilitates the development of advanced OCR techniques. The dataset extends the MIMIC-III and MIMIC-IV databases to include a mock repository of electronic health records. These mock records are designed to reflect the most commonly processed formats in healthcare record-keeping, such as those by the Centers for Medicare and Medicaid Services (CMS), including standard medical invoices and provider summary notes for patient visits. The MIMIC dataset, initially in plain text, is converted into an image simulating a scanned document and undergoes synthetic augmentation to simulate OCR errors, offering a unique dataset for training a sequence-to-sequence language model. These errors represent cosmetic flaws, such as oversized headers and bar-code footers, to simulate real-world scenarios and challenges in OCR processing. The augmented text is then provided as input to the model, with the original, unaugmented text as the desired output.
2 CC News: mock docs—newsletter This component is also a new dataset that supports the development of advanced OCR techniques. We have created a mock repository based on CC News Articles from the Hugging Face data repository [16]. These documents are formatted to resemble standard newsletters. Similar to the approach taken with MIMIC-III, this dataset undergoes a process of synthetic error introduction to mimic common OCR challenges. However, the focus here shifts to subtler issues, including noise injection, character recognition errors and difficulties in page segmentation, providing a distinct set of obstacles for training sequence-to-sequence models. As with the previous dataset, the goal is to use the manipulated text for model training, aiming for the restoration of the original, error-free text.
3 OCR2Seq: A learning task for systemic denoising OCR2Seq enhances OCR techniques through an advanced noise correction technique. It is an end-to-end framework for training a model to correct OCR errors, integrating a novel data augmentation strategy with sequence-to-sequence model training. The process begins with ground truth text, converting it into an image or PDF, then introducing noise to create a corrupted version. This corrupted document is processed through pytesseract

[17], an open-source OCR tool that translates images of text into machine-readable text, to generate augmented text that simulates OCR errors. The sequence-to-sequence language model is then trained to predict the original, error-free text from this augmented input, effectively correcting the simulated OCR errors. The use of pytesseract in the data augmentation process of OCR2Seq is chosen for its wide usage, open-source availability, and capability to replicate a broad spectrum of OCR errors, which makes it an ideal tool. Pytesseract is widely known to work well on scanned documents, which our mock documents simulate. The innovative approach of our framework, which combines data augmentation with machine learning, aims to systematically address and correct OCR errors, thus enhancing OCR accuracy across different domains.

Each of these components plays a vital role in the research, providing distinct but complementary perspectives and tools for enhancing OCR technologies. The MIMIC-III and CC News datasets offer practical, real-world scenarios for testing and refining OCR systems, while the OCR2Seq framework provides an innovative approach to improving OCR accuracy through systemic denoising and error correction.

Organization of remaining sections

The remainder of this paper is structured as follows: Sect. 2 provides background information and related work in the areas of text extraction engines, data augmentation, and precision health systems; Sect. 3 describes the methodologies employed in formulating our approach and developing the datasets; Sect. 4 delves into two case studies, each applying the methodologies outlined in the previous section, and provides an analysis of the results; Sect. 5 provides a conclusion of our case studies, including insights into future work.

Related work

This section discusses the related contributions and work that lay the foundation for exploring our proposed methods, focusing primarily on text extraction engines, the challenges in language learning tasks, and their influence on precision health systems. We examine both the successes and hurdles in this domain to establish a comprehensive background for this paper. The exploration begins with text extraction engines, discussing their notable capabilities. This is followed by an examination of data augmentation strategies and the role of language models in precision health research. Finally, we discuss the applications of noise correction facilitated by modern transformer architectures, highlighting their significance in the current landscape of language processing and health informatics.

Text extraction engines

Text extraction engines function as systems that transform various data formats into a character-encoded, human-readable format [2]. This process enables the conversion of a wide range of file types, such as JPEG, WAV, MP3, and others, into text formats that are easily interpretable by computer systems and machine learning algorithms [3]. Among these engines, OCR engines are particularly prominent. OCR systems are

adept at converting text from a variety of image file types into plain text, making them a widely used tool for data ingestion in numerous fields and industries, especially prior to the widespread implementation of digital records. The shift to digital records and information management is relatively recent [18], and many modern neural information retrieval and text mining methods still struggle with insufficient data samples, limiting their performance when compared to more established open-source alternatives [14]. As a result, OCR engines often serve as a vital link between the eras of pre-digital and digital records, facilitating access to a broader range of data. However, the quality of the output from these engines is crucial, as it determines whether the expanded, historical datasets are beneficial or potentially detrimental.

While OCR systems universally aim to extract data from images and convert it to plain text, the specific implementation methods can vary significantly. Generally, OCR systems are structured as data processing pipelines comprising several optimized stages. A typical OCR pipeline includes three main processing stages: a preprocessor, a text extractor, and a post-processor [1, 2]. The design of these systems can vary, with some employing a sequential linkage of components, while others might use probabilistic inference to dynamically adjust and route an image for optimal processing and resource allocation. This variance in design allows for flexibility in handling different types of images and text, ensuring the most efficient processing and resource usage. The process flow of a generic OCR pipeline can be visualized in Fig. 1, illustrating how an image progresses through these stages, from initial preprocessing to final post-processing.

The diversity of OCR engines presents both benefits and challenges that should be carefully considered when developing a tailored application pipeline. Not every OCR system has been trained on extensive datasets, incorporates effective segmentation techniques, or benefits from a supportive external community that can contribute to enhancing its capabilities.

