Journal of Biomedical Informatics 61 (2016) 141–148

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

PDF text classification to leverage information extraction from publication reports

Duy Duc An Bui [a,b,*], Guilherme Del Fiol [a], Siddhartha Jonnalagadda [b]

[a] Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, USA
[b] Department of Preventive Medicine-Health and Biomedical Informatics, Northwestern University, Chicago, IL, USA

ARTICLE INFO

Article history:
Received 23 September 2015
Revised 22 March 2016
Accepted 31 March 2016
Available online 1 April 2016

Keywords:
Text classification
Natural language processing
Document analysis
Machine learning

ABSTRACT

Objectives: Data extraction from original study reports is a time-consuming, error-prone process in systematic review development. Information extraction (IE) systems have the potential to assist humans in the extraction task, however majority of IE systems were not designed to work on Portable Document Format (PDF) document, an important and common extraction source for systematic review. In a PDF document, narrative content is often mixed with publication metadata or semi-structured text, which add challenges to the underlining natural language processing algorithm. Our goal is to categorize PDF texts for strategic use by IE systems.

Methods: We used an open-source tool to extract raw texts from a PDF document and developed a text classification algorithm that follows a multi-pass sieve framework to automatically classify PDF text snippets (for brevity, texts) into TITLE, ABSTRACT, BODYTEXT, SEMISTRUCTURE, and METADATA categories. To validate the algorithm, we developed a gold standard of PDF reports that were included in the development of previous systematic reviews by the Cochrane Collaboration. In a two-step procedure, we evaluated (1) classification performance, and compared it with machine learning classifier, and (2) the effects of the algorithm on an IE system that extracts clinical outcome mentions.

Results: The multi-pass sieve algorithm achieved an accuracy of 92.6%, which was 9.7% ($p < 0.001$) higher than the best performing machine learning classifier that used a logistic regression algorithm. *F*-measure improvements were observed in the classification of TITLE (+15.6%), ABSTRACT (+54.2%), BODYTEXT (+3.7%), SEMISTRUCTURE (+34%), and MEDADATA (+14.2%). In addition, use of the algorithm to filter semi-structured texts and publication metadata improved performance of the outcome extraction system (*F*-measure +4.1%, $p = 0.002$). It also reduced of number of sentences to be processed by 44.9% ($p < 0.001$), which corresponds to a processing time reduction of 50% ($p = 0.005$).

Conclusions: The rule-based multi-pass sieve framework can be used effectively in categorizing texts extracted from PDF documents. Text classification is an important prerequisite step to leverage information extraction from PDF documents.

1. Introduction

Systematic reviews (SRs) are important expert-synthesized information sources to enable evidence-based medicine practice [1]. However, the production and updating of SRs are often costly, slow, and unable to keep pace with the rapid growth of the biomedical literature [2,3]. The total expenditure of the Cochrane Collaboration, a prominent SR development organization, in fiscal year 2011 was $2.4 million [4], and that number increased to $3.9 million in 2013 [5]. Citations indexed in PubMed® have grown from 4 million (pre-1975) to 22 million today [3]. It takes 2.5–6.5 years for a primary study publication to be included and published in a new SR [6]. About 23% of SRs have not been updated with new evidence within 2 years after the first publication [6], and many clinical questions are not addressed in existing SRs [7]. As a result, SRs quickly become outdated and suboptimal for patient care. This is partially because the SR process involves many labor-intensive manual steps, which face human limitations such as limited time and resources, and human inconsistency and errors. This situation highlights the need for investigating computer techniques to aid humans in SR development.

The SR process involves a series of scientifically rigorous steps [8], such as citation searching [9], abstract screening, full-text screening, data extraction, and article appraisal. Data extraction

* Corresponding author at: University of Utah, Biomedical Informatics, 421 Wakara Way, Ste 140, Salt Lake City, UT 84108-3514, USA.
E-mail addresses: duy.bui@utah.edu, bdaduy@gmail.com (D.D.A. Bui).

http://dx.doi.org/10.1016/j.jbi.2016.03.026
1532-0464/

to generate evidence summaries is one of the most important and time-consuming steps in SR development [8]. Natural language processing (NLP) research in the past decade has investigated techniques to extract study characteristics from biomedical publications [10–19]. Those techniques have the potential to optimize the manual data extraction process; however, there are research gaps that have not been filled. One of the gaps that we choose to address in the present study concerns the heterogeneity of the digital document format. Present information extraction (IE) studies select sources of extraction from MEDLINE® abstracts, PubMed Central® (PMC) archives, and journal websites [12,17,19] available in HyperText Markup Language (HTML) format. However, the data extraction practice requires that the source of extraction be the original full-text study reports [8], and the most common format for full-text reports is the Portable Document Format (PDF).

NLP research on PDF documents faces several challenges. In a PDF document, narrative content is often mixed with publication metadata (header, footer, author information, journal information, etc.) and semi-structured text (tables or figures). Publication metadata are often not relevant to the extraction goal and can add noise to the NLP system. Semi-structured text can contain relevant information, but it does not adhere to grammatical rules and requires different extraction strategies than narrative text. Therefore, categorizing the text snippets (texts) in a PDF document is a necessary first step to design an optimal extraction strategy.

There have been studies on document structure recognition that sought to recover the logical structure from PDF documents. Commonly used approaches were machine learning [20–23] and rule-based or heuristics [24–29]. A rich number of PDF features have been used, including text pattern, format, spatial coordinates, and page boundary. Those methods used different approaches to recognize the PDF structure, and their performances also varied. None of the previous studies have been evaluated with practical real-world applications; therefore, the usefulness of PDF structure recognition for IE or text mining has not been validated.

