International Journal of Medical Informatics 105 (2017) 110–120

Contents lists available at ScienceDirect

International Journal of Medical Informatics

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

A comparison of rule-based and machine learning approaches for classifying patient portal messages

Robert M. Cronin[a,b,c,*], Daniel Fabbri[a,d], Joshua C. Denny[a,b], S. Trent Rosenbloom[a,b,c], Gretchen Purcell Jackson[a,c,e]

[a] Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA
[b] Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA
[c] Department of Pediatrics, Vanderbilt University Medical Center, Nashville, TN, USA
[d] Department of Computer Science, Vanderbilt University, Nashville, TN, USA
[e] Department of Pediatric Surgery, Vanderbilt University Medical Center, Nashville, TN, USA

ARTICLE INFO

Keywords:
Patient portal
Text classification
Natural language processing
Machine learning

ABSTRACT

Objective: Secure messaging through patient portals is an increasingly popular way that consumers interact with healthcare providers. The increasing burden of secure messaging can affect clinic staffing and workflows. Manual management of portal messages is costly and time consuming. Automated classification of portal messages could potentially expedite message triage and delivery of care.

Materials and methods: We developed automated patient portal message classifiers with rule-based and machine learning techniques using bag of words and natural language processing (NLP) approaches. To evaluate classifier performance, we used a gold standard of 3253 portal messages manually categorized using a taxonomy of communication types (i.e., main categories of informational, medical, logistical, social, and other communications, and subcategories including prescriptions, appointments, problems, tests, follow-up, contact information, and acknowledgement). We evaluated our classifiers' accuracies in identifying individual communication types within portal messages with area under the receiver-operator curve (AUC). Portal messages often contain more than one type of communication. To predict all communication types within single messages, we used the Jaccard Index. We extracted the variables of importance for the random forest classifiers.

Results: The best performing approaches to classification for the major communication types were: logistic regression for medical communications (AUC: 0.899); basic (rule-based) for informational communications (AUC: 0.842); and random forests for social communications and logistical communications (AUCs: 0.875 and 0.925, respectively). The best performing classification approach of classifiers for individual communication subtypes was random forests for Logistical-Contact Information (AUC: 0.963). The Jaccard Indices by approach were: basic classifier, Jaccard Index: 0.674; Naïve Bayes, Jaccard Index: 0.799; random forests, Jaccard Index: 0.859; and logistic regression, Jaccard Index: 0.861. For medical communications, the most predictive variables were NLP concepts (e.g., Temporal Concept, which maps to 'morning', 'evening' and Idea or Concept which maps to 'appointment' and 'refill'). For logistical communications, the most predictive variables contained similar numbers of NLP variables and words (e.g., Telephone mapping to 'phone', 'insurance'). For social and informational communications, the most predictive variables were words (e.g., social: 'thanks', 'much', informational: 'question', 'mean').

Conclusions: This study applies automated classification methods to the content of patient portal messages and evaluates the application of NLP techniques on consumer communications in patient portal messages. We demonstrated that random forest and logistic regression approaches accurately classified the content of portal messages, although the best approach to classification varied by communication type. Words were the most predictive variables for classification of most communication types, although NLP variables were most predictive for medical communication types. As adoption of patient portals increases, automated techniques could assist in understanding and managing growing volumes of messages. Further work is needed to improve classification performance to potentially support message triage and answering.

* Corresponding author at: Department of Biomedical Informatics, Vanderbilt University Medical Center, 2525 West End Blvd, Suite 1475, Nashville, TN, 37232, USA.
E-mail address: robert.cronin@vanderbilt.edu (R.M. Cronin).

http://dx.doi.org/10.1016/j.ijmedinf.2017.06.004
Received 15 February 2017; Received in revised form 13 June 2017; Accepted 20 June 2017

I. Informational Needs or Communications
- A. Normal Anatomy and Physiology
- B. Problems (Diseases or Observations)
  1. Definition
  2. Epidemiology
  3. Risk factors
  4. Etiology
  5. Pathogenesis/natural history
  6. Clinical presentation
  7. Differential diagnosis
  8. Related diagnoses
  9. Prognosis
- C. Management
  1. Goals/strategy
  2. Tests
  3. Interventions
  4. Sequence/timing
  5. Personnel/setting
- D. Tests
  1. Definition
  2. Goals
  3. Physiologic basis
  4. Efficacy
  5. Indications/contraindications
  6. Preparation
  7. Technique/administration
  8. Interpretation
  9. Post-test care
  10. Advantages/benefits
  11. Costs /disadvantages
  12. Adverse effects
- E. Interventions
  1. Definition
  2. Goals
  3. Mechanism of action
  4. Efficacy
  5. Indications/contraindications
  6. Preparation
  7. Technique/administration
  8. Monitoring
  9. Post-intervention care
  10. Advantages/benefits
  11. Costs/disadvantages
  12. Adverse effects