Data augmentation

Data augmentation is a technique that generates synthetic variations of samples from existing datasets, aiming to enhance the decision-making boundaries in specific inference tasks [12]. This approach, which introduces augmented data into training, acts as a form of implicit regularization. By diversifying the range of samples for a particular labeled task, the risk of overfitting is reduced, leading to a more generalized and robust

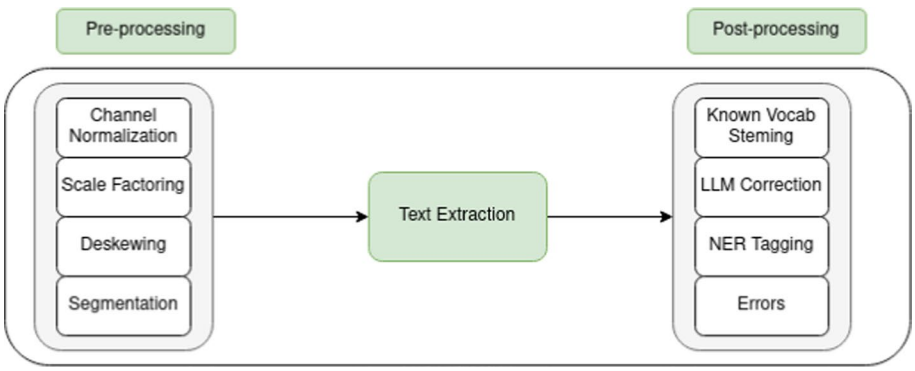

Fig. 1 Example of generic OCR pipeline design strategy

decision boundary that performs better on new, unseen data. This method is especially popular for pre-training versatile models that are later fine-tuned for specific applications in downstream or transfer learning scenarios. Among the most common forms of data used in this context are images and text. We explore various techniques for implementing data augmentation in these two modalities.

Generative text data augmentation

In its most basic form, generative text data augmentation involves creating synthetic data that slightly or moderately deviates from the original data. The extent of variation introduced through these mutations is typically optimized to enhance a model's tolerance or performance for a specific application. For instance, overly drastic changes in a dataset's distribution due to augmentation might degrade model performance, while minimal changes could lead to overfitting, as noted in sources [12]. A prominent example of such tailored data augmentation can be seen in the use of token-based language models [7, 19].

Transformer architectures, recognized for their strength and robustness in deep learning, are based on encoder-decoder structures and feature a multi-head attention mechanism. This mechanism is proficient at capturing complex nuances in sequential, probabilistic, and layered contexts within batched query inputs [20]. Common augmentation strategies leveraging this architecture typically focus on token substitution and masked language modeling. These strategies involve generating samples constrained by a specific corpus. However, such approaches can detrimentally impact the performance of downstream tasks, particularly when inferring tokens not encountered during training. Text with inherent 'noise' often results in decreased task efficiency and performance compared to clean, well-tokenized text [8, 10]. This performance degradation arises not from the language model itself, but from the tokenization logic. Most tokenizers utilize special tokens, including masked tokens, special characters, and tokens for unknown elements. When a tokenizer encounters a token outside its domain-specific vocabulary, such as an erroneous word or a mistranslated token-piece, it often processes it as an ambiguous 'context to unknown' mapping. This increases the ambiguity of the unknown token/sequence association and can adversely affect task performance [7]. To address these issues, new strategies have emerged, such as using Masked Language Modeling (MLM) over tokenless models like Google's ByT5, introducing common errors into a domain's corpus, and employing autoregressive fitting for language correction prior to tokenization.

Denoising auto encoders and data augmentation

Denoising Auto Encoders (DAEs) are a framework that capitalize on the concept of reducing feature space before reconstructing it. This process allows for the variations in corrupted input sequences of augmented samples to be refined and simplified prior to reconstruction. One of the methods employing this technique is Transformers and Sequential Denoising Auto-Encoder (TSDAE), which introduces an arbitrary noise function to deliberately corrupt data points. The key actions of these noise functions include substituting, adding, or removing characters within a sample, thereby introducing variance at the feature distillation stage. Additionally, open-source libraries like NLPAug

play a pivotal role. NLPAug acts as a simulator for typical errors encountered in digital inputs, audio, and visual data extraction, making it valuable for data augmentation and regularization tasks [21]. By generating more sophisticated errors, many applications have experienced improvements in classification performance. Moreover, this approach has led to innovative methods for unsupervised domain adaptation, leveraging the combined strengths of data augmentation and the distillation capabilities of DAEs.

Selecting appropriate augmentation methods

One of the primary objectives in employing learning and feature distillation techniques is to fine-tune the extent of knowledge transferred from the teacher (encoder) to the student (decoder). Effective data augmentation can play a crucial role in this optimization process, provided that the augmentation adds relevant and 'good' data to the training set. The effectiveness of data augmentation can be evaluated by assessing a model's ability to generalize in the presence of noise. In the context of language models, factors like tokenization schemes and vocabulary size are often crucial considerations. For architectures like Bidirectional Encoder Representations from Transformers (BERT), which rely heavily on language perplexity as a key performance indicator due to pretraining tasks like masked language modeling, the quality of both individual tokens and entire sentences is critical in determining the success of an augmentation strategy. However, introducing noise, if not done carefully or if excessively applied, can impede these corrective tasks. Excessive noise can potentially hinder rather than help the model, leading to decreased performance.

Precision health systems

In the clinical healthcare sector, about eighty percent of the data consists of unstructured information, which is often entered manually. This includes various data types, such as text and images, which are often underutilized and eventually end up in large data repositories before being offloaded and discarded. A major challenge in integrating modern natural language processing techniques into healthcare is the limited open-source support and restricted public access to data, a contrast to other fields. This limitation hinders crowd-sourced contributions in integrating tools, developing methods, and making pre-trained large language models available to the public. To address this issue, the MIMIC critical care database was released publicly, facilitating research that uses its clinical text to create publicly accessible, pre-trained language models. This initiative fosters community involvement in adapting sophisticated language processing techniques for healthcare applications.