In the present research, we present an alternative approach to recognizing PDF structure. We used the PDFBox tool [30] to extract raw texts and metadata from PDF files and applied a novel automated classification technique to categorize text into high-level document structures. PDFBox is a free open-source tool commonly used in NLP solutions for PDF documents [31–33]. In our preliminary study, PDFBox outperformed other open-access PDF-to-text conversion tools in extracting unlossy texts and maintaining the correct text order. The proposed algorithm employed the rule-based multi-pass sieve framework to perform the classification. For evaluation, we used the PDF reports used in the development of Cochrane systematic reviews. First, the classification algorithm performance was measured and compared against machine learning approaches. Then, the algorithm was applied to and evaluated in an IE use case.

2. Materials and methods

Our study has three main parts: (a) development of a gold standard for PDF text classification and outcome extraction task, (b) development of a multi-pass sieve algorithm for PDF text classification, and (c.1) evaluation of the performance of the multi-pass sieve algorithm and comparison with a machine learning approach and (c.2) evaluation of the impact of PDF text classification on IE performance. The system architecture and the study design are summarized in Fig. 1.

2.1. Gold standard

For being close to actual systematic review data extraction, we developed a gold standard from the published Cochrane systematic reviews. More specifically, Cochrane reviews on the subject "heart and circulation" that were published after October 2014 were retrieved from the Cochrane Library web interface. In each review, we located the included primary studies and searched for the PDF reports. To narrow the research focus to clinical trials and facilitate the IE task, we excluded nonrandomized control trials and studies that had been reported in multiple publications.

To build the text classification corpus, we used the PDFBox tool to extract raw texts from the PDF reports. Then, we used the GATE annotation tool [34] to annotate text snippets into five categories: TITLE, ABSTRACT, BODYTEXT, SEMISTRUCTURE, and METADATA. The METADATA labels were assigned to text snippets related to citation information, such as authorship, journal name, header/-footer, and references. The SEMISTRUCTURE labels were assigned to text snippets that consisted of tables or figures. The TITLE and ABSTRACT labels were assigned to snippets that were the title and abstract of the document, respectively. The remainder of the text snippets were assigned the BODYTEXT label.

In the next stage, we developed the gold standard for IE of study outcomes. Study outcomes are the extracted data elements commonly reported in the evidence summary of Cochrane reviews. They are the measurements used to assess a study hypothesis. We started with the outcome values reported in the Cochrane extraction template and extended by reviewing full-text manuscripts to validate and supplement the gold standard. The extraction template produced by systematic reviewers were not designed as an NLP gold standard, and might contain human extraction errors [35]. Therefore, a second validation with the original report is necessary to enhance the reliability of the system evaluation. We looked for all exact mentions of outcome concepts in the document. An outcome concept can be reported in multiple slightly different ways. We grouped all mentions co-referred to the same concept to a single unique entry to evaluate recall at the concept level.

2.2. Text snippet classification

Raw texts extracted from a PDF document using PDFBox have similar characteristics as manually copying and pasting texts from a PDF reader. Visual structures are lost, and texts are broken down into multiple lines of text snippet (Fig. 2). However, the essential text order is well maintained for NLP research. This work attempts to automatically assign document-level contexts to those text snippets, which is valuable information for designing NLP solutions for full-texts.

We investigated the rule-based multi-pass sieve approach and a set of machine learning approaches for automated text classification. To enhance the system accuracy, we utilized some external NLP and knowledge resources. First, we used standard features of PDFBox to extract raw texts, font type, font size, page number, and paragraph number. Second, we used the Stanford's Named Entity Recognition (NER) module [36] to detect PERSON, ORGANIZATION, and LOCATION entities in text. This feature was used in recognizing publication metadata such as authorship, affiliation, and bibliography. Lastly, MEDLINE® resources such as abstract, title, and author metadata were also used. Comparing MEDLINE® resources with full texts was a useful way to determine which text portions belong to the title, abstract, or author metadata. Those resources are shared and reused in both the multi-pass sieve and ML classification techniques.

2.3. The multi-pass sieve algorithm

We followed the multi-pass sieve framework to automatically classify text snippets into one of five categories. In previous studies, the multi-pass sieve framework has been successfully applied

