International Journal of Medical Informatics 144 (2020) 104302

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

Automatic classification of scanned electronic health record documents

Heath Goodrum [a], Kirk Roberts [a], Elmer V. Bernstam [a,b,*]

[a] School of Biomedical Informatics, The University of Texas Health Science Center at Houston, TX, United States
[b] Division of General Internal Medicine, McGovern Medical School, The University of Texas Health Science Center at Houston, TX, United States

ARTICLE INFO

Keywords:
Electronic health records
Scanned documents
Classification
Optical character recognition
Machine learning
Patient safety

ABSTRACT

Objectives: Electronic Health Records (EHRs) contain scanned documents from a variety of sources such as identification cards, radiology reports, clinical correspondence, and many other document types. We describe the distribution of scanned documents at one health institution and describe the design and evaluation of a system to categorize documents into clinically relevant and non-clinically relevant categories as well as further sub-classifications. Our objective is to demonstrate that text classification systems can accurately classify scanned documents.

Methods: We extracted text using Optical Character Recognition (OCR). We then created and evaluated multiple text classification machine learning models, including both "bag of words" and deep learning approaches. We evaluated the system on three different levels of classification using both the entire document as input, as well as the individual pages of the document. Finally, we compared the effects of different text processing methods.

Results: A deep learning model using ClinicalBERT performed best. This model distinguished between clinically-relevant documents and not clinically-relevant documents with an accuracy of 0.973; between intermediate sub-classifications with an accuracy of 0.949; and between individual classes with an accuracy of 0.913.

Discussion: Within the EHR, some document categories such as "external medical records" may contain hundreds of scanned pages without clear document boundaries. Without further sub-classification, clinicians must view every page or risk missing clinically-relevant information. Machine learning can automatically classify these scanned documents to reduce clinician burden.

Conclusion: Using machine learning applied to OCR-extracted text has the potential to accurately identify clinically-relevant scanned content within EHRs.

1. Introduction

Unstructured scanned documents are frequently found in electronic health records (EHRs) [1]. Documents may be scanned into the EHR for multiple reasons including supporting clinical care, administration or regulatory compliance. Administrative documents include driver's licenses, insurance cards, and payment records. Examples of clinically relevant scanned documents include prescriptions, radiology reports, laboratory results, and depression screenings.

Ideally, scanned documents are deposited into appropriate EHR categories. Thus, scanned laboratory tests are in the same folder as structured laboratory tests, separated from administrative records. However, sometimes documents from a single source, such as an outside provider, are scanned into a single document and placed into general folders such as "outside records." Clinically-relevant documents may be harder to find when mixed with non-clinically relevant documents. Clinicians may miss important, time-sensitive information, such as abnormal laboratory results. Thus, uncategorized clinically relevant scanned documents may pose a risk to patient safety [2,3].

In this paper, we describe the distribution of scanned documents at our institution. Then, we describe the design and evaluation of a system to categorize scanned documents into clinically relevant and non-clinically relevant categories as well as further sub-classifications. Specifically, we evaluated optical character recognition (OCR) and text classification models trained on documents that were previously manually classified to determine whether this approach could accurately classify scanned documents within an EHR.

* Corresponding author at: School of Biomedical Informatics, The University of Texas Health Science Center at Houston, 7000 Fannin St., Suite 600, Houston, TX 77030, United States.
E-mail address: Elmer.V.Bernstam@uth.tmc.edu (E.V. Bernstam).

https://doi.org/10.1016/j.ijmedinf.2020.104302
Received 28 June 2020; Received in revised form 18 September 2020; Accepted 9 October 2020
Available online 17 October 2020

2. Background and significance

To our knowledge, this is the first published attempt to automatically classify scanned documents within the EHR. However, prior work has addressed text/image classification and separately, to a lesser extent, scanned documents within the EHR.

2.1. Text and image classification

Document classification, including free text and images, is a well-studied problem in computer science and information technology (IT) [4]. Document classification algorithms accept some type of data and output a class label that describes the data. Binary classification algorithms assign one of only two possible labels. For example, a binary classification algorithm might accept scanned document pages and determine whether the document page is clinically relevant vs. not. In contrast, multiclass classification algorithms assign one of multiple possible labels. For example, the document page might be classified as a laboratory result vs. radiology report vs. insurance card.

With more and more data being collected, classification may be seen as a first step in creating information from data. Data mining and machine learning techniques have been successful in classifying a wide range of different data types. There are numerous examples of biomedical classification including detecting and classifying lesions in mammograms [5], classifying images of skin lesions into skin cancer classes [6], and sentiment analysis of HPV vaccine-related tweets [7].

In text classification, the field has shifted from high-dimensional, sparse representations (such as "bag of words"), to low-dimensional, dense representations ("embeddings"). The latter range from context-free word embeddings such as word2vec and GloVe to contextual embeddings such as BERT [8–12]. With word embeddings, words that share a semantic relationship are also similar when projected into a multi-dimensional vector space.

2.2. Images and scanned documents within electronic health records

There has been relatively little work on the use of scanned documents within EHRs. Scanned documents are a type of unstructured data contained in the EHR with features similar to both text and images. The prevalence of scanned documents in EHRs has not been evaluated.

In 2003, document imaging was described as a "valuable bridge" for importing historical paper medical records into new EHR systems [13]. Including scanned documents in the EHR allows these documents to be accessed by multiple clinicians and administrative staff simultaneously, and decreases the need to store physical records [14]. While it was likely assumed that scanned documents would be a temporary "bridge" on the path to complete EHR interoperability, in 2020 they are still very much a part of clinical reality and are likely to persist for years into the future.

A fundamental issue with scanned images is that they are not searchable since they do not contain discrete data elements [14]. Thus, it may be difficult to locate specific information (e.g., locate a specific test result) or determine whether something has been done (e.g., has a specific test been done in this patient?). Identifying a subset of scanned documents that are clinically-relevant has the potential to ease clinician burden by reducing the amount of information to read and may improving patient safety by increasing the likelihood that clinically relevant information is in fact read by the clinician.

2.3. General domain OCR document processing

Although there has been relatively little academic research regarding OCR in the context of EHRs, there has been a great deal of work in the general domain under the broad category of digital image processing. An extensive review of this work is beyond the scope of this paper. However, generally OCR applications can be classified based on the type of target document such as hand written vs. machine printed vs. special applications (e.g., specific languages, historical). Although specifics vary, OCR pipelines generally include image acquisition (in our case, scanning), pre-processing, segmentation, feature extraction, classification and post-processing [15]. Pre-processing involves manipulating the image in hopes of improving OCR performance and may include edge enhancement, thresholding and noise reduction. Segmentation in this context refers to dividing the image into regions (e.g., containing text vs. not). Feature extraction refers to transforming each character to a vector representation. Then some classifier, such as neural network or support vector machine (SVM), is applied to the vectors in order to recognize characters. Finally, the pipeline may include a post-processing step such as spell checking.