Due to the prevalence of unstructured data formats in healthcare records, such as faxes or scanned images, there is a need for extensive preprocessing, text correction, and quality checks to effectively use this data in standard machine learning and artificial intelligence applications. This paper proposes expanding the MIMIC-III and MIMIC-IV databases to include mock electronic health records that mimic common formats such as those of the CMS. This expansion would involve recreating patient notes, claims, and laboratory documents as plain images, offering a means for accessible, modifiable, and reproducible documents. Such an initiative would aid in advancing reliable text extraction methods for clinical natural language processing.

As interest in precision health research escalates, there is an increasing need to aggregate various types of data, such as insurance claims, billing invoices, and patient health records [18]. This surge, coupled with the challenge of managing large volumes of heterogeneous unstructured data, calls for an intelligent and automated approach to effectively apply artificial intelligence in these contexts. A key strategy to streamline this process and expedite the annotation of documents and entities within unstructured text is the use of application-specific tasks trained on domain-specific language transformers [22]. By tailoring these tasks to mirror the domain expertise of professionals who label and annotate data in precision health systems, we can assess the efficacy and practicality of these specialized tasks against both general and domain-focused language embedding spaces.

The digitization of healthcare and medical records has led to the creation of clean, structured data repositories. However, in the domain of clinical notes, laboratory event logging, and other specialized data sources that underpin the administrative aspects of healthcare, much of this information remains unstructured [22]. The structure of this data often varies based on the context in which it is analyzed. These sources, encompassing patient notes, lab procedures, event records, and medical billing or claim invoices, each possess unique characteristics. This variety poses a challenge in developing uniform methods for extracting meaningful information from these documents.

A specific challenge encountered in applying language transformation to these documents is the inadvertent introduction of noise due to policy-driven censorship aimed at protecting patient data, often referred to as Protected Health Information (PHI).Consequently, processing text in its entirety may not always be feasible, depending on the data access rights of a particular entity or task. In light of these constraints, domain-specific language transformers, especially those tailored for clinical use, could offer more flexible solutions. This paper will examine how such specialized transformers can help navigate these challenges.

Building on recent advancements, this paper explores the transformation of general BERT uncased models into specialized versions like BioBERT, BlueBERT, and ClinicalBERT. We also examine the potential of ByT5 and language transformers enhanced with medical terminology prompts, introducing novel approaches in functional classification. The objective is to utilize natural language processing and transformers for accelerating clinical and administrative tasks. These tasks include relevance search and classification of invoices, as well as the classification of unstructured documents, thereby leveraging the power of language models to streamline these complex processes.

Our experiments evaluate the effectiveness of using contextual patterns derived from one language transformer and applying them to another for specialized classification tasks. This approach aims to replicate domain-specific knowledge to enhance document classification processes. The datasets employed in our experiments feature medical language in a mixed context. This means the documents, including medical service invoices and clinical event texts from doctors and nurses, contain both medical terminology and standard English. These texts are sourced from the MIMIC-III and MIMIC-IV databases, along with a custom dataset of assorted clinical invoices gathered from various online platforms. These data structures are presented in various formats, such as relational tables, electronic forms, and images. They were first preprocessed, then processed

through an OCR engine, and finally prepared for labeling and embedding. This comprehensive approach allows for a thorough examination of the applicability and efficiency of language transformers in handling diverse and complex medical data structures.

Methods

The design of our datasets is meticulously crafted to facilitate the generative creation of images suited for OCR analysis. These images are deliberately embedded with common cosmetic flaws known to induce more errors in the output of their data processors. We have developed this pipeline following the principles of a simple abstract factory software design pattern. This choice allows us to maintain both the structural integrity of the documents and the fundamental logic for base assignment, while simultaneously implementing various behavioral flaws to generate our augmented samples. In assessing the effectiveness of sequence-to-sequence learning, we use these specially crafted data points in an auto-regressive learning framework. This framework aims to discern and learn from the patterns inherent in the deliberately introduced imperfections. Each experiment within our study is structured to encompass three primary processes: ingestion, transformation, and document generation.

In Fig. 2, the pipeline designed for processing CC News data is depicted, featuring a modular approach with document structure manipulation and OpenCV [23] noise modules as interchangeable components. This modular design is instrumental in enabling the production of documents with variable structural attributes and imperfections, which is vital for the realistic replication of OCR errors. The flexibility of the pipeline components to be substituted or altered allows for extensive adaptability across different OCR engines and error scenarios, which is pivotal for the training of OCR error correction models. For example, pytesseract could be replaced with MultiMedia Optical Character Recognition (MMOCR) [24].

The methodology, while broad in scope, is intentionally crafted to accommodate a range of document conditions and noise patterns. This approach ensures that the pipeline can effectively mimic the variability encountered in actual OCR processing environments. The innovative aspect of this work is not merely in the individuality of each component but in their collective integration into a cohesive system that has been shown to improve the training outcomes for OCR error correction models more efficiently than conventional methods.

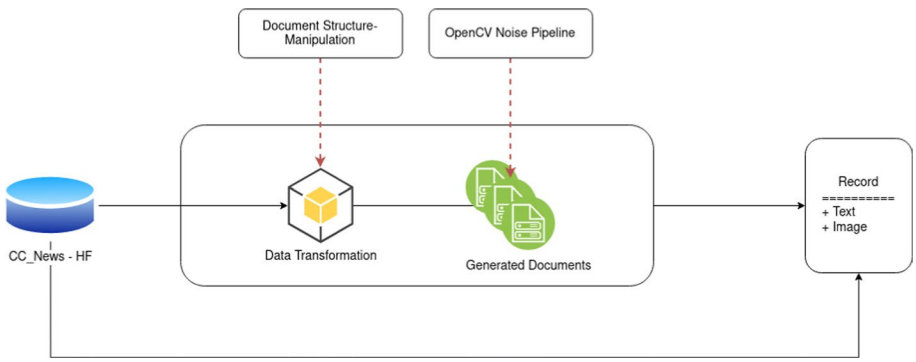

Fig. 2 Logic workflow of generative augmentation pipeline

CC news generative data

In typical applications, generative data augmentation spans various multi-modal data types, such as audio, images, and text. For our study, we concentrated on the last two: images and plain text. We developed a method to generate samples in these formats and then evaluated their quality and practicality in comparison to a controlled version of the data. This controlled version was specifically modified to mimic the most common errors encountered during document translation. Factors included in our assessment were font size variation, image resolution differences, character encoding issues, and challenges in page segmentation. This comparison aims to understand the effectiveness of our augmentation approach against real-world challenges in OCR data processing.