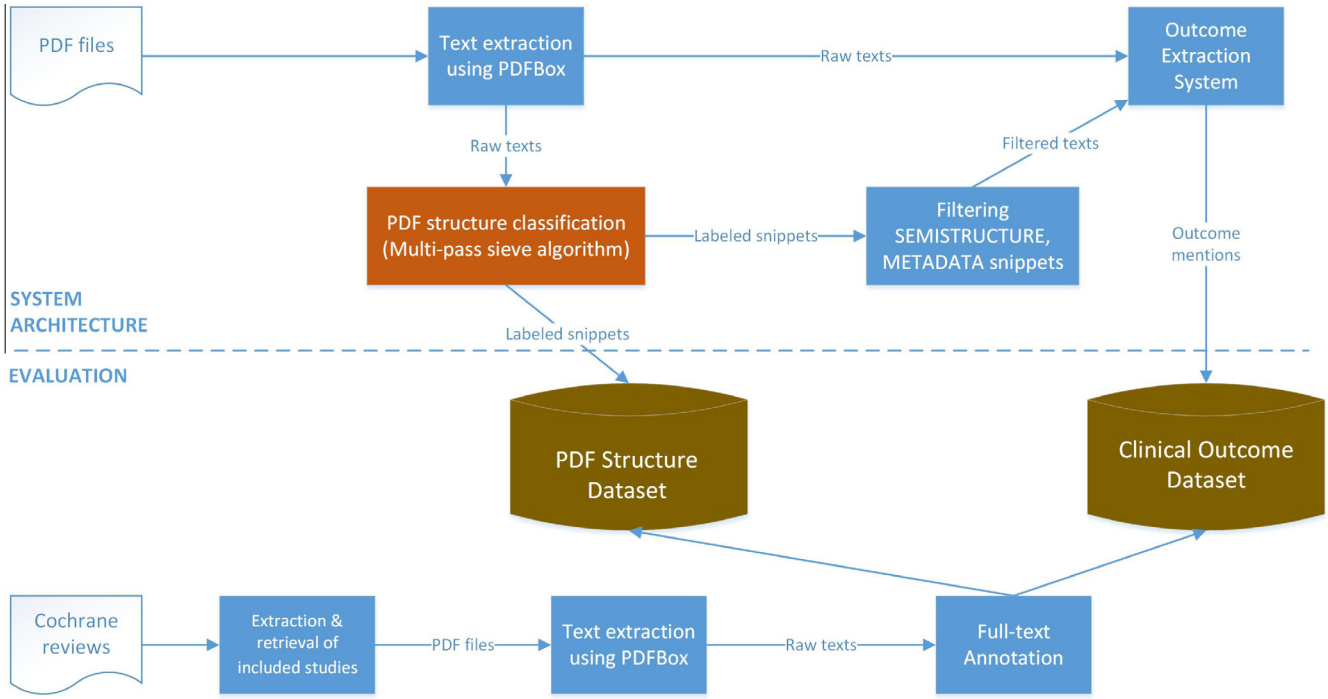

Fig. 1. The system architecture and the study design.

Title The Effect of Low Molecular Weight Heparin on
Title Survival in Patients With Advanced Malignancy
Metadata Clara P.W. Klerk, Susanne M. Smorenburg, Hans-Martin Otten, Anthonie W.A. Lensing,
Metadata Martin H. Prins, Franco Piovella, Paolo Prandoni, Monique M.E.M. Bos, Dick J. Richel,
Metadata Geertjan van Tienhoven, and Harry R. Bu¨ller
Abstract A B S T R A C T
Abstract Purpose
Abstract Studies in cancer patients with venous thromboembolism suggested that low molecular
Abstract weight heparin may prolong survival. In a double-blind study, we evaluated the effect of
Abstract low molecular weight heparin on survival in patients with advanced malignancy without
Abstract venous thromboembolism.
.....
Metadata J Clin Oncol 23:2130-2135. © 2005 by American Society of Clinical Oncology
Body Text INTRODUCTION
Body Text Tumor-mediated activation of the hemo-
Body Text static system has been implicated in both
Body Text the formation of tumor stroma and the
.....
Semi-Structure Table 1. Baseline Characteristics of the Patients Characteristic
Semi-Structure Nadroparin (n = 148) Placebo (n = 154)
Semi-Structure No. of Patients %
Semi-Structure Age, years
Semi-Structure Median 63 64

Fig. 2. An example of text classification output. The text portion highlighted in red is filtered out before passing to the information extraction system.

to solve co-reference resolution problem [37,38]. The framework favors applying multiple independent sieves passing through the document several times to perform the classification. The alternative machine learning approaches use a single-pass model, which performs the classification in a single pass through the document. An advantage of the multi-pass sieve model is the flexibility of breaking the complex task into multiple sub-tasks (or sieves). Each sieve is specialized to a specific task, which makes it convenient for testing, debugging and controlling for precision. To adapt the multi-pass sieve framework to text classification task, we implemented sieves, which included five main configurations: Begin condition, Pass condition, Stop condition, Directionality, and Repetition. The configurations are varied in different sieves and optimized to recognize a specific target label. +Begin Condition: If one of the composing rules is met, the sieve triggers the discovery process. Many of the begin conditions are dictionary matching rules such as looking up a prebuilt section heading collection. For instance, to identify author metadata, we look for snippets having patterns such as "correspondence to:," "author affiliations," and "financial disclosures".

+Directionality: This defines the direction to which the sieve moves to compute the next snippets in the document. Typically, the directionality for specific labels is statically configured to either UP or DOWN direction. For TABLE and FIGURE labels, the directionalities are dynamically configured. The sieve chooses direction dynamically based on an examination of the surrounding contexts.

+Pass Condition: The sieve assigns the target label to the snippet if the pass condition is met.

+Stop Condition: The sieve stops the discovery process if the condition is met. Stop condition prevents the sieve from aggressively expanding to other sections. A frequently used stop condition is the first failure of the Pass Condition, but there are other rules, such as matching common content headings and maximum page number limit.

+Repetition: This configuration defines the number of times the sieve is repeated. The sieve can be repeated one or many times. A difficult case involves the sieve recognition of ABSTRACT snippets. Abstract texts are sometimes divided into two clusters of texts. Therefore, we configured the sieve repetition of two times to capture those clusters.

Table 1 describes the full multi-pass sieve algorithm. The recognition of METADATA labels is subdivided into recognition of HEADER, FOOTER, KEYWORD, AUTHOR, JOURNAL, and REFERENCE. We built FIGURE and TABLE sieves to recognize SEMISTRUCTURE text snippets. Last, all unlabeled snippets were assigned to BODYTEXT labels.

2.4. Machine learning classifiers

We implemented and tested a representative set of machine learning algorithms to compare with the multi-pass sieve algorithm. Several machine learning algorithms that have been implemented and integrated in the data mining software WEKA were used in this study [39]. The Support Vector Machine algorithm, Sequential Minimal Optimization (SMO), with polynomial kernel and radial basis function (RBF) kernel were tested at various exponent, gamma values, and hyper-parameter c values. In addition, we tested Naïve Bayes, J48, and Logistic Regression algorithms with their default configurations.

We developed a machine learning model as summarized in Table 2. In addition to bag of words features that have been used in designing machine learning models as well as used as baseline approaches [40–43], we also developed additional features that take advantage of external resources (MEDLINE® metadata, PDFBox, and Stanford's NER) that were used in the multi-pass sieve algorithm.