Once the document is converted to digital text using OCR, it can be classified into categories. Multiple approaches have been tried in the general domain including a variety of document representations leveraging Term Frequency – Inverse Document Frequency (TF-IDF), latent Dirichlet allocation (LDA) and Doc2Vec, but no representation is found to consistently outperform others across tasks [16]. Similarly, multiple classification approaches have been used including linear models such as SVMs [17]. Recently, various approaches based on neural networks have become popular.

3. Methods

We evaluated the scanned documents within an Allscripts EHR system at a single urban healthcare institution over a one-year period from 01-01-2018 to 12-31-2018. This EHR system is used by a network of over 100 outpatient, multi-specialty clinics across the greater Houston area. A single scanned document within an EHR may contain one or more pages. Some scanned documents such as depression screenings were relatively standardized across clinics. However, many clinics created custom forms such as local History & Physical templates to suit their particular needs. Custom document types (or categories) could also be created. At the time of this analysis, there were 129 document types. In this study, we included the scanned documents belonging to the 41 most frequent document types in the EHR. These 41 document types contained 3,352,969 documents (93.8 %) out of a total 3,574,783 documents created during 2018. The 90 document types excluded from our study contained fewer than 400 documents per document type.

When document type labels contained obvious misspellings, truncations or naming variants, we standardized spelling and combined the labels into a single category. We combined six pairs of document types, specifically: "Disclosure Authoriza" and "Disclosure Authorizations", "External Medical Rec" and "External Medical Records", "H&P for Surgeries" and "History Physical", "Questionaire/Consent" and "Questionnaire/Consents", "Referrals/Authorizat" and "Referrals/Authorzations", and "Drivers License" and "Drivers Liscense". Additionally we excluded, "External Medical Records" since these contain a mixture of other document types. We then reviewed 10 randomly-selected documents from each document type in order to understand the contents of each document type and identify categorization errors.

A subset of 65,860 documents (192,074 pages) of documents created during a one-month period from 01-01-2018 to 02-01-2018 were used to train and evaluate a classification model. We observed that scanned document images varied depending on the source (e.g., a particular clinic or hospital) with various logos and variable formatting. However, most scanned documents in the EHR contained text, we hypothesized that clinically-relevant documents of a particular type would have relatively similar text (e.g., chest x-ray reports would have a description of the lungs) and thus chose to classify scanned documents based on their text content using Optical Character Recognition (OCR). This also allowed us to leverage extensive prior work on text classification. This study was approved by the Committee for the Protection of Human Subjects (the UTHSC-H IRB) under protocol HSC-SBMI-13-0549.

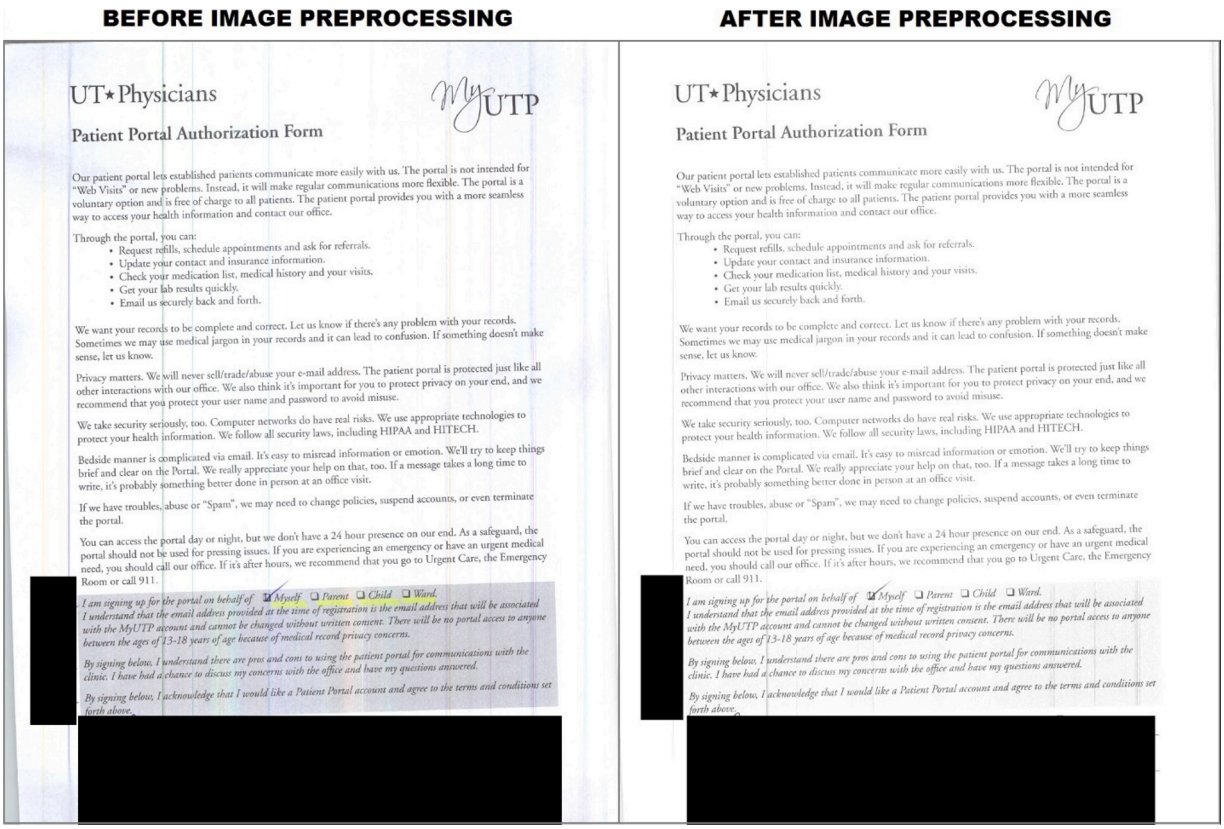

Fig. 1. An example of a document before and after image preprocessing steps were applied. (Note: black regions indicate removal of identifiable information.).

3.1. Image preprocessing

In order to improve OCR performance, we transformed each image using the Open Source Computer Vision library (OpenCV) prior to text extraction [18]. Each image was loaded and transformed into grayscale, a technique that has been shown to increase OCR performance with Tesseract [19]. The majority of scanned documents were already grayscale, though some scans of identification or insurance cards contained color. Scan quality was variable. Next, we increased contrast by 20 %. Finally, we ran one iteration of the 'erosion' transformation, which slightly increases the thickness of each character in order to improve recognition. Fig. 1 shows an example document before and after transformations.