Ingestion Our process begins by acquiring a large, domain-agnostic textual data source that is compatible with general-purpose text processing tools without the need for fine-tuning. For this task, we chose news data from the CC News dataset found in Hugging Face's data repository. This particular dataset was selected due to the existing summarization and manual annotation it provides, which aids in downstream learning tasks. These features allow us to analyze the quality of our augmentations against a well-established benchmark. To ingest this data efficiently, we utilized the dataset's native API, which offers general control and efficient utilization of computational resources. Future research could explore optimal parallelization strategies for enhanced performance during ingestion, possibly employing tools like Spark or Polars APIs.

Transformation The key benefit of using an OCR engine's multimodal capabilities to generate our text output, rather than creating it ourselves, lies in the control mechanism offered by document templating. We employed the FPDF2 library to establish a versatile template for producing our plain text image documents, with tunable parameters unique to each object. Shared instance traits, such as font size, resolution, and spacing, are considered object-independent parameters. In contrast, object-dependent parameters include document shape, page segmentation, and layout imperfections. We add complexity to these templates by using the PDF object in the FPDF2 library and applying the abstract factory method to create intricate documents.

The success of this case study hinges on not just the volume of errors generated by replicating system noise, but also on the quality of this noise. Previous methods have involved attaching generative noise through random sampling and insertion/deletion at the inference stage of a corrective system. Meanwhile, other correction techniques without probabilistic inference rely on fixed rule sets to address various error types encountered during manual text processing.

Generation The final phase of our methodology is document generation. We apply the preprocessing steps laid out for each task sequentially to create a basic document. These documents are then generated as plain text articles, each paired with the desired displayed text and a metadata entry. This metadata includes subclassification information for each document, along with the accurate text associated with our generated images. This approach allows for a comprehensive understanding of the documents' content and the effectiveness of our generative processes.

MIMIC-III(IV) generative data

MIMIC versions III and IV are clinical care databases that have undergone processes of de-identification, transformation, and re-modeling, structured as modules for various observational layers. Our approach utilizes the relational structure of these databases to generate a layer that facilitates the mocking and augmentation of clinical documents. With the rapid integration of clinical data systems in healthcare, many existing infrastructures require manual data correction and extensive document ingestion. The primary goal of these systems is to transform vast repositories of documented information into formats that can be interpreted, processed, and modeled by computational systems for various downstream applications.

However, the private nature of many of these data repositories often leaves the task of intelligent, corrective ingestion to individual organizations. These organizations may lack the necessary resources to effectively address this challenge. Our objective is to bridge this gap by introducing a publicly supported solution that can be integrated into the workflows of industry researchers and engineers.

Ingestion In the initial phase of implementing the MIMIC-IV 'mock docs', we generate two types of unstructured document simulations: one is structured medical invoice data in the form of CMS-1500 tables, and the other consists of free-text clinical lab reports. The primary source for reconstructing these documents is the Admissions table, which serves as the foundation for synthesizing the required information. This synthesis involves organizing unique patient visits by criteria such as admission status, diagnosis, and insurance provider, and subsequently aligning them with Current Procedural Terminology (CPT) events and Lab events to formulate a mock generative persona. This persona equates to fabricated patient information associated with each admission ID for document creation.

For data manipulation, we utilize the Polars API, known for its efficient handling of columnar data structures and its capability to support parallel processing at the column level. This ensures quicker data retrieval and processing for the initial staging tables. Alternatively, Structured Query Language (SQL) can be utilized for similar data structuring and manipulation, offering a more traditional approach to applying these methods.

Transformation The MIMIC database's primary strength lies in its adaptable design and comprehensive relational structure. This versatility enables researchers to examine the patient care infrastructure with varying levels of detail and tailor the data to suit specific downstream applications. In the process of creating mock documents, we combine data from the Admissions table with the CPT Events table. This involves joining and grouping the data by admission ID to simulate sequential billing entries.

To further enhance the realism of these mock documents, each entry is linked to a persona generated randomly using data from the 1990 s census, along with contact details sourced from the open address dataset. This approach of generating personas, which includes a considerable time gap, random reassignment of gender, and arbitrary contact information, significantly reduces the already low probability of accidentally replicating real patient data. This method is crucial for maintaining patient confidentiality and adhering to the de-identification standards set by the HITECH Act, which builds upon the privacy protections established by HIPAA.

Generation The concept of mocking goes beyond creating patient personas to include the physical replication of documents. This stage involves simulating various cosmetic imperfections in the documents to ensure compliance with PHI regulations. These mock documents are produced using the tabular data structure created in the transformation stage, with each document receiving specific design parameters. This ensures that every persona is represented in a manner consistent with real documents issued by healthcare providers or laboratories.

The method involves taking data entries at the row level and compiling them into a billing invoice format that corresponds to each patient visit. This process not only creates realistic documents but also allows us to introduce controlled variations in the data. These variations are designed to mimic errors that could occur in intelligent data processing systems like OCR engines or Text-to-Text Transformers. For instance, in this experiment, we manipulate aspects of the CMS-1500 billing invoice data structure and plain text prescriptions, such as image size, resolution, and segmentation, to test how these factors impact data processing at the generative stage of document creation. This approach provides valuable insights into the challenges and potential solutions for processing complex medical documents.