The number of text snippets used for training is significantly large (16,546 snippets in 24 documents). Training snippets all at

Table 1
The full description of multi-pass sieve algorithm (R = Rule).

Target element	Begin condition	Pass condition	Stop condition	Direction/repetition
HEADER	**R1:** Begin of page AND **R2:** Repeat more than 1 times	**R3:** Same paragraph with previous line	**R4:** Fail Pass Condition **R5:** Match common section headings	Direction: DOWN Repeat: UNLIMITED
FOOTER	**R5:** End of page AND **R2**	**R3**	**R4** **R5**	Direction: UP Repeat: UNLIMITED
KEYWORD	**R6:** Match keywords headings	**R3**	**R4** **R5**	Direction: DOWN Repeat: 1
TABLE	**R7:** Match table common headings **R7.1:** Treat the following lines in the paragraph as the Table captions	**R3** **R8:** Same font with previous line AND **R9.1:** NOT contain document main font **R10:** Contain sequence of number pattern (e.g., Mean age 46 87) **R11** Contain mathematical and reporting symbols (±*<>†‡)	**R4** **R5** **R9.2:** Contain document main font AND **R12**: Contain predication/verb	Direction: DYNAMIC (e.g., choose the direction with the largest number of numeric patterns) Repeat: UNLIMITED
FIGURE	**R13:** Match figure common headings **R13.1:** Treat the following lines in paragraph as the Figure captions	**R3** **R8** AND **R9.1** **R11**	**R4** **R5** **R9.2** AND **R12**:	Direction: DYNAMIC Repeat: UNLIMITED
TITLE	**R13:** Begin of paragraph AND **R14:** Contain in MEDLINE's title	**R3** **R14**	**R4** **R5** **R15:** Page number > 2	Direction: DOWN Repeat: 1
ABSTRACT	**R16:** Match the Abstract heading. **R17:** Contain in MEDLINE's abstract	**R3** **R17**	**R4** **R5** **R6** **R15**	Direction: DOWN Repeat: 2
REFERENCE	**R18:** Match reference common headings. **R19:** Prefix by a number AND **R20:** Contain PERSON or ORGANIZATION entities.	**R3** **R8**	**R4**	Direction: DOWN Repeat: UNLIMITED
AUTHOR	**R20:** Contain in MEDLINE's authors **R21:** Match common Authors information headings (e.g., correspondence to:, author affiliations, financial disclosures, etc.)	**R3** **R8** AND **R9** **R22:** Contain LOCATION, PERSON, ORGANIZATION entities **R23:** Contain publication predications (submitted, published, supported, received, accepted, etc.)	**R4** **R5**	Direction: DOWN Repeat: UNLIMITED
JOURNAL	**R24:** Match Journal metadata headings (e.g., original article, print issn:, link available on, etc.)	**R3** **R23** **R25:** Match URL, IP address, price, DOI patterns	**R4** **R5**	Direction: DOWN Repeat: UNLIMITED

Table 2
Text snippet classification machine learning features.

Variable	Type	Description
text_length	Numeric	Length of text snippet in number of characters
paragraph_number	Numeric	Paragraph position number
page_number	Numeric	Page number
IsPrimaryFont	Boolean	Whether the text snippet uses the most prominent font in the document
font_size	Numeric	Font size of the text snippet
containInMedlineTitle	Boolean	Whether the snippet is contained in the MEDLINE® title
containInMedlineAbstract	Boolean	Whether the snippet is contained in the MEDLINE® abstract
containInMedlineAuthor	Boolean	Whether the snippet is contained in the MEDLINE® author metadata
numPERSON	Numeric	Number of PERSON entities
numLOCATION	Numeric	Number of LOCATION entities
numORGANIZATION	Numeric	Number of ORGANIZATION entities
Bag-of-words features	Numeric	Frequencies of each word

once become inefficient and difficult to scale. In our experiments, a computer with 16 GB of RAM ran out of memory when training SVMs with all the text snippets. Training all snippets was not feasible due to significant need for system memory. This issue raises a scalability limitation when training the ML model in larger and more diverse datasets. To address this problem, we treated each document as a single classifier, and used the majority voting approach proposed by John Wiley & Sons (2004) to combine the individual classifiers [44]. A snippet was assigned to a class if it received the majority of votes from multiple document classifiers. If there were ties, random assignment following a uniform random distribution was conducted.

2.5. Outcome extraction system

To measure the impact of the classification algorithm on an IE system, we used a homegrown PICO (Population, Intervention, Comparison, and Outcome) extraction system (citation pending). The goal of this system is to extract PICO data elements from full-text PDF reports to aid in SR development. The full description of the system is beyond the scope of this report; therefore, we present a brief description of one of the most mature components, the outcome extraction system. In short, the outcome extraction system is composed of two main stages: sentence selection and noun phrase chunking and filtering. The first stage accepts raw text input from any source and splits the input into multiple sentences by using Stanford NLP's sentence splitter. From those sentences, we selected only sentences that potentially contain outcome information (e.g., contain definitive phrases: "outcomes were," "study end points were") or contain reporting phrases (e.g., "statistically different," "was improved," "was measured"). In the second stage, we used Stanford's parser to generate a Penn tree and extract all noun phrase mentions. Since noun phrase extraction might detect exceedingly long phrases in complex sentences, we filtered phrases that have more than 10 words. Last, we applied a set of regular expression rules and semantic tests to collect outcome mentions. Regular expression rule looks for surrounding contexts [e.g., rate of (\\S+), incidence of (\\S+), etc.] to determine candidate mentions. In semantic test, we used Metamap [45] to map text snippets to UMLS concepts restricted to the following semantic types: "Finding," "Sign or Symptom," "Laboratory or Test Result," and "Disease and Syndrome."

