YJBIN 2503
29 January 2016
ARTICLE IN PRESS
No. of Pages 8, Model 5G

Journal of Biomedical Informatics xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Toward better public health reporting using existing off the shelf approaches: A comparison of alternative cancer detection approaches using plaintext medical data and non-dictionary based feature selection

Suranga N. Kasthurirathne [a,*], Brian E. Dixon [b,c], Judy Gichoya [d], Huiping Xu [c], Yuni Xia [d], Burke Mamlin [b,d], Shaun J. Grannis [b,d]

[a] Indiana University School of Informatics and Computing, Indianapolis, IN, USA
[b] Regenstrief Institute, Indianapolis, IN, USA
[c] Indiana University Fairbanks School of Public Health, Indianapolis, IN, USA
[d] Indiana University School of Medicine, Indianapolis, IN, USA

A R T I C L E I N F O

Article history:
Received 30 April 2015
Revised 18 January 2016
Accepted 20 January 2016
Available online xxxx

Keywords:
Public health reporting
Decision models
Cancer
Pathology
Feature selection
Data preprocessing

A B S T R A C T

Objectives: Increased adoption of electronic health records has resulted in increased availability of free text clinical data for secondary use. A variety of approaches to obtain actionable information from unstructured free text data exist. These approaches are resource intensive, inherently complex and rely on structured clinical data and dictionary-based approaches. We sought to evaluate the potential to obtain actionable information from free text pathology reports using routinely available tools and approaches that do not depend on dictionary-based approaches.
Materials and methods: We obtained pathology reports from a large health information exchange and evaluated the capacity to detect cancer cases from these reports using 3 non-dictionary feature selection approaches, 4 feature subset sizes, and 5 clinical decision models: simple logistic regression, naïve bayes, k-nearest neighbor, random forest, and J48 decision tree. The performance of each decision model was evaluated using sensitivity, specificity, accuracy, positive predictive value, and area under the receiver operating characteristics (ROC) curve.
Results: Decision models parameterized using automated, informed, and manual feature selection approaches yielded similar results. Furthermore, non-dictionary classification approaches identified cancer cases present in free text reports with evaluation measures approaching and exceeding 80–90% for most metrics.
Conclusion: Our methods are feasible and practical approaches for extracting substantial information value from free text medical data, and the results suggest that these methods can perform on par, if not better, than existing dictionary-based approaches. Given that public health agencies are often under-resourced and lack the technical capacity for more complex methodologies, these results represent potentially significant value to the public health field.

1. Introduction

The widespread adoption of electronic medical records has resulted in increased availability of free text clinical data, usually in the form of plaintext reports dictated or typed by clinicians, for secondary use. Because free text clinical data must be converted to actionable information to realize its full value, analyzing and extracting pertinent information from unstructured clinical text has become an increasingly important activity within the healthcare industry.

Various approaches for obtaining actionable information from unstructured free text generally attempt to address the challenges of both identifying and contextualizing concepts of interest, so-called "named entities". Identifying named entities, a process termed "named entity recognition" (NER), can be performed using either dictionary-based or non-dictionary approaches. Dictionary-based approaches for NER rely on medical ontologies while non-dictionary approaches derive named entities from less formal sources such as clinician's empirical knowledge or from source data being analyzed.

* Corresponding author at: Indiana University School of Informatics and Computing, 535 W. Michigan Street, IT 475, Indianapolis, IN 46202, USA. Tel.: +1 (317) 278 4636.
E-mail address: snkasthu@iupui.edu (S.N. Kasthurirathne).

http://dx.doi.org/10.1016/j.jbi.2016.01.008
1532-0464/© 2016 Published by Elsevier Inc.

While dictionary-based approaches have the advantage of using lists of pre-vetted entities that reflect concepts of interest, no single medical ontology has been designed to comprehensively reflect entities for a specific illness/condition, nor grouped in a hierarchical structure that makes their selection an efficient process. Consequently, deriving concepts from existing ontologies to accurately identify specific conditions requires considerable expertise and manual effort.

Achieving accurate NER in plaintext reports is a significant bottleneck in text mining, especially when using dictionary based approaches [12]. Dictionary based NER performance measures have been found to be well below levels acceptable for routine use in clinical and research contexts [11,18]. Further, given that controlled vocabularies and ontologies routinely evolve (Bodenreider, 2008); Vreeman [20], dictionary based approaches for NER often require manual curation to accurately reflect constantly changing terminology. These challenges suggest that dictionary-based approaches require high maintenance, and may not yield a satisfactory cost benefit when applied in the medical domain. Conversely, performing NER using non-dictionary machine learning approaches (Jiang et. al., 2011) can mitigate the challenges of dictionary-based methods by leveraging data on hand to minimize the reliance on complex and constantly changing sources of external knowledge.