Measurement and evaluation

In the initial phase of this work, we determine the exact axis of observation. Our method requires us to consider performance at both the character and line levels, especially because of the segmented nature of our extraction engines. For this purpose, we have chosen two main metrics: CER and WER. These metrics are vital for evaluating the ability of our system to handle character changes (insertions, deletions, etc.) compared to standard OCR engines, and to assess line-level performance. We utilize CER and WER, both rooted in Levenshtein distance calculations, across word and line levels of our OCR data. This approach is key to effectively determining the efficacy of our augmentation strategy.

Edit distance

In our study, we assess the Levenshtein edit distance to identify differences between text strings, focusing on three key transformations: insertions, deletions, and substitutions. Insertions are characters added to the string, deletions are characters removed, and substitutions are instances where the string remains the same length but has different characters.

The error rate (ERR) is quantified using the following equation:

$$ERR = \frac{\sum_i T_i}{N}; \quad T = \{S, D, I, \ldots, Z\}$$

where the set T comprises the types of errors considered:

S : the number of substitutions,
D : the number of deletions,
I : the number of insertions,
Z : any other specific text transformation under observation, and
N : the total number of data samples.

The precision of our analysis depends on the assigned task, reflecting through metrics such as CER and WER.

CER provides a granular assessment of an OCR process by closely comparing recognized characters to a benchmark. Ideal for transformer-based models that bypass tokenization, CER allows for dissecting text down to the smallest unit, ensuring meticulous evaluation.

WER, on the other hand, gauges the fidelity of word sequences during document extraction. It is a cornerstone in language modeling, supporting a range of tasks from word insertion to sequence translation, thereby playing a pivotal role in refining sequence prediction models.

These metrics, derived from the edit distance equation we discussed, serve as essential indicators of OCR accuracy at varying levels of text analysis.

Case studies: results and discussion

To assess the value of different augmentation methods for our dataset, we employed three distinct transformations. First, we used the arbitrary noise function from the TSDAE within the sentence transformers library. Next, we utilized NLPAug, an open-source framework known for its parameterized data transformations that mimic common OCR errors. Finally, we introduced our unique approach OCR2SEQ as an augmentation technique, which involves the generation of mock text documents processed through a standard OCR engine. We hypothesize that the errors stemming from actual text extraction will yield a more effective sample set for language models, leading to a better-trained and more general decision-making framework.

We then proceeded to evaluate the effectiveness of each augmentation method. The aim was to not only measure their overall impact but also to determine if the methods are statistically distinct from each other, thereby ensuring a diverse augmentation approach. To maintain the quality of our original dataset, we carefully balanced the introduction of variability. We analyzed the sample distributions from each transformation, focusing on their error rates to ascertain their statistical differences. For this comparison, we opted for the Kolmogorov-Smirnov (KS) test, as it suited the segmented nature of the data derived from page chunks during document extraction. This allowed us to identify each augmentation method as a valuable and independent strategy.

Mock docs CC news sanalysis

We began by examining the error rate distributions that resulted from applying the three different augmentation strategies to our training data. There was a noticeable shift in the mean computational edit distance for each variant compared to its original version. The tsdae_comp_distance strategy (TSDAE) shows a tendency towards larger, yet more consistent edits. However, it has the least impact on erroneous vocabulary, primarily due to the simplistic nature of noise addition or removal at the sentence level. In contrast, the comp_distance approach (OCR2SEQ) indicates smaller edits and has the broadest distribution. The probability density distributions for these single-page CER edits are shown in Fig. 3.

We then assessed the impact on vocabulary by analyzing the summary of vocabulary disjoint per augmentation strategy, as displayed in Fig. 4. From this plot, one

could infer that the 'title_disjoint_score' method (OCR2SEQ) potentially introduces the most variability in vocabulary disjointness, as it has the broadest interquartile range.

Further, we quantified the overall edit distance summary per strategy, an essential metric for evaluating the augmentation impact, which is shown in Fig. 5. The box plot shows that ocr_edit (OCR2SEQ) tends to introduce the most changes, as indicated by its broadest interquartile range.

Additionally, we conducted a KS 3-way analysis to statistically validate the distinctness of each augmentation strategy, with the results summarized in Table 1. From the table, it is clear that all three pairwise comparisons show statistically significant differences, which means that each augmentation strategy can be said to affect the text data in a significantly different way.