3.2. Optical character recognition (OCR)

We used Tesseract OCR [20], an open source OCR software package via a python wrapper library (pytesseract). Each page of a document was processed separately. Example output of OCR is shown in Fig. 2.

3.3. Classification models

Since OCR was processed at the page level, to perform document-level classification, text from all pages of a single scanned document record were concatenated into a single text document. This text was labelled with the document type of the source document. The extracted text was then used to create features for traditional classification models as well as to evaluate different embeddings with deep learning models.

For the traditional machine learning models, we evaluated the following feature sets:

1 Count vectors, which represent the frequency of words in each class.
2 Term Frequency – Inverse Document Frequency (TF-IDF) created at the word level. (TF-IDF Word)
3 TF-IDF created at the bi- and tri-gram word level. (TF-IDF 2/3-gram Word)
4 TF-IDF created at the bi- and tri-gram character level. (TF-IDF 2/3-gram Char)

These vectors were created using the Scikit-learn TfidfVectorizer and CountVectorizer functions [21]. Each of these were then evaluated by creating different supervised classification models:

1 Multinomial Naïve Bayes (results presented in supplemental materials tables 2–4)
2 Logistic Regression
3 Random Forest

The training and test sets were made up of a 75 percent and 25 percent split of the document level data respectively. All models were created using Scikit-learn. The multinomial Naïve Bayes model used default parameters. Logistic Regression was implemented using a SAGA solver, and one-vs-rest (OvR) for multi-classification. Random Forest was implemented using default parameters except the number of trees was increased from 10 (default) to 250.

For the deep learning model, we used a pre-trained version of BERT, called ClinicalBERT [22]. ClinicalBERT is an extension of BERT-Base, but further pre-trained on MIMIC-III [23] notes to better adapt the language model to EHR text. BERT-Base utilizes 12 transformer encoder layers with 12 attention heads, for a total of 110 million parameters. As is standard with BERT-based text classification, a single linear layer with a cross-entropy loss is used on top of the ClinicalBERT model for the classification tasks described in this paper. The deep learning model was created using the TensorFlow, PyTorch, and Transformers python libraries [24–26]. The same training and test sets as the traditional

```
UT* Physicians
UTP
Patient Portal Authorization Form

Our patient portal lets established patients communicate more easily with us. The portal is not intended for
"Web Visits" or new problems. Instead, it will make regular communications more flexible. The portal is a
voluntary option and is free of charge to all patients. The patient portal provides you with a more seamless
way to access your health information and contact our office.

Through the portal, you can:
* Request refills, schedule appointments and ask for referrals.
* Update your contact and insurance information.
* Check your medication list, medical history and your visits.
* Get your lab results quickly.
* Email us securely back and forth.

We want your records to be complete and correct, Let us know if there's any problem with your records.
Sometimes we may use medical jargon in your records and it can lead to confusion. If something doesn't make
sense, let us know.

Privacy matters. We will never sell/trade/abuse your e-mail address. The patient portal is protected just like all
other interactions with our office. We also think it's important for you to protect privacy on your end, and we
recommend that you protect your user name and password to avoid misuse.

We take security seriously, too. Computer networks do have real risks. We use appropriate technologies to
protect your health information. We follow all security laws, including HIPAA and HITECH.

Bedside manner ts complicated via email. It's easy to misread information or emotion. We'll try to keep things
brief and clear on the Portal. We really appreciate your help on that, too. Ifa message takes a long time to
write, it's probably something better done in person at an office visit.

If we have troubles, abuse or "Spam", we may need to change policies, suspend accounts, or even terminate
the portal.

You can access the portal day or night, but we don't have a 24 hour presence on our end. As a safeguard, the
portal should not be used for pressing issues. If you are experiencing an emergency or have an urgent medical
need, you should call our office. If it's after hours, we recommend that you go to Urgent Care, the Emergency
Room or call 911.

I am signing up for the portal on behalf of Wiyself Q Parent Child OO Ward.

L understand that the email address provided at the time of registration ts the email address that will be associated
with the MyUTP account and cannot be changed without written consent. There will be no portal access to anyone
between the ages of 13-18 years of age because of medical record privacy concerns.

| . * . . .
By signing below, I understand there are pros and cons to using the patient portal for communications with the

os + . .
clinic. | have had a chance to discuss my concerns with the office and have my questions answered.
:

By signing below, I acknowledge that I would like a Patient Portal account and agree to the terms and conditions set
forth above,

Signature Date Relationship
Printed Name Email Address
```

Fig. 2. Sample results of OCR after processing the image from Fig. 1.

classifiers were used. However, 10 % of the training set was used to generate a validation set used during model training for early stopping. Additionally, this model was trained with a learning rate of 2e-5 and a batch size of 6, for 10 epochs.

All models were evaluated using recall (number of true positives divided by the number of true positives and false negatives), precision (number of true positives divided by the number of true positives and false positives), F1 (harmonic mean of recall and precision) and accuracy (number of correctly predicted out of all records). Also, for each class we report the classes that were most frequently incorrectly identified for that class. This error analysis helped us determine which classes should be grouped together in the next step.

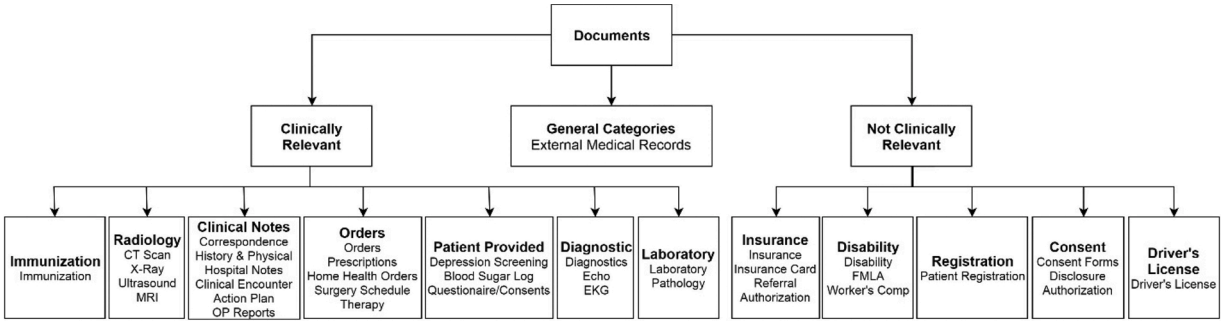

Fig. 3. Hierarchy of documents showing binary classification and sub-classification.