2.6. Evaluation approach

We used the gold standard and methods described to test two hypotheses: H1: In the classification of PDF texts, the rule-based multi-pass sieve approach is more accurate than the machine learning approach; and H2: Using automated text classification to filter semi-structures and publication metadata improves performance of information extraction from full-text publications when compared to off-the-shelf PDF Box extraction.

To test H1, we randomly divided the gold standard into a 50–50 random split of documents. The first half of the dataset was used to train the ML classifier and to develop rules for the multi-pass sieve algorithm; the other half was used for the evaluation. Standard text classification evaluation metrics such as accuracy, recall, precision, and *F*-measure were calculated at the token level, with accuracy used as the primary endpoint. There are short and long text snippets. We evaluate at token level to better represent the amount of information irrespective of the length of text snippets. We used Wilcoxon signed-rank test to assess the significance of the differences between the two approaches in terms of the accuracies averaged across documents.

To test H2, we setup the experiment with two study arms. The first arm used raw texts extracted from PDF reports by using PDFBox. The second arm used the multi-pass sieve algorithm to categorize raw texts and filter all SEMISTRUCTURE and METADATA snippets before passing them to the IE system. Both arms were tested against the evaluation set of the gold standard described earlier. Recall, precision, *F*-measure, number of split sentences, and processing time are reported, with *F*-measure being the primary endpoint. We considered a correct mention if it contained phrases that appeared in the gold standard. Wilcoxon signed-rank test was used to test the significance of the performance difference per document between the two study arms.

3. Results

We constructed a gold standard composed of 48 published reports that were included in eight Cochrane reviews. Those reports represent the publication formats of 34 different journals. A follow-up analysis showed that only 64% of studies have contents available in HTML pages, while all of them can be downloaded as PDF reports. All of them are randomized controlled trials, but only 16% have posted structured results on ClinicalTrials.gov. Raw text extraction using PDFBox generated 33,307 lines of text snippets, from which we were able to annotate 157 (0.5%) TITLE snippets, 1230 (3.7%) ABSTRACT snippets, 17,711 (53.2%) BODYTEXT snippets, 5596 (16.8%) SEMISTRUCTURE snippets, and 8613 (25.9%) METADATA snippets. In the outcome extraction dataset, we were able to manually annotate 204 outcome mentions, with a rate of 4.2 outcome mentions per document.

Classification performances of various machine learning classifiers are summarized in Table 3. The best classifier used a logistic regression algorithm and achieved an accuracy of 82.9%. The multi-pass sieve approach achieved an average accuracy of 92.6% over 24 documents. This is a significant improvement of 9.7% ($p < 0.001$) over the logistic regression classifier. According to a power analysis, to reach a statistical power of 80% for the effect size we found, a sample of 16 documents would be needed.

Table 4 shows the detailed performance comparison of the logistic regression classifier and the multi-pass sieve algorithm. For certain data elements, the *F*-measures for the multi-pass sieve approach were much better than those for the machine learning classifier (TITLE +15.6%, ABSTRACT +54.2%; BODYTEXT +3.7%; SEMISTRUCTURE +34%; METADATA +14.2%).

For outcome extraction task, the IE system that operated on PDF texts after filtering out SEMISTRUCTURE and METADATA snippets had better performance than off-the-shelf PDF Box extraction (Table 5). The improvement on recall was not significant (+1%; $p = 0.11$) while precision was significantly improved

Table 3
Classification performances of various machine learning models.

	SVM polynomial kernel					SVM RBF kernel					Naïve Bayes	J48	Logistic regression
c	*e* = 1	*e* = 2	*e* = 3	*e* = 4	*e* = 5	*g* = 0.01	*g* = 0.02	*g* = 0.03	*g* = 0.04	*g* = 0.05	79.3	68.6	**82.9**
1	80.9	75.4	67.3	63.5	51.5	74.9	73.8	73.4	73.8	73.9			
2	80.9	75.5	65.3	63.3	50.7	72.6	74.5	75	75.3	75.8			
3	80.9	76.2	66.1	64.7	53.2	74.7	75.5	76.1	76.6	77.6			
4	81	76.4	67.1	65.3	54.1	75.2	75.9	76.9	78.1	79			
5	81	76.9	67.4	65.2	54.4	75.3	76.6	77.8	79.1	79.8			

The highest classification performance is marked in bold.

Table 4
Performance comparison of the multi-pass sieve approach versus the logistic regression classifier.

	Logistic regression classifier				Multi-pass sieve approach			
	Accuracy	Recall	Precision	*F*-measure	Accuracy	Recall	Precision	*F*-measure
TITLE (%)	82.9	89	62.3	69.7	92.6 ($p < 0.001$)	95.8	80.3	85.3
ABSTRACT (%)		21.3	57.5	28.9		86.6	82.2	83.1
BODYTEXT (%)		92	88	89.3		95.9	90.6	93
SEMISTRUCTURE (%)		50.4	70.9	54.9		88.6	91.5	88.9
METADATA (%)		81	72.9	73.1		82	96.8	87.3

(+4.3%; $p < 0.001$). *F*-measure increased significantly by 4.1% ($p = 0.002$). Most notably, filtering publication metadata and semi-structured texts reduced the number of sentences to be processed by 44.9% ($p < 0.001$), which corresponds to a processing time reduction of 50% ($p = 0.005$). In our testing, the most time consuming step of unfiltered branch was the syntactic parsing the lengthy non-grammar texts such as table.