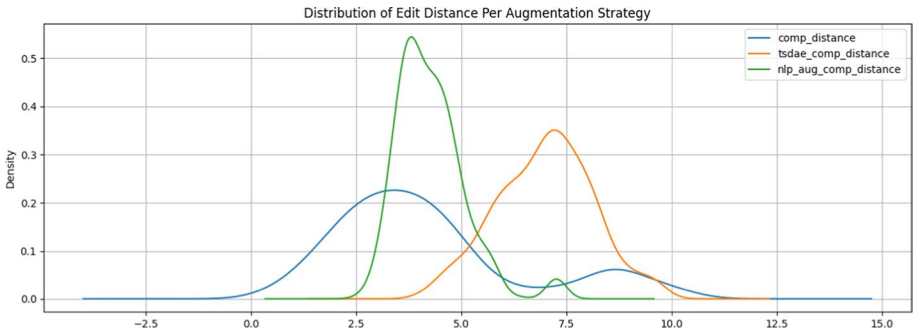

Fig. 3 Probability density distribution for single page CER edit distance

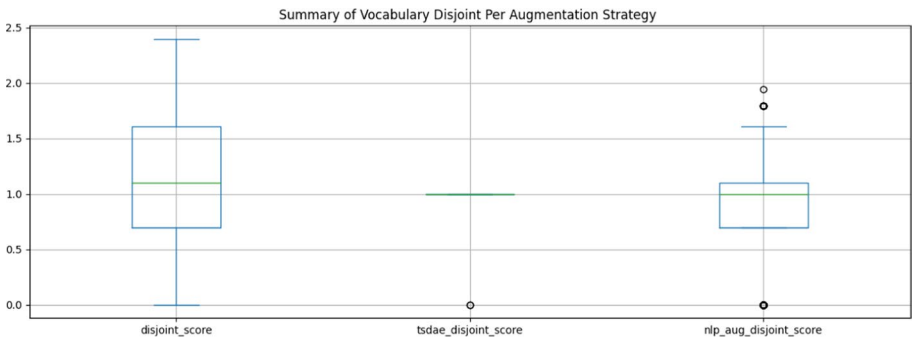

Fig. 4 Augmentation strategy box plot for single page WER

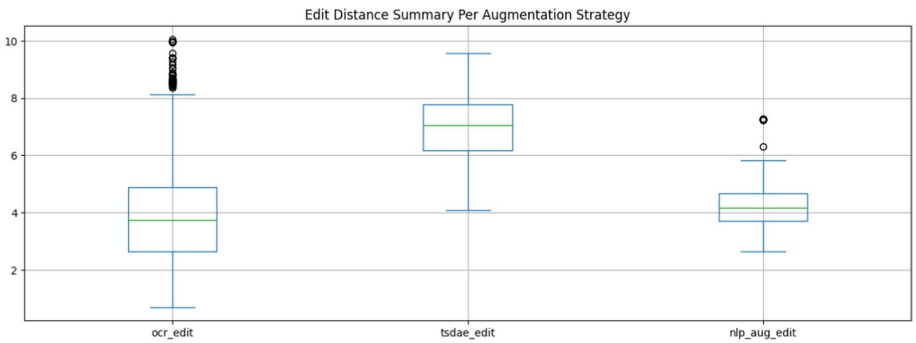

Fig. 5 Edit distance summary statistics

Table 1 Kolmogorov-Smirnov 3-way analysis on CC news documents

Group 1	Group 2	KS distance	p-value
tsdae edit	ocr edit	0.4705	1.85×10^{-200}
ocr edit	nlpaug edit	0.9715	0
nlpaug edit	tsdae edit	0.852	0

Table 2 Summary statistics of each data augmentation strategy for CC news documents

	ocr edit	tsdae edit	NLPAug edit
Mean	4.209219	6.984248	4.274464
Atd	2.263670	1.137885	0.816949
Min	0.693147	4.094345	2.639057
Max	10.065861	9.581076	7.261225

The summary statistics for each data augmentation approach revealed notable differences in terms of mean, standard deviation, minimum, and maximum values, which are detailed in Table 2. These statistical measures emphasize the distinct behavior of each augmentation strategy. Specifically, the standard deviation values offer insights into the consistency of each method, with higher standard deviation indicating greater variability introduced by the augmentation process. These statistics are not just numbers. They represent the potential of each method to handle the complexity of real-world data in training robust language models.

Our comparative analysis indicates that while TSDAE introduces noise at a macro level, impacting the edit distance more significantly, NLPAug's micro-level control allows for a nuanced approach to error introduction. Our OCR2SEQ augmentation method, however, stands out by adding a rich diversity of errors, as evidenced by the wider range of unique noises shown in Fig. 5. This diversity could be key in creating more resilient models capable of understanding and correcting a wider array of errors seen in real-world OCR applications.

Mock Docs MIMIC-IV analysis

In our analysis of the MIMIC-IV mock documents, a noticeable shift was observed in the distribution patterns compared to the plain text documents sourced from CC News. This altered distribution tended to be bimodal after implementing our trio of data augmentation techniques. The probability density distributions for single and multi-page CER edits are shown in Fig. 6 and Fig. 7, respectively. These figures illustrate the impact of our augmentation techniques on the error rate distributions for both single and multi-page documents. In Fig. 6, tsdae_edit (TSDAE) is associated with notable edits but demonstrates a narrower range of variation. Conversely, ocr_edit (OCR2SEQ) exhibits a wider spectrum of smaller-scale edits. Moving to Fig. 7, a comparable trend is observed: TSDAE continues to generate more pronounced edits with a tighter distribution, whereas OCR2SEQ maintains its tendency to produce a diverse array of minor edits.

We then explored factors contributing to this pattern, which was consistently observed across all three data transformations. One key finding was the impact of