3.4. Grouping and re-evaluation

We labelled each document type as either clinically-relevant or not clinically-relevant and evaluated the performance of all models. Specifically, Consent Forms, Disability, Disclosure Authorizations, Driver's License, FMLA, Insurance, Insurance Card, Patient Registration, and Referral Authorizations were classified as not clinically relevant, while the remaining classes were all classified as clinically relevant (Fig. 3).

We then grouped each document type into 12 classes. These 12 classes were: Immunization, Radiology, Clinical Notes, Orders, Patient Provided, Diagnostic, Laboratory, Insurance, Disability, Registration, Consent, and Driver's License. This created a 3-level hierarchy (Fig. 3). After this grouping, all models were once again evaluated using these new classifications. Since most documents were composed of multiple pages, we compared the performance of training and testing at the page level (where each page within a document is treated as its own document) as well as at the document level (where text from all pages is concatenated together and treated as one document). This results in a total of 6 models (3 levels of classification x 2 levels of document and page) per machine learning model.

3.5. Text pre-processing

We evaluated the change in performance from the previous step, after applying two different text pre-processing methods to the text after the OCR step. The first text pre-processing method (a) consisted of a three step process.

A Text Pre-Processing Method 1 (PP1)
B Commonly misrepresented letters as numbers during OCR were replaced if at least one of the neighboring characters was a letter, and the other neighboring character is either a letter or a white space. These replacements were '5' to 'S', '0' to 'O', '2' to 'Z', '4' to 'A', and '8' to 'B'.
C All remaining numbers and punctuation characters were removed from the text.
D Stop words as defined in the Python Natural Language Toolkit (NLTK) library [27] were removed from the text.
E Text Pre-Processing Method 2 (PP2)

Since we knew that some spelling errors may be introduced via the OCR process, we ran a spell checker (SymSpell) using the default provided English word frequency dictionary over each word in every document and replaced the word with the first suggestion in the event that a spelling mistake was identified [28].

The text pre-processing steps and traditional machine learning models were run on an Intel Xeon CPU E5-2603 v2 @1.80 GHz. The BERT models were run on an NVIDIA TESLA V100 32GB GPU.

3.6. Statistical significance comparisons

We conducted two sets of statistical significance tests. First, we compared different models. Specifically, we evaluated the statistical significance of the difference between the best performance without text pre-processing (labelled Baseline in Table 2) of Random Forest with Count Vectors vs. Logistic Regression with TF-IDF Word vs. ClinicalBERT using McNemar's test, implemented via the python statsmodels library [29,30]. If significant, these comparisons are represented by *, ** and *** under "Baseline" in Table 2.

Second, we compared the effects of text-preprocessing. Specifically, we compared the baseline models within each of the best performing models, with the two different models that included text pre-processing for that model (PP1 and PP2), using the same statistical testing procedure. If significant, these comparisons are represented by #, ##, ### under "PP1" (Baseline vs. PP1) and/or "PP2" (Baseline vs. PP2) in Table 2.

Table 1
Top 41 document frequencies in 2018 by type, with descriptive statistics on page counts (prior to combining any document types).

Document Type	Document Count	Max # Pages	Mean # Pages	Median # Pages	Std dev # Pages
Action Plan	8824	28	1.5	1	1.5
Blood Sugar Log	14,413	107	4.3	2	5.9
Clinical Encounter	307,647	260	1.9	1	1.9
Consent Forms	375,153	72	4.6	4	3.8
Correspondence	115,335	207	3.5	2	4.7
CT Scan	9593	34	2.3	2	1.5
Depression Screening	52,048	21	1.3	1	0.6
Diagnostics	13,696	67	2.1	1	2.7
Disability	11,280	103	4.6	3	4.6
Disclosure Authoriza	81,168	81	2.2	1	2.2
Disclosure Authorizations	27,584	321	2	2	2.7
Drivers License	172,085	20	1.9	2	0.5
Drivers Liscense	81,018	26	1.4	1	0.5
Echo	15,139	47	2.8	3	1.3
EKG	31,634	110	1.5	1	1.4
External Medical Rec	43,038	2565	11.1	6	29.3
External Medical Records	61,144	800	11.8	6	20.4
FMLA	9583	58	5.7	5	4.4
H&P for Surgeries	21,923	54	4.6	2	5.4
History Physical	25,254	116	2.2	2	2
Home Health Orders	34,989	99	5.2	4	4.6
Hospital Notes	26,386	306	6.8	4	10.4
Immunizations	41,586	168	1.7	1	1.4
Insurance	105,725	104	2.3	1	2.9
Insurance Card	315,849	34	2	2	0.8
Laboratory	78,895	232	2.8	2	3.3
MRI	27,131	49	2.3	2	1.4
OP Reports	31,298	58	3	3	1.9
Orders	256,744	145	2.4	1	3.7
Pathology	19,226	94	2	2	1.9
Patient Registration	244,917	36	1.5	1	1
Prescriptions	156,930	100	2.2	1	2.8
Questionaire/ Consent	86,030	61	2.6	2	2.6
Questionnaire/ Consents	114,747	60	2.8	2	2.3
Referrals/ Authorizat	31,711	172	3.6	2	4.9
Referrals/ Authorzations	108,385	218	5	3	5.7
Surgery Schedule	42,480	102	7.4	5	7.6
Therapy	98,037	79	3.6	3	3.2
Ultrasound	29,302	43	2.2	2	1.7
Worker's Comp	16,102	264	3	1	6.1
X-Ray	18,523	91	1.9	1	1.9

4. Results

For all document types listed in Table 1 the minimum number of pages in each document was one page. The EHR contained a "catch all" category called "External Medical Records." Based on informal manual review, this category contained a wide variety of content that would otherwise fit into other existing document types. We identified two instances of documents incorrectly classified out of the 340 documents that were reviewed. One X-ray was misclassified as a hospital note, and one insurance document was misclassified as a driver's license. Similarly there were instances of documents that were correctly classified but may have been more appropriately placed into a different category. For example, an insurance card was identified as "Insurance," a category which reflects documentation pertaining to insurance, as opposed to the "Insurance Card" category which contained only insurance cards. We did not identify any errors where a clinically relevant document was in a

Table 2
Accuracy results for the best-performing models at three different levels of classification, at both the page and document level. This shows the baseline results, and results after applying different text pre-processing (PP) methods to the text, prior to classification.