4. Discussion

4.1. PDF text classification

We designed and evaluated a rule-based multi-pass sieve approach to categorize texts extracted from PDF documents. The approach is an alternative to the machine learning algorithms that are commonly used in text classification studies [43,46,47]. Overall, the multi-pass sieve classifier significantly outperformed the best machine learning classifier (accuracy 92.6% vs. 82.9%, $p < 0.001$). Machine learning approaches tend not to perform well on imbalanced datasets [48,49]. For this task of document-level classification, the distribution of text labels is not symmetrical (0.5% TITLE, 3.7% ABSTRACT, 53.2% BODYTEXT, 16.8% SEMISTRUCTURE, and 25.9% METADATA). This imbalanced dataset could lead to training models biased toward the majority labels. The multi-pass sieve framework addresses such problem through high precision sieves that target minority labels. To classify large labels, the framework allows breaking the task into multiple sieves with each sieve targeting a specific aspect of the data. That approach can control for precision, while recall can be improved by stacking multiple sieves. In the present study, the label-specific performances of the multi-pass sieve algorithm are significantly better, especially for minority labels (*F*-measure: TITLE +15.6%, ABSTRACT +54.2%, SEMISTRUCTURE +34%, and MEDADATA +14.2%).

Table 5
Comparison of IE performance operated on original PDF texts vs PDF texts after filtering SEMISTRUCTURE and METADATA snippets.

	Original texts extracted from PDF reports	PDF texts after filtering SEMISTRUCTURE and METADATA snippets
Recall (%)	95.8	96.8 ($p = 0.11$)
Precision (%)	43.1	47.4 ($p = 0.004$)
F Score (%)	68.5	62.6 ($p < 0.002$)
Average number of sentences	256	141 ($p < 0.001$)
Average processing time (minute)	1.16	0.58 ($p = 0.005$)

Our dataset also showed that 36% of studies published in PDF format did not have an HTML version available. This is further confirmed with Cochrane Heart Group's systematic reviewers, who stated that PDF is often the preferred choice due to wider adoption and availability offline. Therefore, IE systems need to operate on PDF documents to effectively support systematic review development. The annotation results showed that 26% of text snippets are publication metadata, and 17% are semi-structured texts. These findings confirm the heterogeneity problem in PDF reports. There are off-the-shelf tools developed to help detect the PDF logical structure. However, our preliminary studies could not find one that meets the needs of data extraction either because the classification schema did not match the needs of data extraction, or because tools did not perform well in our systematic review dataset.

The multi-pass sieve framework proposed in this study has several strengths. First, its accuracy was higher than the best performing machine learning classifier by 9.7%. Second, the framework is flexible and extendable. Developers have the flexibility to create and add new sieves and rules to target new data elements. Rules are organized at different stages to facilitate maintainability and extension. Third, the algorithm is intuitive, i.e. it operates in a way similar to human screening, in which documents are scanned for the prominent signatures (e.g., heading, caption) and then examine the contents.

4.2. Clinical outcome extraction

Our baseline system achieved 95.8% recall and 43.1% precision. While the recall is adequate, precision needs further improvement. The use of the classification algorithm to filter publication metadata and semi-structured texts improved recall by 1% and precision by 4.3%. The difference in recall was not significant, since the baseline recall was very high with little room for improvement. The precision improvement corresponds to a reduction of 17% in the number of false-positive mentions; therefore, the algorithm would considerably reduce the number of mentions that reviewers would need to correct in a semi-automated data extraction process.

A subsequent analysis showed that texts without filtering sometimes have publication metadata and semi-structured texts embedded within body text fragments, breaking up sentences. Detecting and filtering those non-prose texts improved the

performance of the sentence splitter algorithm (e.g., Stanford sentence splitter). Defining correct sentence boundaries is an important prerequisite step for most NLP systems. Incorrect sentence boundaries negatively impact subsequent NLP pipelines such as syntactic parsing and phrase chunking. Moreover, the filtering reduced the number of sentences to be processed by 44.9%. This reduction improved the efficiency of the NLP approach by reducing processing time (50%) without causing performance loss.

This study evaluated impacts of the PDF structure recognition on an IE system. However, the proposed technique is also potentially useful in other areas, such as information retrieval, automated document classification, and library management. These areas share the PDF heterogeneity problem that might degrade the performance of any text processing approaches.

4.3. Limitations

The algorithm takes advantage of MEDLINE® resources such as title, abstract, and author metadata, which reduces the applicability of the method to documents not indexed in MEDLINE®. However, our focus was on biomedical research publications and the majority of these resources are openly available in the MEDLINE® database.

This study did not test an exhaustive list of machine learning algorithms, their optimization parameters, and ensemble approaches to combine classifiers. Unexplored ML optimizations include using feature selection and sampling techniques to reduce dimensionality. Other ensemble approaches such as Bayes Optimal Classifier [50] and Stacking [51], have not been investigated. The ML baseline could be improved with sequential ML models such as Hidden Markov Model [52] and Conditional Random Field [53]. Further parameter optimization of ML models might also boost the classification performance, but with the risk of overfitting the model over the training set. Nevertheless, given the abundance of ML models and optimization methods, finding the best model to a single problem requires a significant amount of experimental work and often lacks generalizability. Our rule-based approach using the multi-pass sieve framework can serve as alternative or a complement to a machine learning solution, especially for classifying minority labels.

This study used the outcome extraction module to perform an intrinsic evaluation. The impacts of filtering on other data elements, such as sample size and intervention are unknown. Our analysis confirmed that filtering affects the performance of the sentence-splitter, which impacts any extraction methods that rely on the assumption of a correct sentence boundary.

We did not use semi-structured texts as the source of extraction, although they might contain outcome mentions. Semi-structured texts require different extraction strategies that rely more on pattern-matching and dictionary-matching than on syntactic parsing and chunking. Because we deemed recall to be satisfactory, an additional source of extraction was considered unnecessary.