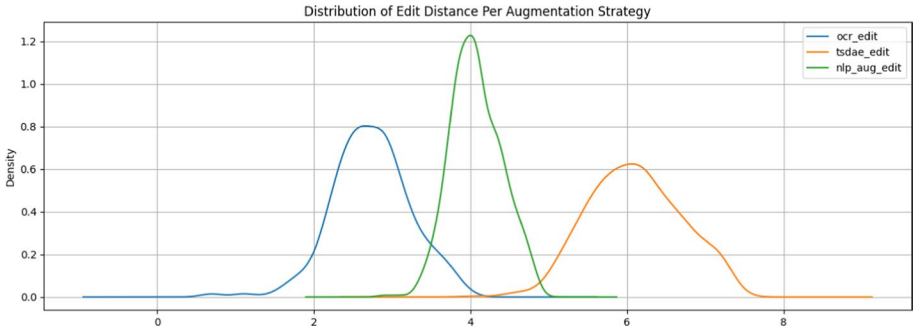

Fig. 6 Probability density distribution for single page CER edit distance

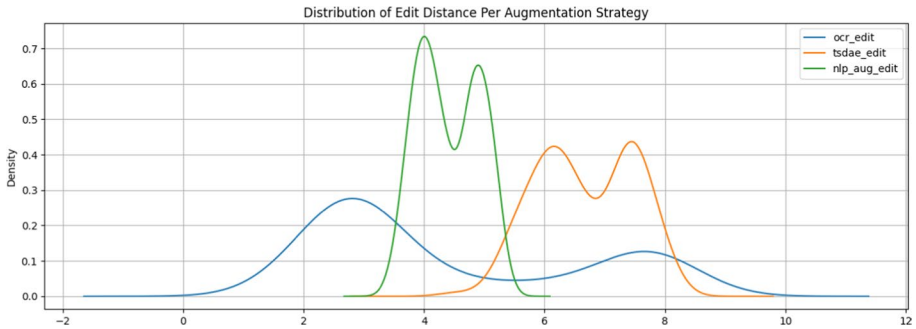

Fig. 7 Probability density distribution for multi-page CER edit distance

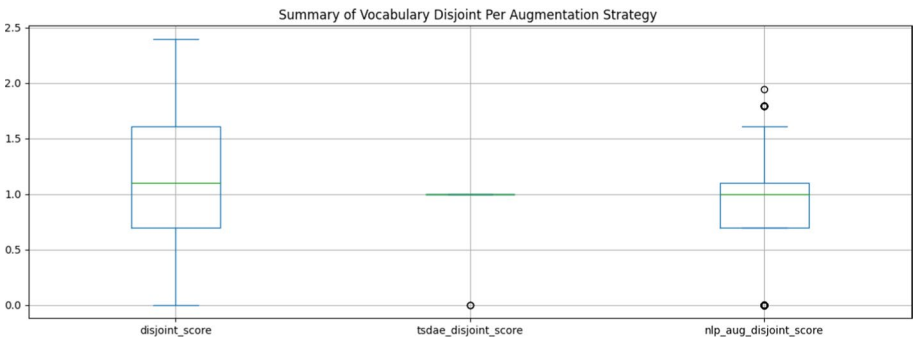

Fig. 8 Augmentation strategy box plot for single page WER

using increasingly large text sources, which necessitated mapping onto multiple pages. This approach seemed to directly influence the distribution pattern, as shown in the box plots for single and multi-page WER in Fig. 8 and Fig. 9. In Fig. 8, the box plot illustrates that disjoint_score (OCR2SEQ) exhibits the most variation in modifications, evidenced by its wide interquartile range. However, in Fig. 9, the nlp_aug_wer approach (NLPAug) demonstrates the greatest spread in values, as shown by its expansive interquartile range.

To examine more thoroughly this shift in distribution, we adjusted our sample to ensure a uniform page count and consistent transformation process for each page. We hypothesized that the noise introduced by the TSDAE and NLPAug techniques might

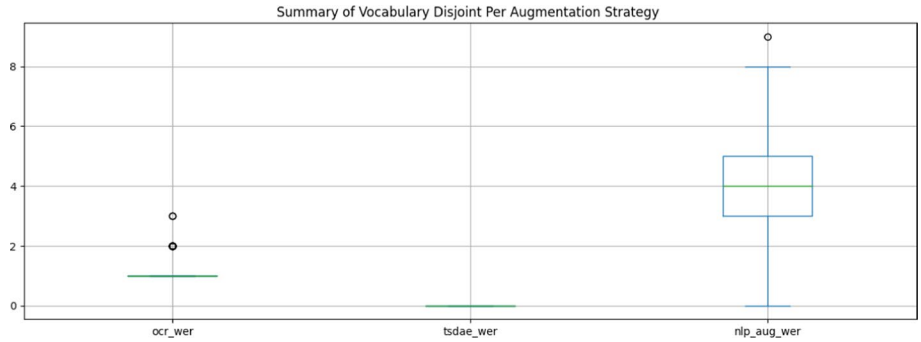

Fig. 9 Augmentation strategy box plot for multi-page WER

vary from page to page. The KS 3-way analysis, summarized in Table 3, provided statistical validation of the distinctness of each augmentation strategy, underscoring the effectiveness of our methodology.

Our findings, supported by low p-value scores and high KS scores, confirm the significant difference between each augmentation strategy on data quality. This is evidenced by the unique edit distances and disjoint sets for each sample, summarized in Table 4. These differences in distribution are crucial for validating the effectiveness of various augmentation types in enhancing machine learning denoising tasks.

It is essential to clarify that our evaluation relies on key metrics such as word error rate (WER), character error rate (CER), and edit distance, with distributions depicted in Figs. 3 to 5 for the CC News dataset and Figs. 6 to 9 for MIMIC. OCR2Seq's ability to consistently achieve the lowest error across different data types is a testament to its efficacy and versatility.

Conclusion

This study demonstrates OCR2SEQ's significant advancement in OCR technology through its multi-modal generative augmentation, setting a new benchmark in the field. OCR2SEQ's ability to outperform TSDAE and NLPAug is particularly notable in processing data with sparse characters and domain-specific vocabularies. Its application proves invaluable in sectors such as healthcare and library sciences, where

Table 3 Kolmogorov-Smirnov 3-way analysis on MIMIC documents

Group 1	Group 2	KS Distance	p-value
tsdae edit	ocr edit	0.939	0
ocr edit	nlpaug edit	0.825	0
nlpaug edit	tsdae edit	0.946	0

Table 4 Summary statistics of each data augmentation strategy for mimic mocuments

	ocr edit	tsdae edit	nlpaug edit
Mean	2.732798	6.103634	4.073298
Std	0.500384	0.591656	0.323122
Min	0.693147	4.043051	2.890372
Max	3.970292	7.433667	4.875197

the accuracy of OCR is not just beneficial but essential. The effective integration of OCR2SEQ with complex datasets underscores its ability to enhance the interpretation capabilities of machine learning models.

The innovative approach of OCR2SEQ in simulating realistic text extraction errors through its augmentation strategies stands as a testament to its potential in transforming the OCR landscape. By effectively addressing common OCR challenges, OCR2SEQ paves the way for more robust and reliable text-to-text transformations, contributing significantly to the reduction of error rates in OCR processes. The system's effectiveness in various data scenarios highlights its potential as a versatile tool in the ever-evolving field of data processing.

