Expert Systems with Applications 41 (2014) 1498–1508

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Automatic text classification to support systematic reviews in medicine

J.J. García Adeva [a,*], J.M. Pikatza Atxa [a], M. Ubeda Carrillo [b], E. Ansuategi Zengotitabengoa [b]

[a] Erabaki Group, Department of Computer Languages and Systems, University of the Basque Country UPV/EHU, Manuel Lardizabal 1, 20018 Donostia-San Sebastián, Spain
[b] Donostia University Hospital, 20014 Donostia-San Sebastián, Spain

ARTICLE INFO

Keywords:
Medical systematic reviews
Machine learning
Text mining
Text classification

ABSTRACT

Medical systematic reviews answer particular questions within a very specific domain of expertise by selecting and analysing the current pertinent literature. As part of this process, the phase of screening articles usually requires a long time and significant effort as it involves a group of domain experts evaluating thousands of articles in order to find the relevant instances. Our goal is to support this process through automatic tools. There is a recent trend of applying text classification methods to semi-automate the screening phase by providing decision support to the group of experts, hence helping reduce the required time and effort. In this work, we contribute to this line of work by performing a comprehensive set of text classification experiments on a corpus resulting from an actual systematic review in the area of Internet-Based Randomised Controlled Trials. These experiments involved applying multiple machine learning algorithms combined with several feature selection techniques to different parts of the articles (i.e., titles, abstract, or both). Results are generally positive in terms of overall precision and recall measurements, reaching values of up to 84%. It is also revealing in terms of how using only article titles provides virtually as good results as when adding article abstracts. Based on the positive results, it is clear that text classification can support the screening stage of medical systematic reviews . However, selecting the most appropriate machine learning algorithms, related methods, and text sections of articles is a neglected but important requirement because of its significant impact to the end results.

1. Introduction

Medical Systematic Reviews support the conversion of medical research into practice by bringing together the collection of existing studies that are relevant to a specific medical question. This synthesis of current evidence benefits different stakeholders such as clinicians and policymakers.

Although Systematic Reviews started as early as the 18th century (Lind, 1753), their production exploded after the second half of the 20th century along with a significant increment of publications in medical, nursing, and allied health care (Shonjania & Bero, 2001). Unfortunately, the significant growth of clinical trials in the last decades, has not been matched by a suitable number of systematic reviews produced (Bastian et al., 2010). An analysis of the situation at the time revealed that because the amount of work required to produce reviews is increasing, there was a majority of systematic reviews with many years out of date (Shojania et al., 2007).

The general process for creating a systematic review is based on three main steps: (i) conducting broad searches in the relevant literature, (ii) manually screenning titles and abstract of retrieved citations, and (iii) reviewing full articles of those citations identified as relevant. No matter how critical and necessary these steps are, they are very time consuming, especially the screening of citations and the review of candidate studies.

Multiple text mining techniques have been gaining popularity over the past years as a consequence of the ever increasing amount of available digital documents of unstructured text and, thus, the necessity of analysing their content in flexible ways (Hearst, 1999). From these techniques, one of the most prominent is text classification using machine learning, which consists of automatically predicting one or more suitable categories for unstructured texts written in natural language (e.g., English, Spanish, etc.). Text classification is currently a major research area with many commercial and research applications in a large number of domains. Medicine is one of the most evident areas where text mining methods have multiple applications, such as the discovery of new literature (Swanson, 1986), concept-based search (Ide, Loane, & Demner-Fushman, 2007), or automatic bibliographic update in clinical guidelines (Iruetaguena et al., 2013).

This work was motivated by the hypothesis that text classification could assist the production of Systematic Reviews by supporting reviewers in their process of manually screening published articles. Although this assumption is not new, as there has been recently an incipient while still modest body of research in this

[*] Corresponding author. Tel.: +34 943 018153.
E-mail address: jjga@ehu.es (J.J. García Adeva).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.08.047

direction (Thomas, McNaught, & Ananiadou, 2011), our contribution is focused on: (i) studying the application of a comprehensive selection of machine learning algorithms, (ii) combining these algorithms with multiple feature selection methods and different numbers of features, (iii) selecting different parts of citations (i.e., title, abstract, or both), and (iv) applying these methods to the medical domain of Internet-Based Randomised Controlled Trials.

In such a way, an automatic text classification system could be trained with a set of articles from the medical domain in question after the collection of studies had been already manually screened. As these articles describing primary studies had been manually labelled as either *relevant* or *irrelevant*, they fit well with the paradigm of a two-class text classifier. Once the system was trained, it was ready to automatically classify unseen articles, therefore providing input into the screening process similarly to a human expert. In consequence, this system would not aim at replacing the persons involved in the decision process but to complement and assist them. Contrary to other previous studies covered by Section 4, where they directly selected either the abstract of the full article to train and test the classifiers, we were interested in investigating what sections of the articles provided the best results. We also applied a bigger variety of classifiers than other previous studies, in addition to multiple feature selection methods.