4.4. Future work

Areas that demand future research include testing the classification algorithm on a diverse set of PDF documents, validating the usefulness in other text mining research such as information retrieval, document classification, and IE of other data elements, and using semi-structured texts instead of excluding them.

5. Conclusion

We present an alternative approach for PDF structure recognition by using PDFBox to extract raw texts and a multi-pass sieve algorithm for classification. The multi-pass sieve algorithm achieved a higher accuracy than the more commonly used machine learning classification approach. The multi-pass sieve algorithm also improved the performance of an IE system compared to off-the-shelf PDF extraction. PDF structure recognition unlocks the door to conduct text mining research on PDF files, an important information source for biomedical research.

Conflicts of interest

The authors have no conflicts of interest to declare.

Acknowledgments

This work was made possible by funding from National Library of Medicine Grants (R00LM011389 and R01LM011416-01). We are also thankful to the guidance provided by the US Satellite of the Cochrane Heart Group led by Dr. Mark Huffman.

References

[1] D.L. Sackett, W.M. Rosenberg, J.A. Gray, R.B. Haynes, W.S. Richardson, Evidence based medicine: what it is and what it isn't, BMJ (Clinical research ed). 312 (7023) (1996) 71–72.
[2] M. Ware, M. Mabe, An overview of scientific and scholarly journal publishing, The STM Report (2009).
[3] Statistical Reports on MEDLINE®/PubMed® Baseline Data. Available from: <https://www.nlm.nih.gov/bsd/licensee/baselinestats.html>.
[4] M.E. Schaafsma, The Cochrane Collaboration Treasurer's Report, 2012.
[5] Limited CTC, Directors' Reports and Financial Statements, 2013.
[6] K.G. Shojania, M. Sampson, M.T. Ansari, J. Ji, S. Doucette, D. Moher, How quickly do systematic reviews go out of date? A survival analysis, Ann. Intern. Med. 147 (4) (2007) 224–233.
[7] P. Bragge, O. Clavisi, T. Turner, E. Tavender, A. Collie, R.L. Gruen, The global evidence mapping initiative: scoping research in broad topic areas, BMC Med. Res. Methodol. 11 (1) (2011) 92.
[8] J.P. Higgins, S. Green, Cochrane Handbook for Systematic Reviews of Interventions, Wiley Online Library, 2008.
[9] D.D. Bui, S. Jonnalagadda, G. Del Fiol, Automatically finding relevant citations for clinical guideline development, J. Biomed. Inform. (2015).
[10] R.L. Summerscales, Automatic Summarization of Clinical Abstracts for Evidence-based Medicine, Illinois Institute of Technology, 2013.
[11] K.-C. Huang, I.J. Chiang, F. Xiao, C.-C. Liao, C.C.-H. Liu, J.-M. Wong, PICO element detection in medical text without metadata: are first sentences enough?, J Biomed. Inform. 46 (5) (2013) 940–946.
[12] F. Boudin, J.-Y. Nie, J.C. Bartlett, R. Grad, P. Pluye, M. Dawes, Combining classifiers for robust PICO element detection, BMC Med. Inform. Decis. Mak. 10 (2010) 29.
[13] H. Zhu, Y. Ni, P. Cai, Z. Qiu, F. Cao, Automatic extracting of patient-related attributes: disease, age, gender and race, Stud. Health Technol. Inform. 180 (2012) 589–593.
[14] D.P.A. Corney, B.F. Buxton, W.B. Langdon, D.T. Jones, BioRAT: extracting biological information from full-length papers, Bioinformatics (Oxford, England) 20 (17) (2004) 3206–3213.
[15] Verspoor K, Mackinlay A, Cohn JD, Wall ME. Detection of protein catalytic sites in the biomedical literature. In: Pac Symp Biocomput., 2013, pp. 433–444.
[16] J. Hakenberg, R. Leaman, N.H. Vo, S. Jonnalagadda, R. Sullivan, C. Miller, et al., Efficient extraction of protein–protein interactions from full-text articles, IEEE/ACM Trans. Comput. Biol. Bioinform. 7 (3) (2010) 481–494.
[17] W. Hsu, W. Speier, R.K. Taira, Automated extraction of reported statistical analyses: towards a logical representation of clinical trial literature, in: AMIA Annual Symposium Proceedings/AMIA Symposium AMIA Symposium, 2012, pp. 350–359.
[18] B. de Bruijn, S. Carini, S. Kiritchenko, J. Martin, I. Sim, Automated information extraction of key trial design elements from clinical trial publications, in: AMIA Annual Symposium Proceedings/AMIA Symposium AMIA Symposium, 2008, pp. 141–145.
[19] S. Kiritchenko, B. de Bruijn, S. Carini, J. Martin, I. Sim, ExaCT: automatic extraction of clinical trial characteristics from journal publications, BMC Med. Inform. Decis. Mak. 10 (2010) 56.
[20] R. Kern, K. Jack, M. Hristakeva, M. Granitzer, TeamBeam meta-data extraction from scientific literature, D-Lib Magazine 18 (7) (2012) 1.
[21] M. Granitzer, M. Hristakeva, R. Knight, K. Jack, R. Kern (Eds.), A comparison of layout based bibliographic metadata extraction techniques, Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics, ACM, 2012.
[22] H. Han, C.L. Giles, E. Manavoglu, H. Zha, Z. Zhang, E. Fox (Eds.), Automatic document metadata extraction using support vector machines, Digital Libraries 2003 Proceedings 2003 Joint Conference on;, IEEE, 2003.