Future work should focus on broadening OCR2SEQ's use in diverse fields and enhancing its algorithms for increased efficiency and adaptability. Moreover, it is essential for OCR2SEQ to adhere to strict data protection and patient confidentiality standards, particularly in sectors like healthcare where the ethical handling of data is crucial. Addressing these ethical and privacy concerns is not only essential for compliance but also for maintaining public trust in OCR technology.

Abbreviations

BERT	Bidirectional encoder representations from transformers
CER	Character error rate
CMS	Centers for medicare and medicaid services
CPT	Current procedural terminology
DAE	Denoising auto encoder
ICD	Internatioal classification of diseases
KS	Kolmogorov-smirnov
LLM	Large language model
MMOCR	MultiMedia optical character recognition
OCR	Optical character recognition
OCR2Seq	Optical character recognition to sequence
PHI	Protected health information
PII	Personally identifiable information
RAG	Retrieval augmented generation
SQL	Structured query language
TSDAE	Transformers and sequential denoising auto-encoder
WER	Word error rate

Acknowledgements
We would like to thank the internal reviewers of this manuscript

Author contributions
ML searched for relevant papers and drafted the manuscript. All authors provided feedback to ML and helped shape the work. ML,JDP, and JLL prepared the manuscript. TMK introduced this topic to ML and helped to complete and finalize the work. All authors read and approved the final manuscript.

Funding
Not applicable

Availability of data and materials
Not applicable

Declarations

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Competing interests
The authors declare that they have no competing interests.

Received: 8 March 2024 Accepted: 16 April 2024
Published online: 13 June 2024

References

1. Patel C, Patel A, Patel D. Optical character recognition by open source OCR tool tesseract: a case study. Int J Comput Appl. 2012;55(10):50–6.
2. Smith R. An overview of the tesseract OCR engine. In: Ninth International Conference on Document Analysis and Recognition (ICDAR 2007). IEEE. 2007;2: 629–633.
3. Jockers ML, Underwood T. Text-mining the humanities. A new companion to digital humanities. Wiley Online Library. 2015;291–306.
4. Lihui F, Underwood T. The core issues and latest progress of current digital humanities research: An interview with ted underwood. Foreign Lit Stud. 2021;43(6):1.
5. Cleland I, Han M, Nugent C, Lee H, Zhang S, McClean S, Lee S. Mobile based prompted labeling of large scale activity data. In: Ambient Assisted Living and Active Aging: 5th International Work-Conference, IWAAL 2013, Carrillo, Costa Rica, December 2-6, 2013, Proceedings 5, 2013; 9–17. Springer.
6. Berabi B, He J, Raychev V, Vechev M. Tfix: Learning to fix coding errors with a text-to-text transformer. In: International Conference on Machine Learning. 2021; 780–791. PMLR.
7. Srivastava A, Makhija P, Gupta A. Noisy text data: Achilles-heel of bert. In: Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020), 2020;16–21.
8. Xue L, Barua A, Constant N, Al-Rfou R, Narang S, Kale M, Roberts A, Raffel C. Byt5: towards a token-free future with pre-trained byte-to-byte models. Trans Assoc Comput Linguist. 2022;10:291–306.
9. Papanikolaou Y, Staib M, Grace J, Bennett F. Slot filling for biomedical information extraction. arXiv. 2021. https://doi.org/10.48550/arXiv.2109.08564.
10. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res. 2020;21(1):5485–551.
11. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. J Big Data. 2019;6(1):1–48.
12. Shorten C, Khoshgoftaar TM, Furht B. Text data augmentation for deep learning. J Big Data. 2021;8:1–34.
13. Were MC, Sinha C, Catalani C. A systematic approach to equity assessment for digital health interventions: case example of mobile personal health records. J Am Med Inform Assoc. 2019;26(8–9):884–90.
14. Johnson A, Bulgarelli L, Pollard T, Horng S, Celi L.A, Mark R. Mimic-iv. PhysioNet. 2020. https://physionet. org/content/mimiciv/1.0/. Accessed 23 Aug 2021.
15. Johnson AE, Pollard TJ, Shen L, Lehman L.-w.H, Feng M, Ghassemi M, Moody B, Szolovits P, Anthony Celi L, Mark R.G. Mimic-iii, a freely accessible critical care database. Sci Data. 2016;3(1):1–9.
16. Hugging Face Team: Hugging Face Data Repository. https://huggingface.co/datasets. Accessed 22 Feb 2024.
17. Saoji S, Eqbal A, Vidyapeeth B. Text recognition and detection from images using Pytesseract. J Interdiscip Cycle Res. 2021;13:1674–9.
18. Shull JG. Digital health and the state of interoperable electronic health records. JMIR Med Inform. 2019;7(4):12712.
19. Shushkevich E, Alexandrov M, Cardiff J. Bert-based classifiers for fake news detection on short and long texts with noisy data: a comparative analysis. In: International Conference on Text, Speech, and Dialogue. Springer. 2022: 263–274.
20. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Adv Neural Inform Proc Syst. 2017;30.
21. Ma E. NLP Augmentation. GitHub. 2019. https://github.com/makcedward/nlpaug.
22. Li M, Lv T, Chen J, Cui L, Lu Y, Florencio D, Zhang C, Li Z, Wei F. Trocr: transformer-based optical character recognition with pre-trained models. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2023;37:13094–102.
23. Bradski G. The OpenCV Library. Dr. Dobb's. J Softw Tools. 2000;120:122–5. https://opencv.org.
24. Kuang Z, Sun H, Li Z, Yue X, Lin TH, Chen J, Wei H, Zhu Y, Gao T, Zhang W, Chen K, Zhang W, Lin D. MMOCR: a comprehensive toolbox for text detection, recognition and understanding. arXiv. 2021. https://doi.org/10.48550/arXiv.2108.06543.

Publisher's Note