Although new approaches for processing unstructured clinical data are routinely published, there remains a paucity of practical, generalizable, evidence-based best practices addressing approaches for obtaining actionable information from unstructured clinical text. Further, much of the work performed in this space has been conducted in the clinical informatics realm, and there is shortage of methodology studies specifically addressing needs in the public health realm.

Consequently, this study seeks to assess the practical use of existing "off the shelf" text analysis and information-mining methods to generate actionable information from free text clinical resources to address problems affecting the population/public health space. As a demonstration of our work, we sought to assess how these approaches could improve case reporting to cancer registries using unstructured clinical data.

Cancer registries play a significant supporting role in public health activities by integrating cancer case information for multiple purposes including determining population-based cancer incidence, initiating survival and mortality reporting, identifying at-risk populations, and supporting research studies on comparability, clustering, and the adequacy of cancer surveillance [2,23]. However, cancer reporting activities are often delayed and incomplete [1,6,23], yielding delayed ascertainment of cases, which limits the value of cancer registry data and its use [19]. Prior studies have demonstrated that automated methods for identifying a variety of public health reportable cases can effectively improve the timeliness and completeness of case reporting [8,15].

The purpose of this study was to evaluate the accuracy of cancer case identification within plaintext clinical reports using off the shelf tools and machine learning NER approaches. By evaluating alternate approaches that vary the level of clinician expertise required, we sought to assess the performance of various automated cancer case detection approaches having varying levels of human guidance and pave the way for further research into practical applications for the public health space.

2. Materials and methods

We sought to evaluate our work using data collected by the Indiana Network for Patient Care (INPC), a large Health Information Exchange (HIE) serving major hospitals of Indiana [14]. The INPC serves public health by scrutinizing incoming HL7 laboratory messages for results of public health interest using dictionary-based approaches, and reports them to the state and county health departments [16]. However, it has no mechanism to perform similar reporting using plaintext data. We sought to assess non-dictionary cancer detection using plaintext pathology reports collected by the INPC. Pathology reports were used due to (a) their completeness and availability and (b) their suitability for identifying cancer diagnoses.

We sampled 7000 heterogeneous plaintext pathology reports distributed across seven diverse health systems, representing over 30 hospitals within the INPC. Clinicians performed a manual review of these reports and tagged them as either positive or negative for the presence of a cancer diagnosis. Next, we sought to identify specific tokens associated with the presence or absence of a cancer diagnosis using these labeled results.

2.1. Preparation of the master feature vector

A Perl script was written to parse each plaintext report and identify the number of unique tokens present in the entire report set. Of these, tokens that appear only once or twice in the entire set of reports were removed due to their low prevalence. We also identified and removed all stop words appearing in the token list using the Perl Lingua Stopwords module [5]. Next, we used the Negex algorithm [3] to identify the context of use (positive or negative) for each remaining token. The remaining tokens were stemmed using the Perl Lingua Stem module [4]. We counted the presence of each token in positive and negated contexts and used this data to prepare an input vector for each pathology report. Each token was represented by two digits in the master feature vector – the number of positive occurrences and the number of negative occurrences of each token per report. Subsets of the master feature vector would be used for decision modeling based on token subsets selected by each feature selection approach.

2.2. Selection of feature subsets

We used 3 non-dictionary feature selection approaches: (a) manual, (b) informed, and (c) automated to create feature subsets from the master feature vector.

2.2.1. Manual feature selection

Clinicians selected feature subsets based on their domain expertise. Two experienced clinicians independently created prioritized lists of tokens that would suggest the presence of a cancer diagnosis in a pathology report. The clinicians then compared their ranked lists and resolved any conflicts. In the event of a disagreement, a third clinician served as a tiebreaker. Using this process, we identified 20 top tokens for automated cancer case detection.

2.2.2. Informed feature selection

In contrast to the manual feature selection approach, the informed approach provided clinicians with summary statistics for each token. Combining this information with their own domain expertise, two clinicians independently reviewed and selected subsets of prioritized tokens for analysis. A third clinician adjudicated any disagreements. The summary statistics supplied to clinicians to aid in feature selection included:

$$\text{Positive Coverage} = P_X/R_P \tag{1}$$

$$\text{Negative Coverage} = N_X/R_N \tag{2}$$

$$\text{Coverage Ratio} = (P_X/R_P)/(N_X/R_N) \tag{3}$$

$$\text{Combined Term Frequency} = (O_X/R_{ALL}) \tag{4}$$

$$\text{Inverse Document Frequency} = \log(R_{ALL}/R_X) \tag{5}$$

Please cite this article in press as: S.N. Kasthurirathne et al., Toward better public health reporting using existing off the shelf approaches: A comparison of alternative cancer detection approaches using plaintext medical data and non-dictionary based feature selection, J Biomed Inform (2016), http://dx.doi.org/10.1016/j.jbi.2016.01.008