[23] M.T. Luong, T.D. Nguyen, M.Y. Kan, Logical structure recovery in scholarly articles with rich document features, Multimedia Storage Retrieval Innovat. Digital Lib. Syst. (2012) 270.
[24] A. Constantin, S. Pettifer, A. Voronkov (Eds.), PDFX fully-automated PDF-to-XML conversion of scientific literature, Proceedings of the 2013 ACM symposium on Document Engineering, ACM, 2013.
[25] F. Kboubi, A.H. Chabi, M.B. Ahmed (Eds.), Table recognition evaluation and combination methods, Document Analysis and Recognition, 2005 Proceedings Eighth International Conference on, IEEE, 2005.
[26] H. Chao, J. Fan, Layout and content extraction for pdf documents, in: Document Analysis Systems VI, Springer, 2004, pp. 213–224.
[27] S. Klampfl, K. Jack, R. Kern, A comparison of two unsupervised table recognition methods from digital scientific articles, D-Lib Mag. 20 (11) (2014) 7.
[28] S. Klampfl, R. Kern, An unsupervised machine learning approach to body text and table of contents extraction from digital scientific articles, in: Research and Advanced Technology for Digital Libraries, Springer, 2013, pp. 144–155.
[29] E. Oro, M. Ruffolo (Eds.), PDF-TREX: An approach for recognizing and extracting tables from PDF documents, Document Analysis and Recognition, 2009 ICDAR'09 10th International Conference on, IEEE, 2009.
[30] Apache PDFBox – A Java PDF Library 2015. Available from: <https://pdfbox.apache.org/>.
[31] J. Beel, B. Gipp, A. Shaker, N. Friedrich, SciPlore Xtract: extracting titles from scientific PDF documents by analyzing style information (Font Size), in: Research and Advanced Technology for Digital Libraries, Springer, 2010, pp. 413–416.
[32] M. Garcia-Remesal, V. Maojo, J. Crespo, A knowledge engineering approach to recognizing and extracting sequences of nucleic acids from scientific literature, Conf. Proc. IEEE Eng. Med. Biol. Soc. 2010 (2010) 1081–1084.
[33] U. Schäfer, B. Kiefer, Advances in deep parsing of scholarly paper content, Springer, 2011.
[34] T. Kenter, D. Maynard, Using Gate as an Annotation Tool, University of Sheffield, Natural Language Processing Group, 2005.
[35] A.P. Jones, T. Remmington, P.R. Williamson, D. Ashby, R.L. Smyth, High prevalence but low impact of data extraction and reporting errors were found in Cochrane systematic reviews, J. Clin. Epidemiol. 58 (7) (2005) 741–742.
[36] J.R. Finkel, T. Grenager, C. Manning (Eds.), Incorporating non-local information into information extraction systems by gibbs sampling, Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 2005.
[37] K. Raghunathan, H. Lee, S. Rangarajan, N. Chambers, M. Surdeanu, D. Jurafsky, et al., Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2010.
[38] S.R. Jonnalagadda, D. Li, S. Sohn, S.T. Wu, K. Wagholikar, M. Torii, et al., Coreference analysis in clinical notes: a multi-pass sieve with alternate anaphora resolution modules, J. Am. Med. Inform. Assoc. 19 (5) (2012) 867–874.
[39] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA data mining software: an update, SIGKDD Explor. Newsl. 11 (1) (2009) 10–18.
[40] M.A.M. García, R.P. Rodríguez, L.E.A. Rifón, Biomedical literature classification using encyclopedic knowledge: a Wikipedia-based bag-of-concepts approach, PeerJ. 3 (2015) e1279.
[41] C. Soguero-Ruiz, K. Hindberg, J. Rojo-Alvarez, S.O. Skrovseth, F. Godtliebsen, K. Mortensen, et al., Support Vector Feature Selection for Early Detection of Anastomosis Leakage from Bag-of-Words in Electronic Health Records, 2014.
[42] R. Xu, Y. Hirano, R. Tachibana, S. Kido, Classification of diffuse lung disease patterns on high-resolution computed tomography by a bag of words approach, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2011, Springer, 2011, pp. 183–190.
[43] D.D. Bui, Q. Zeng-Treitler, Learning regular expressions for clinical text classification, J. Am. Med. Inform. Assoc. 21 (5) (2014) 850–857.
[44] L.I. Kuncheva, Combining Pattern Classifiers: Methods and Algorithms, John Wiley & Sons, 2004.
[45] A.R. Aronson, Metamap: Mapping Text to the Umls Metathesaurus, NLM, NIH, DHHS., Bethesda, MD, 2006.
[46] M.H. Song, S.H. Kim, D.K. Park, Y.H. Lee, A multi-classifier based guideline sentence classification system, Healthcare Inform. Res. 17 (4) (2011) 224–231.
[47] S. Sohn, M. Torii, D. Li, K. Wagholikar, S. Wu, H. Liu, A hybrid approach to sentiment sentence classification in suicide notes, Biomed. Inform. Insights 5 (Suppl. 1) (2012) 43–50.
[48] N. Japkowicz, S. Stephen, The class imbalance problem: a systematic study, Intell. Data Anal. 6 (5) (2002) 429–449.
[49] J. Van Hulse, T.M. Khoshgoftaar, A. Napolitano (Eds.), Experimental perspectives on learning from imbalanced data, Proceedings of the 24th International Conference on Machine Learning, ACM, 2007.
[50] S. Tong, D. Koller (Eds.), Restricted Bayes Optimal Classifiers, AAAI/IAAI, 2000.
[51] D.H. Wolpert, Stacked generalization, Neural Networks 5 (2) (1992) 241–259.
[52] S.R. Eddy, Hidden markov models, Curr. Opin. Struct. Biol. 6 (3) (1996) 361–365.
[53] D. Pinto, A. McCallum, X. Wei, W.B. Croft (Eds.), Table extraction using conditional random fields, Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2003.