	Document Level (Accuracy)								
	Logistic Regression - TF-IDF Word			Random Forest - Word Count			ClinicalBERT		
	Baseline	PP1	PP2	Baseline	PP1	PP2	Baseline	PP1	PP2
All Classes	0.877	0.880	0.866$^{\#\#}$	0.883**	0.884	0.880	0.908***	0.913	0.897$^{\#\#}$
12 Classes	0.916	0.919	0.908$^{\#\#}$	0.918	0.922	0.918	0.942***	0.949$^{\#\#}$	0.938
Binary Classes	0.957	0.958	0.955	0.954	0.959$^{\#}$	0.958	0.967***	0.973$^{\#\#\#}$	0.971
	Page Level (Accuracy)								
	Logistic Regression - TF-IDF Word			Random Forest - Word Count			ClinicalBERT		
	Baseline	PP1	PP2	Baseline	PP1	PP2	Baseline	PP1	PP2
All Classes	0.781	0.780	0.763$^{\#\#\#}$	0.779	0.784	0.778	0.845***	0.839$^{\#}$	0.827$^{\#\#\#}$
12 Classes	0.829	0.823$^{\#\#}$	0.808$^{\#\#\#}$	0.827	0.833$^{\#\#}$	0.826	0.885***	0.882	0.872$^{\#\#\#}$
Binary Classes	0.910	0.908	0.905$^{\#\#}$	0.905***	0.910	0.909	0.936***	0.935	0.931$^{\#\#\#}$

Statistical significance between models is indicated as *, **, and *** for comparisons between baseline models (Random Forest Baseline vs. Logistic Regression Baseline vs. ClinicalBERT Baseline) and $^{\#}$, $^{\#\#}$, and $^{\#\#\#}$ for comparisons of text preprocessing methods (Baseline vs. PP1 vs. PP2), indicating a p-value $<= 0.05$, 0.01, and 0.001 respectively.

Table 3
Performance of the ClinicalBERT model (best performing model) for all document types. The last two columns indicate the classes most frequently misidentified for the document type and the percentage of occurrence. The most frequent misclassification is in bold font when it is not in the same 12-class category as the correct document type.

Document Type	Precision	Recall	F1	n	Most Frequent Misclassification	% of Most Frequent Misclassification
Action Plan	0.908	0.892	0.900	111	Correspondence	0.78 %
Blood Sugar Log	0.928	0.932	0.930	234	**Clinical Encounter**	0.80 %
CT Scan	0.948	0.936	0.942	156	**Correspondence**	0.78 %
Clinical Encounter	0.906	0.906	0.906	500	Action Plan	4.59 %
Consent Forms	0.906	0.926	0.916	500	Disclosure Authorizations	2.99 %
Correspondence	0.761	0.825	0.792	475	**Disability**	8.29 %
Depression Screening	0.985	0.981	0.983	468	**Clinical Encounter**	0.80 %
Diagnostics	0.913	0.851	0.881	222	**Correspondence**	1.94 %
Disability	0.663	0.663	0.663	193	FMLA	24.00 %
Disclosure Authorizations	0.943	0.924	0.934	991	Consent Forms	7.44 %
Driver's License	0.958	0.986	0.972	998	**Blood Sugar Log**	1.70 %
EKG	0.955	0.971	0.963	485	Echo	0.63 %
Echo	0.984	0.981	0.983	318	Diagnostics	0.48 %
FMLA	0.686	0.69	0.688	174	Disability	18.65 %
History Physical	0.983	0.972	0.978	727	Correspondence	1.17 %
Home Health Orders	0.956	0.96	0.958	500	**Correspondence**	1.75 %
Hospital Notes	0.947	0.922	0.934	500	**Home Health Orders**	1.59 %
Immunizations	0.968	0.926	0.947	652	**QuestionaireConsents**	3.10 %
Insurance	0.852	0.810	0.830	741	ReferralAuthorizations	6.10 %
Insurance Card	0.912	0.952	0.932	750	Insurance	3.41 %
Laboratory	0.91	0.923	0.917	691	**Orders**	2.52 %
MRI	0.91	0.915	0.912	410	X-Ray	2.71 %
OP Reports	0.971	0.942	0.956	499	**Ultrasound**	1.46 %
Orders	0.813	0.776	0.794	500	Therapy	4.39 %
Pathology	0.94	0.91	0.925	277	Laboratory	1.85 %
Patient Registration	0.957	0.94	0.949	501	**Insurance**	2.13 %
Prescriptions	0.802	0.821	0.812	459	Orders	3.77 %
Questionnaire/Consents	0.93	0.96	0.945	750	**Immunizations**	1.28 %
Referral Authorizations	0.802	0.829	0.815	714	Insurance	4.83 %
Surgery Schedule	0.934	0.904	0.919	500	**Insurance**	1.42 %
Therapy	0.861	0.906	0.883	498	**Worker's Comp**	3.45 %
Ultrasound	0.877	0.887	0.882	476	MRI	2.67 %
Worker's Comp	0.897	0.875	0.886	208	**ReferralAuthorizations**	1.08 %
X-Ray	0.902	0.839	0.869	317	Ultrasound	3.12 %
Macro Average	0.899	0.895	0.897	16,495	–	–

Table 4
Results of best model, ClinicalBERT, after converting classes to either clinically relevant or not clinically relevant, and applying text pre-processing method 1.

	Precision	Recall	F1	n
Clinically Relevant	0.975	0.984	0.980	10,725
Not Clinically Relevant	0.970	0.954	0.962	5770
Macro Average	0.972	0.969	0.971	16,495
Accuracy	0.973			

non-clinically relevant classification.

Appendix Table 1 provides a short description of each document type, the number of document pages, and the average number of tokens per page.

The OCR required over one week of computer time running on a multi-core machine running in parallel. Training the ClinicalBERT model took over 30 h.

Table 2 shows the primary classification results. This includes only the best-performing feature representation for Random Forest and

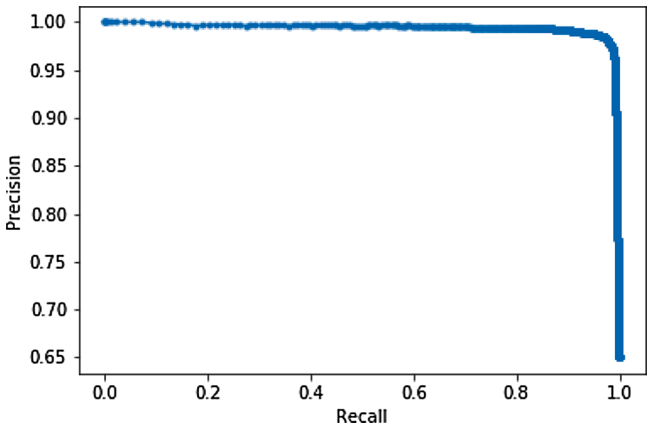

Fig. 4. Precision-Recall Curve for best performing model, ClinicalBERT with text pre-processing method 1 applied to documents at the document level, on binary clinically relevant classification.

