Text Mining in Healthcare

Applications and Opportunities

By Uzma Raja, PhD; Tara Mitchell; Timothy Day, PhD; and J. Michael Hardin, PhD

KEYWORDS

Text mining, electronic clinical records, domain knowledge, predictive models, natural language processing.

ABSTRACT

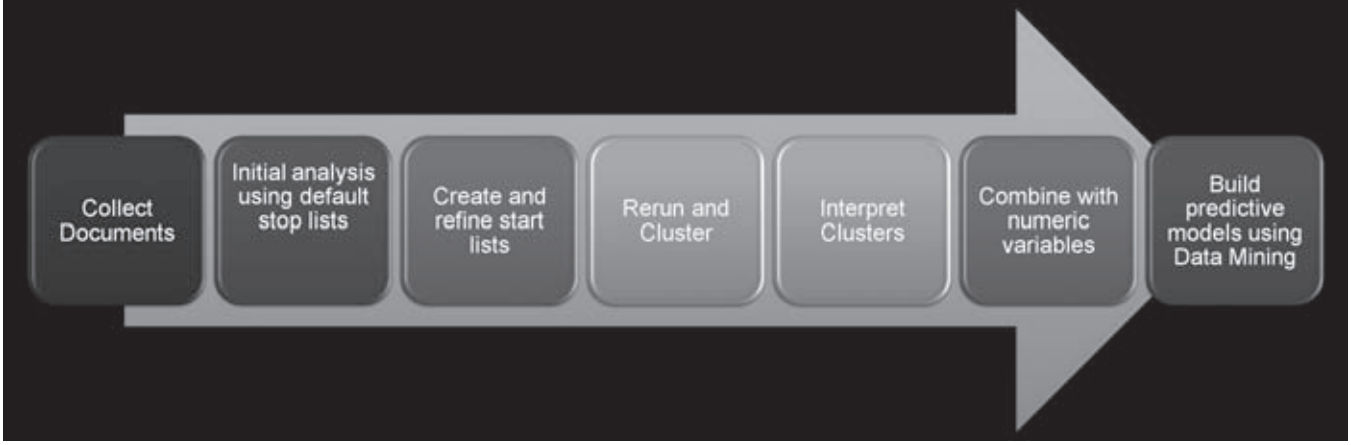
Healthcare information systems collect massive amounts of textual and numeric information about patients, visits, prescriptions, physician notes and more. The information encapsulated within electronic clinical records could lead to improved healthcare quality, promotion of clinical and research initiatives, fewer medical errors and lower costs. However, the documents that comprise the health record vary in complexity, length and use of technical vocabulary. This makes knowledge discovery complex. Commercial text mining tools provide a unique opportunity to extract critical information from textual data archives. In this paper, we share our experience of a collaborative research project to develop predictive models by text mining electronic clinical records. We provide an overview of the text mining process, examples of existing studies, experiences of our collaborative project and future opportunities.

Text mining refers to the discovery of knowledge from textual data. Text contains abundant qualitative information that is difficult to use in statistical modeling. However, traditional-model building requires quantifiable, tangible information. Text mining converts text into numeric form, which allows it to be used for analysis. There are several text mining algorithms suitable for a variety of problem domains. Text mining has been used in sociology and communication to extract the intangible information hidden in words. The question is whether text mining can be used to improve healthcare quality.

Figure 1 shows the general strategy of building predictive models supplemented by text analysis.

The text-mining process begins with collecting the documents to be analyzed. Domain knowledge plays a vital role in extracting knowledge from text. A field expert decides on the criticality of a word occurrence. Data extraction and cleaning is a labor-intensive process that requires the careful attention of domain experts to ensure that the data is valid and the information is complete. Most text-mining tools provide two options for analysis: Ignore commonly used words that do not carry value (such as prepositions or articles) or analyze an exclusive list of terms. The tool that we used allows the creation of start lists and stop lists. A stop list contains terms that are to be ignored during the analysis. This provides a more reliable analysis of term occurrences. Once an initial run is performed, the results present all the terms that occur in the document, their frequency and the "relationship" of the terms. For example, if the term "shortness of breath" always appears with the term "heart attack," it's indicative that the patients have these symptoms in common.