II. Medical Needs or Communications
- A. Appointments/scheduling
- B. Medical equipment
- C. Personnel/referrals
- D. Prescriptions
- E. Problems
- F. Follow-up
- G. Management
- H. Tests
- I. Interventions

III. Logistical Needs or Communications
- A. Contact information/communication
- B. Facility/policies
- C. Insurance/billing
- D. Medical records
- E. Personal documentation
- F. Health information technologies
- G. Tests
- H. Interventions
- I. Transportation

IV. Social Needs or Communications
- A. Acknowledgment
- B. Complaints
- C. Emotional need or expression
- D. Relationship communication
- E. Miscellaneous

V. Other

Fig. 1. The taxonomy of consumer health information communication types [17,33,34].

1. Introduction

Patient portals, online applications that allow patients to interact with their healthcare providers and institutions, have had increasing adoption because of consumer demand and governmental regulations [1]. Secure messaging is one of the most popular functions of patient portals, and this function allows individuals to interact with their healthcare providers [2–9]. The increasing burden of secure messaging has been demonstrated in multiple settings with providers having a few messages per week to multiple messages per day within a few years after patient portal implementation [10–13]. This increasing burden can affect staffing and workflows of a clinic and health care providers [14]. Being able to determine the contents of these messages in an automated fashion could potentially mitigate this burden.

Prior research has demonstrated that users express diverse health-related needs in portal messages, and substantial medical care is delivered through portal interactions [10,15–19]. Message content can contain informational (e.g., what is the side effect of simvastatin?), logistical (e.g., what time does the pharmacy open?), medical (e.g., I am having a new numbness in my legs), and social (e.g., Please thank your nurse for his care of my wife) communication types. Classification of messages may aid message management with triage to appropriate resources or personnel. Identifying when medical care is delivered in patient portal messages could support online compensation models beyond the limited codes for online care transitions and telehealth services [20].

Categorization of portal message content can be viewed as a text classification problem. The most popular methods employed for text classification include a manual approach, where a human will classify each message, or automatically, through rule-based approaches based on words or phrases that appear in the text or machine learning techniques (e.g., logistic regression, random forests, support vector machines) [21–27]. Text classification applications have been evaluated in the health care domain. Several studies have demonstrated the ability to classify unstructured text written by medical personnel [28,29]. These studies demonstrated excellent areas under the operator-receiver curve over 0.88. Researchers have also attempted to identify adverse drug reactions from consumer-generated text [30–32]. Although text classification has been successfully done for consumer-generated text from online forums and social media, this approach has not been applied to and evaluated for secure messages from patient portals.

A limited number of studies have classified portal messages in primary care settings only using manual methods [15,16]. North et al. manually classified 323 messages, demonstrating 37% were medication related, 23% were symptom related, 20% were test related, 7% were medical questions, 6% were acknowledgements, and 9% had more than one issue [16]. Haun et al. asked senders to classify their messages in predefined categories and observed the following distribution, although user-assigned categories were not consistently applied accurately: 59% general (i.e., condition management/report, specialty/procedure request, correspondence request, medication refill request,), 24% appointments (i.e., confirmations, cancellations, specialty appointment requests), and 16% refill and medication inquiries [15]. As patient portal and secure messaging adoption increases, understanding the content of these messages and their implications for provider workload becomes more important.

Our research team has previously evaluated different methods of automatically classifying messages sent through a patient portal used by multiple specialties at a large, tertiary care institution. We compared the performance of basic classification and machine learning approaches to determine the major types of portal message content using a gold standard of 1000 portal messages classified by the semantic types of communications within the message, including informational, medical, logistical and social communication types [17]. We discovered that automated methods have promise for predicting major semantic types of communications, but this work was limited because of the small data set, and it did not include an analysis of what features (e.g., words, concepts, semantic types) are most important for the machine learning classifiers or an attempt to classify messages beyond the major categories in a rich hierarchy containing many communication subtypes. In this manuscript, we expand on our previous work by comparing automated approaches to message classification using a substantially larger gold standard and the full semantic communication type hierarchy (Fig. 1), and determining the features important for classification.