Logistic Regression. These results are for classification at the document level (top) as well as the individual page level (bottom) separately. As shown in Table 2, the ClinicalBERT model performed best in every category. However it is worth pointing out how well the Random Forest using only simple count vectors as input features, as well as the Logistic Regression model using TF-IDF word features were able to perform similarly for the classification task at the document level. The difference in performance between the traditional and ClinicalBERT models was larger at the page level. The performance of all models, including naïve Bayes, is provided in Appendix Table 2 (page level, before text processing), Appendix Table 3 (page level, text-pre-processing) and Appendix Table 4 (document level, after text pre-processing). Table 3 provides a breakdown of the best performing model (ClinicalBERT) by each document type along with the most common error document class, for the baseline evaluation at the document level for all classes (see also Appendix Table 5).

A review of the most frequent error classes for each document type (Table 3), confirmed some expected relationships between document types. For example, the content of administrative "FMLA" and "Disability" documents were very similar and can be reasonably grouped into a unified category. In addition, "Consent Forms" and "Disclosure Authorizations," and separately imaging reports like "X-Ray," "MRI," "Ultrasound," and "CT Scan" were also very similar.

After grouping like document classes into the 12 defined classes as shown in Fig. 3, the overall performance of all models increased. The accuracy of the ClinicalBERT model increased to 0.942, the Random Forest model with count vectors to 0.918, and the Logistic Regression model with TF-IDF vectors to 0.916 at the document level, and 0.885, 0.827, and 0.829 at the page level respectively. Results for all models evaluated at this level are additionally shown in Appendix Table 2. Finally, the best performance was achieved when evaluating at the binary classification level, deciding between if a document was clinically relevant or not. The accuracy of the ClinicalBERT model increased to 0.967, the Random Forest model with count vectors to 0.954, and the Logistic Regression model with TF-IDF vectors to 0.957 at the document level, and 0.936, 0.910, and 0.905 at the page level respectively.

The results from both text pre-processing methods applied to the text after OCR for these select top performing models are also shown in Table 2. The PP1 (Text Pre-Processing Method 1) column corresponds to the results after the first post processing method was applied, likewise PP2 (Text Pre-Processing Method 2) corresponds to the second post processing method. Results for all models after text pre-processing method 1 are in Appendix Table 3. Additionally, results for all models after text pre-processing method 2 are in Appendix Table 4. Most models saw a minor improvement when applying text pre-processing method 1. Conversely, most models saw a decrease in performance when applying text pre-processing method 2.

Fig. 4 shows the precision-recall curve for the ClinicalBERT model on the binary classification task of separating clinically-relevant documents from non-clinically relevant documents at the document level, after text pre-processing method 1 had been applied. This is the overall best performing model. Table 4 also shows a breakdown of performance for each class for this model.

5. Discussion

We were able to accurately classify scanned documents into categories using a combination of optical character recognition (OCR) and machine learning. A ClinicalBERT model trained on already-classified scanned documents can distinguish clinically relevant from not clinically relevant documents with an accuracy of 0.973. Notably, even a relatively simple Logistic Regression model using TF-IDF vectors as input features was able to accurately identify clinically relevant documents. There is a tradeoff between the number of classes and the overall accuracy of the system. When classifying into 12 classes, grouping similar document types together, the best performing model had an accuracy of 0.949. At the lowest level of classification, with one document type per class, the best model performed with an overall accuracy of 0.913.

As shown in Table 1, the average number of pages in the "External Medical Records" category at our institution was much higher than any other document type. In order for a physician to view these documents, they must click through page by page within the EHR. Given realistic time constraints, it would be difficult for clinicians to review each scanned page. This causes a clinical problem, because that a physician is responsible for information within all parts of a patient's EHR. Given that the "External Medical Records" category contained largely the same document types as the already-classified documents, our method trained on the already-classified documents may reduce clinician workload by automatically identifying clinically relevant documents. To put it another way, the system can screen out documents that are not clinically relevant.

There are multiple ways that our approach could be implemented in practice. At 100 % recall, precision is approximately 65 % (Fig. 4). Thus, a conservative implementation strategy would be to configure the

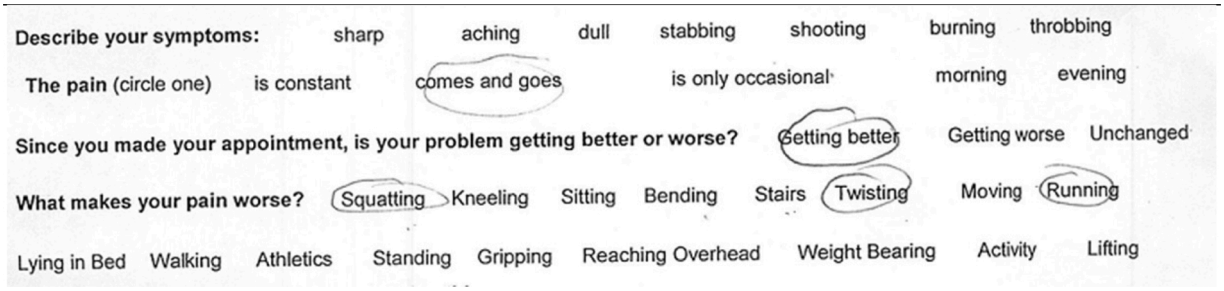

Fig. 5. Example of scanned document where pen marks interfere with text.

system to reduce the burden on clinicians by moving some non-clinically-relevant documents into a general "administrative document" category with minimal risk of missing a clinically-relevant document. An alternative strategy is to use the system to categorize "External Medical Records" into more meaningful categories. However, these documents likely have a different class distribution than the 40 document types classified here. Further, documents from an external source will likely look different. Both these issues may lower classification accuracy.

To accurately model the practical task of categorizing the content of the "External Medical Record" document type, we experimented with classifying each page separately. This is important as the external documents often contain multiple documents in one file (e.g., a scan of all the patient's records in one scanned file, with many types of documents). However, this still leads to issues as it can be difficult to separate documents of the same type (e.g., two X-Rays reports in a row) as well as issues with the final page of a document, which is often much shorter than a full page and thus more difficult to classify. Many scanned documents included some type of pen marking (example in Fig. 5), making it difficult for the OCR process to recognize words. Further, a variety of content such as fax cover sheets were sometimes included with the scanned documents.