Fig. 1: A general strategy for model building using text mining



Another effective analysis is through the use of a start list. The start list restricts the analysis to a defined list of terms. For example, if we are exploring the relationship between smoking and cancer, we can perform an analysis exclusively on a list of terms related to "cancer" and "smoking."

Once the analysis is complete, it provides the ability of clustering documents based on the terms used. The identification number of the cluster and the distance between the clusters provides useful numeric data that can be used to represent textual information in a traditional model-building process.

TEXT MINING IN HEALTHCARE

Several research studies have focused on the processing of textual information available in healthcare datasets. Following is a brief overview of studies that highlight the significance of textual data and its suitability in research settings:

One notable research initiative in was performed at the Vanderbilt Clinic, in New York City.[1] The objective was to determine if a natural language processing (NLP) program could automatically code functional status information in accordance with the International Classification of Functioning, Disability and Health (ICF) requirements. Automated coding is an obvious choice for these types of initiatives. Coding is extremely important for reimbursement purposes and record-keeping. It also is a tedious and time-consuming process. If this could be accomplished accurately with technology it would save the medical facilities substantial resources.

The researchers extended the existing NLP Medical Language Extracting and Encoding (MedLEE) to code rehabilitation discharge summaries. Ten ICD-9 codes were pre-selected for their known relationship to changes in functional status. Evaluations were performed by the NLP system, expert coders and non-expert coders. They found that the NLP system coded with results similar to that of human coders. This is a promising finding for research into automated coding for ICD-9 codes, which are the main basis for reimbursement in a majority of healthcare settings.

A study conducted at the University of Utah used a modified version of MedLEE, as well as a phrase-matching algorithm, to extract data for research initiatives.[2] Most electronic records are dictated in a narrative form. Manually retrieving specific data for research can be time-consuming and expensive. The purpose of this study was to extract data related to adverse events connected to central venous catheter placement. Adverse events can include a variety of serious complications, such as infections and a collapsed lung. Tests were conducted using each method individually and then using them together on a sample of records that had been manually reviewed beforehand. The trials using the individual methods were unsuccessful. The phrase-matching algorithm was not specific enough and the NLP system was not sensitive enough. They produced positive prediction values of 6.4 percent and 6.2 percent, respectively. However, when used together the results were promising. They yielded a 72 percent sensitivity and 80.1 percent specificity, which are acceptable values. This study shows potential for using NLP systems to automate research data extraction.

Event detection is another significant area of research. Hazlehurst et al. performed a study to identify vaccine reactions for the Vaccine Safety Datalink Project (VSD).[3] The VSD is a partnership between the CDC and eight large HMOs to investigate adverse events following immunization by analyzing medical care databases and patient medical records. In this study, the VSD used a modified version of the NLP system, MediClass, which had been programmed with the knowledge necessary to detect possible vaccination reactions.

It achieved both a high sensitivity and specificity percentage. Compared with other methods used by clinicians, this system significantly improved the positive predictive value. Studies such as these are especially important because the ultimate goal is to migrate to a system that can predict such occurrences in future.

Recently, text mining tools have been utilized in healthcare research. Cerrito and Cerrito analyzed electronic medical records from the emergency department of a hospital over a six-month period using text mining.[4] The authors found that similar complaints were treated differently, depending on the physician on-

Table1: Pathology clusters using a dedicated start list.