This paper is organised as follows. Section 2 describes the methods used in this work to automatically classify articles. Section 3 describes the manual process for performing systematic reviews in medicine and how it can be supported by text classification. Previous efforts in this area of research are described in Section 4. The design and analysis of the experiments proposed to validate our hypothesis is provided by Section 5 The paper concludes with Section 6, which also suggests some ideas for future work.

2. Text classification

Text mining consists of discovering of previously unknown information from existing text resources (Hearst, 1999). It is also called intelligent text analysis, text data mining, or knowledge-discovery in text. Text mining is related to data mining, which intends to extract useful patterns from structured text or data usually stored in large database repositories. Instead, text mining searches for patterns in unstructured natural language texts (e.g., books, articles, e-mail messages, Web pages, etc.). Text mining is a multidisciplinary field that includes several tasks such as text analysis, clustering, categorisation, summarisation, or language identification.

Text classification is one of key text mining tasks that has gained significant popularity over the last decade or so. One of the main reasons for it is the increasing amount of digital documents available and thus the necessity to access their content in flexible ways (Sebastiani, 2002). Text classification is also referred to as Text Categorisation, Document Classification, or even Topic Spotting. The current approach to text classification is applying the machine learning paradigm that uses of a set of previously categorised documents to automatically build a categoriser by learning from this data (i.e., inductive inference). As part of this whole process, each text document is represented by a feature vector, thus dismissing the order of words and other grammatical issues, as this representation is able to retain enough useful information for the classification task (Salton, 1989).

The next sections describe the sequential steps that shape text classification.

2.1. Document preprocessing

The preprocessing stage starts by tokenising documents. In this step, a text document is transformed into smaller units known as words or terms. It is common that the process also involves the removal of certain characters such as non-alphabetical ones, as well as converting them into lower case. After tokenisation, there are two further steps performed: removal of stop words and stemming of words.

A stop word is a term that is considered not to add significant semantic meaning to sentences. Therefore, they can be safely removed without affecting the whole meaning of the sentence. They mainly consist of topic-neutral words like articles and prepositions.

Stemming is the process of normalising words by applying morphological rules that allow a speaker to derive variants of the same idea to evoke an action (i.e., verb), an object or concept (i.e., noun), or a property (i.e., adjective) (Lovins, 1968). For example, the words *activate*, *activating*, *activeness*, *activation* are derived from the same stem *activ* and all share an abstract meaning of action or movement. Stemming does the reverse process, deducing the stem from a fully suffixed word according to its morphological rules. These rules concern morphological and inflectional suffixes. The former type usually changes the lexical category of words whereas the latter indicates plural and gender. Because most languages have a large number of word stems, applying this technique will most probably reduce the number of global unique terms in all the documents.

These three preprocessing procedures described above (tokenisation, stop-word removal, and stemming) are highly dependent on the language in question. Therefore, the preprocessing of documents can be considered to be language dependent.

2.2. Document modelling

After the documents have been preprocessed, the extracted information from each document is used to build a model representing that particular instance. Feature Selection contributes to this goal by reducing the overall dimensionality of terms, thus allowing the posterior creation of feature vectors to represent the documents. This step is crucial as machine learning algorithms usually work better on low-dimensional data, and they may require too much time or memory when the dimensionality of the data set is high (Salton, 1989). In other words, Feature Selection consists of choosing the subset that contains the most relevant terms of all the existing ones in the collection of training documents.

Because text classification depends on a well-defined set of categories, Feature Selection can be local or global. Global Feature Selection consists of generating a subset of terms from all the terms in all categories, while local Feature Selection creates a subset for each document category, where the most relevant features of the category are included.

Term Frequency (TF) is a very simple yet effective term evaluation function based on counting how many times each term appears across all documents.

The higher this count, the more relevant this term is considered. Document Frequency (DF) and inverse document frequency (IDF) are very similar to TF and are based on the count of documents each term appears in. The reason for having these two complementary functions is that in some cases, and depending on the characteristics of the document collection, the feature selection may work better when only the terms that appear in the most documents are kept, while in other situation it may be just the opposite.

The term evaluation function χ^2 calculates the dependence between the occurrences of a term and each category based on the number of expected vs observed occurrences.

After feature selection, the documents are then modelled. A very commonly used algebraic model is the Vector Space Model (VSM), which represents text documents in a high-dimensional