OCR occasionally makes errors, creating misspellings within a token (e.g., "Physician" → "Physlclan"). Tokens that do not exist in the pre-trained word embedding vocabulary—referred to as out-of-vocabulary (OOV)—cannot be associated with a vector. Word embeddings are therefore susceptible to spelling errors due to the fact that the misspelled words may not exist within pre-trained word embeddings. Traditional ML models have always had difficulty with character n-grams, whereas for deep learning, character-based models have become more prominent [31].

One recent development is to create word embeddings that are resilient to misspelt words called MOE (Misspelling Oblivious Embeddings) [32], while another approach would be to use a character-based model such as fastText [33]. While BERT is able to support some degree of OOV words and misspellings by breaking unknown words into subgroups, the overall performance of a system with spelling mistakes and OOV is expected to not perform as well as one without. This is why we attempted to correct spelling mistakes using a basic spell checking library. However, this caused the performance of almost all systems to decrease. This may be partially due to the spell checker not being specific to clinical documents, often treating medical acronyms, medications, or terminologies as misspellings. Better performance may be obtained from a medical spell checker [34–36].

In addition to the OCR-related spelling errors, there were also interesting tokens created from the OCR algorithm trying to read logos, handwritten text, lines, or tables. For example, the OCR algorithm tries to read lines one at a time. In the case of tables with extra padding around the words, the OCR may read a line of the table as a series of 'I' characters (e.g., "IIIIIIIIIIII"), or a line as a series of 'm' characters. As was the case with misspellings, these format tokens often occurred consistently in a specific document type. Format tokens contributed weight in the count vector models, but were OOV for the word embeddings. An alternative explanation for the relatively high performance of the simple count vector and TF-IDF models is that a primary linguistic feature needed to classify documents in many cases is a fairly small set of key phrases (e.g., "Patient Portal Authorization Form" in Fig. 1). These are easily captured by discrete count vectors.

It is worth considering the amount of processing power required to train and use these models if deciding to implement them into a clinical system. While not systematically measured and evaluated as part of this study, the Random Forest and Logistic Regression models required much less time and system resources to train compared to the ClinicalBERT models. Some efforts have been made to try and reduce the overall system resources needed for BERT models, such as the creation of a version of BERT called DistilBERT which uses fewer parameters, while maintaining 95 % of BERT's performance [37]. Another computational consideration is OCR processing time, which in our case took a week of computational time to process a single month's worth of data.

5.1. Study limitations

Our study has several limitations. First, we evaluated scanned documents from a single EHR at one institution and our results may not generalize to other institutions. However, the EHR contained data from over 100 individual receiving clinics representing multiple clinical specialties and the scanned documents originated from an unknown (but probably large) number of sending institutions, increasing the likelihood that our findings are generalizable.

Second, our models depend on the use of already-classified external documents for training and testing. However, manual review suggests that these documents are not always consistently classified. Thus, there may be some errors in both the training and test sets. Since the existing classification system depends entirely on clinic staff to manually select the appropriate document type it was expected that some errors would occur. This error is difficult to quantify without a significant manual review effort. While our manual review of 10 documents from each class, 340 documents in total, identified two errors, we recognize that the subset reviewed is too small for determining a true error rate.

Third, we limited our study to the most frequent document types. Future work includes determining how best to handle the "long tail" of infrequent document types (89 types, 6.2 % of total documents). However, the distinction between high-level categories (e.g., clinically-relevant vs. not) may be the more clinically-relevant.

Finally, in some cases our methods did not allows to distinguish the specific causes why performance. Specifically, multiple post-processing approaches were "bundled" into PP1 and PP2. Thus, we could not identify the contribution of each individual component of PP1 and PP2 to overall performance. The optimal post-processing (e.g., isolated benefit of stop words) can be evaluated in future work.

5.2. Study significance

The clinical risk posed by scanned documents is difficult to quantify. For example, we do not know how often clinically-important information is missed when it is present only in scanned documents. We also do not know whether categorizing scanned documents into relatively precise and clinically meaningful categories, as opposed to general categories such as "External Medical Records" will decrease the risk of missing important information.

In future work, we will determine how best to integrate this system into a clinical workflow and whether this approach can reduce risk in practice. For example, it is not clear what false negative rate for identifying clinically-relevant documents is acceptable to clinicians and institutional administrators. An additional area to explore relates to document segmentation. The "external medical records" category often contained multiple documents, sometimes without a clear division between documents. Thus, we will explore the utility of a document segmentation step.

6. Conclusion

A wide variety of scanned documents are commonly included in EHRs. A combination of OCR and machine learning can accurately classify documents into clinically meaningful categories.

Funding statement

This study was supported in part by NCATS Grants UL1 TR003167, U01 TR002393, NLM grant R01 LM011829, NIBIB grant R21 EB029575, PCORI grant ME-2018C1-10963, the Cancer Prevention Research Institute of Texas (CPRIT) Precision Oncology Decision Support Core

Summary points

Already known:

Scanned documents are common in electronic health records (EHRs) with many systems containing millions of scanned pages from a variety of sources, representing many different document types.

Scanned documents may pose a patient safety hazard because they can contain clinically relevant information, but are not searchable. Thus, clinically relevant information may be missed.

What this study added:

We implemented a pipeline for scanned document classification using optical character recognition (OCR) and machine learning.

Scanned documents can be accurately classified into two (clinically relevant vs. not clinically relevant) or more categories representing different document types.

RP150535, CPRIT Data Science and Informatics Core for Cancer Research (RP170668), and the Reynolds and Reynolds Professorship in Clinical Informatics, National institute of Biomedical imaging and Bioengineering (NIBIB: R21EB029575), the Patient-Centered Outcomes Research Institute (PCORI: ME-2018C1-10963).

Contributorship statement

HG, KR and EVB conceived the methods. HG implemented and tested the software used to collect data and perform the analyses. HG, KR and EVB drafted the original version of the manuscript. All authors read and agreed with the analysis and the manuscript.

Declaration of Competing Interest

The authors have no competing interests to declare.

Appendix B. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.ijmedinf.2020.104302.