Cluster	Terms	Frequency
1	+ lesion, endocervical, intraepithelial lesion, intraepithelial, vaginal, malignancy, fungal, atrophy	326
2	bone, + plasma, clonal, bone marrow, leukemia, myeloid, marrow biopsy, immature, bone, marrow	41
3	renal, amputation, kidney, skin, liver, hysterectomy, + artery, vascular, necrotic, fibrosis	143
4	+ mass, + tumor, + carcinoma, + lymph, + node, endometrial, cervix, + ovary, cell block, endometrium	360
5	squamous intraepithelial lesion, vaginalis, trichomonas, vaginalis present, cytopathic effect, mild dysplasia/hpv, hsil, + squamose, thinprep, endocervical	90
6	tubular adenoma, + polyp, rectum, + adenoma, hyperplastic, colon, sigmoid, hyperplastic, descending, polypectomy	81
7	malignancy, inflammation, thinprep, intraepithelial, + lesion, intraepithelial lesion, vaginal, cellular change, fungal, atrophy	100
8	fna, + hyperplasia, thyroid, prostate, prostate, thyroid, colloid, goiter, prostatic, prostatitis	66

call with an optimized treatment plan. It also can lead to the development of protocols to alleviate treatment disparities.

UA/UAB CASE STUDY

The analysis of clinical records requires two major steps: processing large volumes of textual data and developing useful predictions based on the data. The analysis of the textual data requires an understanding of algorithms that can convert this data to useful numeric information. This conversion leads us to datasets that can be coupled with other quantitative information collected and analyzed traditionally to improve the predictions we can make about various outcomes.

Any text-mining project requires domain experts and technical experts. Thus, a collaborative research project between the University of Alabama-Birmingham (UAB) and the University of Alabama (UA) was initiated to explore the applications of text mining in electronic clinical records. SAS Institute Inc. provided the software and licensing support needed for the project.

The UAB dataset contains patient, diagnosis and prescription information. UAB health system has one of the largest information systems in the nation. These records contain vast amounts of unstructured data that is typically overlooked due to the immense size of the medical record collection at the hospital. The facility also houses experts who can extract and interpret data. UA experts in data mining and text mining provided the analytical support needed for the project.

UAB has a Web-based electronic clinical record system. This system has enterprise-wide availability. Clinicians can view labs, documents, reports, demographics etc. They can create clinical documents, edit and sign documents, and have secure physician communication regarding patients. As of April 2003, the system had 13,279 users and data for 661,533 patients. The total number of documents as of August 2006, was ICDA: 1,923,220 and CDA: 4,580,168. This volume of data rules out the possibility of manual extraction of knowledge from these archives.

The data archives contain various types of documents. Each medical record has numerous documents, making a single complete record large and cumbersome. We narrowed our analysis to pathology reports and discharge summaries; both are moderately stylized and contain specialized vocabularies. Since the purpose of the study was to evaluate the applicability of text mining in the domain of clinical records, the choice of documents ensured that inconsistency in document styles and a lack of standardization did not adversely affect the evaluation process.

To reduce complexity, we decided to start with a single portion of the record. We selected pathology reports because those were the medical issues that we had the most experience with and on our initial run we wanted to make sure we got results that we could identify as clinically relevant.

We began our first run with the entire pathology report from a set of 1,500 records. These pathology reports were surgical, cytology and addendum reports and included the clinical summary, gross description, microscopic description and the diagnosis. For the initial run, we did not perform any major manipulations of the data. We were performing an "unsupervised analysis." In such

analysis, no key terms or start lists are defined. Instead, all word occurrences are analyzed to provide an understanding of the dataset. Such studies are useful in new domains where we want to explore the data in its entirety.

Clustering of the documents was used to discover patterns that exist in the textual contents. We wanted to let the document clusters be discovered, but we wanted the clusters to be clinically interesting.

Some important settings to note were the default start/stop lists were used, the terms only in a single document were ignored, parts of speech were tagged and we used term weight entropy to cluster. The first run produced three large clusters that were seemingly useless. By further analyzing the most common words we realized that our results were being skewed by "noise"—words that were clinically irrelevant. We decided to eliminate this noise in accordance with the recommendation that if the text contains a technical vocabulary you should begin with a dedicated start list.[5]

To build a pathologically relevant start list, we went through the list of terms by hand and kept only the diagnostically appropriate terms. We also kept track of interesting phrases—"no significant histopathologic change"—to incorporate into a synonym list.

Table 2: Discharge summary clusters using the RxNorm start list.