References

[1] S. Felt-Lisk, L. Johnson, C. Fleming, et al., Toward understanding EHR use in small physician practices, Health Care Financ. Rev. 31 (2009) 11–22.
[2] A. Friedman, J.C. Crosson, J. Howard, et al., A typology of electronic health record workarounds in small-to-medium size primary care practices, J. Am. Med. Inform. Assoc. 21 (2014) e78–83, https://doi.org/10.1136/amiajnl-2013-001686.
[3] Es Patterson, S. Anders, S. Moffatt-Bruce, Clustering and prioritizing patient safety issues during EHR implementation and upgrades in hospital settings, in: Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care Published Online First, 15 May, 2017, https://doi.org/10.1177/2327857917061028.
[4] N. Chen, D. Blostein, A survey of document image classification: problem statement, classifier architecture and performance evaluation, IJDAR 10 (2007) 1–16, https://doi.org/10.1007/s10032-006-0020-2.
[5] D. Ribli, A. Horváth, Z. Unger, et al., Detecting and classifying lesions in mammograms with Deep Learning, Sci. Rep. 8 (2018) 1–7, https://doi.org/10.1038/s41598-018-22437-z.
[6] A. Esteva, B. Kuprel, R.A. Novoa, et al., Dermatologist-level classification of skin cancer with deep neural networks, Nature 542 (2017) 115–118, https://doi.org/10.1038/nature21056.
[7] J. Du, J. Xu, H. Song, et al., Optimization on machine learning based approaches for sentiment analysis on HPV vaccines related tweets, J. Biomed. Semantics 8 (2017) 9, https://doi.org/10.1186/s13326-017-0120-6.
[8] T. Mikolov, Gs Corrado, K. Chen, et al., Efficient Estimation of Word Representations in Vector Space, 2013, pp. 1–12.
[9] J. Pennington, R. Socher, C. Manning, Glove: global vectors for word representation, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: :Association for Computational Linguistics (2014) 1532–1543, https://doi.org/10.3115/v1/D14-1162.
[10] Y. Shao, S.J. Taylor, N.J. Marshall, et al., Clinical text classification with word embedding features vs. bag-of-words features, IEEE International Conference on Big Data (Big Data) 2018 (2018) 2874–2878, https://doi.org/10.1109/BigData.2018.8622345.
[11] J. Devlin, M.-W. Chang, K. Lee, et al., BERT: pre-training of deep bidirectional transformers for language understanding, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: : Association for Computational Linguistics (2019) 4171–4186, https://doi.org/10.18653/v1/N19-1423.
[12] Y. Si, J. Wang, H. Xu, et al., Enhancing clinical concept extraction with contextual embeddings, J. Am. Med. Inform. Assoc. 26 (2019) 1297–1304, https://doi.org/10.1093/jamia/ocz096.
[13] H. Rhodes, M. Dougherty, American health information management association. Practice brief. Document imaging as a bridge to the EHR, J. AHIMA 74 (56) (2003). A-56G.
[14] E. Liette, C. Meyers, K. Olenik, Is Document Imaging the Right Choice for Your Organization? J. AHIMA 79 (2008) 58–60.
[15] R. Mittal, A. Garg, Text extraction using OCR: a systematic review, 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA) (2020) 357–362, https://doi.org/10.1109/ICIRCA48905.2020.9183326.
[16] D. Kim, D. Seo, S. Cho, et al., Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec, Inf. Sci. 477 (2019) 15–29, https://doi.org/10.1016/j.ins.2018.10.006.
[17] S. Dumais, Using SVMs for text categorization, IEEE Intell. Syst. 13 (4) (1988) 21–23.
[18] G. Bradski, The OpenCV library. Dr Dobb's Journal of Software Tools, 2000.
[19] C.I. Patel, A. Patel, D.T. Patel, Optical Character Recognition by Open source OCR Tool Tesseract: A Case Study, 2012, https://doi.org/10.5120/8794-2784.
[20] A. Kay, Tesseract: an open-source optical character recognition engine, Linux J. 2007 (2007) 2.
[21] F. Pedregosa, G. Varoquaux, A. Gramfort, et al., Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[22] E. Alsentzer, J.R. Murphy, W. Boag, et al., Publicly Available Clinical BERT Embeddings, arXiv:190403323 [cs] Published Online First: 20 June, 2019, http://arxiv.org/abs/1904.03323.
[23] A.E.W. Johnson, T.J. Pollard, L. Shen, et al., MIMIC-III, a freely accessible critical care database, Sci. Data 3 (2016) 160035, https://doi.org/10.1038/sdata.2016.35.
[24] Martín Abadi, Ashish Agarwal, Paul Barham, et al., TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, 2015. http://tensorflow.org/.
[25] Paszke A., Gross S., Massa F., et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. 12.
[26] Wolf T., Debut L., Sanh V., et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv:191003771 [cs] Published Online First: 11 February 2020.http://arxiv.org/abs/1910.03771 (accessed 29 May 2020).
[27] E. Loper, S. Bird, NLTK: the natural language toolkit. Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1, Association for Computational Linguistics, Stroudsburg, PA, USA, 2002, pp. 63–70, https://doi.org/10.3115/1118108.1118117.
[28] W. Garbe, wolfgarbe/SymSpell, 2020 (accessed 1 Jun 2020), https://github.com/wolfgarbe/SymSpell.
[29] Q. McNemar, Note on the sampling error of the difference between correlated proportions or percentages, Psychometrika 12 (1947) 153–157, https://doi.org/10.1007/BF02295996.
[30] S. Seabold, J. Perktold, Statsmodels: Econometric and Statistical Modeling with Python, Austin, Texas, 2010, pp. 92–96, https://doi.org/10.25080/Majora-92bf1922-011.
[31] G. Lample, M. Ballesteros, S. Subramanian, et al., Neural architectures for named entity recognition, Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, California: : Association for Computational Linguistics (2016) 260–270, https://doi.org/10.18653/v1/N16-1030.
[32] A. Piktus, N.B. Edizel, P. Bojanowski, et al., Misspelling oblivious word embeddings, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,

Volume 1 (Long and Short Papers). Minneapolis, Minnesota: : Association for Computational Linguistics (2019) 3226–3234, https://doi.org/10.18653/v1/N19-1326.

[33] P. Bojanowski, E. Grave, A. Joulin, et al., Enriching Word Vectors with Subword Information. arXiv:160704606 [cs], Published Online First: 19 June, 2017 (accessed 14 Dec 2019), http://arxiv.org/abs/1607.04606.

[34] J. Crowell, Q. Zeng, L. Ngo, et al., A frequency-based technique to improve the spelling suggestion rank in medical queries, J. Am. Med. Inform. Assoc. 11 (2004) 179–185, https://doi.org/10.1197/jamia.M1474.

[35] C.J. Lu, A.R. Aronson, S.E. Shooshan, et al., Spell checker for consumer language (CSpell), J. Am. Med. Inform. Assoc. 26 (2019) 211–218, https://doi.org/10.1093/jamia/ocy171.

[36] H. Kilicoglu, M. Fiszman, K. Roberts, et al., An ensemble method for spelling correction in consumer health questions, AMIA Annu. Symp. Proc. 2015 (2015) 727–736.

[37] V. Sanh, L. Debut, J. Chaumond, et al., DistilBERT, A Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter, arXiv:191001108 [cs] Published Online First: 29 February, 2020, http://arxiv.org/abs/1910.01108.