Cluster	Terms	Frequency
1	NONE 62	62
2	+ zosyn, augmentin, + antibiotic, + support, ciprofloxacin, + multivitamin, oxygen, vancomycin, blood, tobacco	50
3	+ lasix, + lanoxin, albuterol, insulin, oxygen, troponin, + potassium, + sodium, + glucose, + sinus	232
4	+ synthroid, zinc, zoloft, nephro□vite, tobacco, + vitamin, alert, + pepcid, + allergy, magnesium	53
5	amylase, lipase, + glucose, + protein, hemoglobin, blood, alkaline phosphatase, + calcium, + potassium, + sodium	117
6	+ dilantin, + prinivil, + zocor, + lopressor, albuterol, + coumadin, prednisone, + antibiotic, hemoglobin, blood	39
7	+ coumadin, heparin, + ambien, tylenol, colace, + vitamin, + allergy, + tablet, + control, + lopressor	108
8	nephro□vite, + renagel, + monopril, enalapril, + pravachol, + tenormin, + pepcid, + lopressor, aspirin, + coumadin	65
9	+ depakote, + haldol, + klonopin, premarin, + desyrel, + remeron, liver, albumin, + neurontin, hemoglobin	30
10	+ control, colace, + tablet, + percocet, tylox, labetalol, lortab, + prinzide, + allergy, + multivitamin	149
11	+ antibiotic, yeast, vancomycin, + zosyn, bactrim, blood, + tequin, tylenol, + protein, + glucose	77
12	bactrim, prograf, prednisone, valcyte, cellcept, flagyl, + multivitamin, + zosyn, + sodium, blood	39
13	+ sleep, + klonopin, + desyrel, paxil, stress, + vitamin, + nitrol, + sinus, oxygen, aspirin	51
14	bile, liver, prednisone, + actigall, nadolol, prograf, paxil, nystatin, premarin, alkaline phosphatase	36
15	carboplatin, lortab, + advil, + robaxin, mestinon, taxol, oxycontin, doxil, decadron, valium	46
16	+ glucotrol, + glucophage, + glucovance, + glucophage xr, + zocor, + plavix, + lipitor, aspirin, + lopressor, insulin	61
17	+ lipitor, aspirin, + nitrol, + plavix, + enzyme, troponin, + prinzide, + lopressor, stress, + tenormin	100
18	+ cordarone, + lead, + sinus, + coumadin, + lopressor, + lipitor, + lanoxin, + lasix, + enzyme, troponin	33

When we ran the pathology reports with this revised start list we had much more promising results. Our most common words changed to things such as "malignancy," "lesion," "benign," "polyp" and "carcinoma." The results, shown in Table 1, indicate noteworthy clusters. For instance, the reports in Cluster 1 are cytologies; Cluster 2 has to do with bone marrow biopsies; Cluster 3 is kidney pathologies; Cluster 4 is tumors; and Cluster 5 is thinprep cytologies. This encouraging result indicated the ability of the text miner to work with specialized vocabulary and complicated data sets.

We realized the significance of a dedicated start list because of the specialized vocabularies in our data. Going through the list of terms to manually develop a start list is inefficient and time-consuming. There were two areas where the start list vocabularies were readily available: malignancies and medications. Since we already had a fairly complete start list for malignancies we decided to move on to medications to avoid a duplication of effort.

To ensure statistical validity of our analysis, we selected a different set of documents. Therefore, we used the discharge summaries to extract medication information. Hospital administrators want to explore drug reconciliation within the hospital and determine the likelihood of re-admittance or other complications after being prescribed certain medications. The start list for this analysis was created using the RxNorm vocabulary from the National Library of Medicine's Unified Medical Language System (UMLS). The RxNorm vocabulary is a comprehensive list of standardized nomenclature for clinical drugs, their ingredients, strengths and forms. We began with the "RXNCONSO" file which is a delimited text file containing the concept and source information for RxNorm. It is the most inclusive of the files in the RxNorm vocabulary.

Using this start list and a set of 15,000 discharge summaries, we attempted a preliminary run through the text miner using the default settings. The results were very encouraging. Some of our most common words were "alcohol," "blood," "aspirin" and "Coumadin." Our clusters also gave promising results.

The resulting clusters are shown in Table 2. In Cluster 1 there

are no terms because these patients did not receive medication. We can assume that the patients in Cluster 2 had some kind of infection because the drugs are antibiotics. In Cluster 6 we see that these patients all received muscle relaxants, and in Cluster 7 we can assume some form of heart problem from the medications listed. This was important because we were able to group patients with similar conditions together.

While all the drugs that were of the same type did cluster together, we wanted certain terms, such as "acetaminophen" and "Tylenol," to be considered a single term. To alleviate this problem, we created a synonym list. A synonym list identifies the synonyms of the terms and uses the information during text mining to correlate the occurrence of synonyms. We used the RxNorm files for this purpose by stripping the RXNCONSO dataset down to the RXCUI and the drug name. Then, using RXNREL, which is the relationship file, we narrowed it down to the RXCUI1, RELA and RXCUI2 variables. RXCUI1 and RXCUI2 are unique drug identifiers and the RELA variable identifies the relationship between two drugs—whether they share an ingredient or of one is a generic form of the other, etc. We wanted to ensure that generic forms were seen as the same as the brand name drug. We therefore sorted the new RXNREL table and kept only those entries whose RELA was tradename. We then merged the two new datasets and dropped the RXCUI and RELA variables to form the synonym list. The use of synonym list resulted in similar clusters. These clusters can now be used with other numeric data to create predictive models.

CONCLUSIONS AND FUTURE DIRECTION

Our initial study indicates that text mining can be an effective tool in healthcare datasets. With shift to electronic clinical records and availability of standardized vocabulary the future of text mining in this domain is bright.

We made extensive use of the existing standard vocabularies collected by the National Library of Medicine in the UMLS system. We believe that the availability of these libraries make text mining a very effective choice in healthcare analytical projects. Currently, we are building predictive models from clinical record archives that can accurately predict future outcomes in various healthcare settings. We supplement our models with the results of text analysis to make them more reliable and robust.

There are many challenges to using text mining on healthcare data, especially in a joint venture similar to ours. Coordinating multiple schedules at multiple institutions can be a challenge. Healthcare studies have the added complexity of HIPAA compliance and protection of confidential information. Yet the need for joint ventures is critical because the domain knowledge and methodology knowledge resides in multiple institutions. With the availability of modern tools and techniques to mask confidential data and to work in geographically dispersed teams, we are hopeful that most of these challenges can be overcome.

It is estimated that between 44,000 and 98,000 people die every year due to medical errors, making text mining a hot research topic. While technology can never replace doctors, it can serve as a very capable redundancy system that could greatly reduce medical error-related deaths or complications. Text mining can minimize adverse drug interactions and uncover correlations between specific patient characteristics and adverse reactions to proposed treatments that may not be evident using clinical research alone. This type of warning system could save countless lives and large sums of money. **JHIM**

Uzma Raja, PhD, is an Assistant Professor at the University of Alabama.

Tara Mitchell is an undergraduate student at the University of Alabama.

Timothy Day, PhD, is a Senior System Analyst at the University of Alabama at Birmingham Health System.

J. Michael Hardin, PhD, is a Professor at the University of Alabama.

REFERENCES

1. Kukafka R, Bales ME, Burkhardt A, Friedman C. Human and automated coding of rehabilitation discharge summaries according to the international classification of functioning, disability, and health. *J Am Med Inform Assoc*. 2006;13(5):508-15.

2. Penz JF, Wilcox AB, Hurdle JF. Automated identification of adverse events related to central venous catheters. *J Biomed Inform*. 2007;40(2):174-82.

3. Hazlehurst B, Mullooly J, Naleway A, Crane B. Detecting possible vaccination reactions in clinical notes. AMIA Annual Symposium Proceedings, 2005.

4. Cerrito P, Cerrito JC. Data and text mining the electronic medical record to improve care and to lower costs. SUGI 31 San Francisco, Calif., 2005.

5. Raja U, Tretter M. Model formulation, validation and testing, using SAS enterprise miner. SAS User Group International (SUGI) 31, San Francisco, Calif., 2006.