TEXT MINING IN HEALTHCARE: APPLICATIONS AND OPPORTUNITIES

Uzma Raja*, Tara Mitchell*, Timothy Day** and J. Michael Hardin*
*University of Alabama, **University of Alabama at Birmingham

INTRODUCTION

Electronic clinical records contain information on all aspects of health care. Healthcare information systems that collect large amounts of textual and numeric information about patients, visits, prescriptions, physician notes etc. The electronic documents encapsulate information that could lead to: improvement in health care quality, promotion of clinical and research initiatives, reduction in medical errors, and reduction in healthcare costs. However, the documents that comprise the health record are wide ranging in complexity, length and use of technical vocabulary, making knowledge discovery complex. Recent availability of commercial text mining tools provides 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, some of the existing studies in the areas, experiences of our collaborative project and the future opportunities in this area. The goal is to evaluate the potential of harnessing the capabilities of text mining in healthcare settings.

TEXT MINING: AN OVERVIEW

Text Mining refers to the discovery of knowledge from textual data. Text contains abundant qualitative information that is difficult to use in statistical modeling. In fields of healthcare, physicians express opinions in terms of words that contain useful information, not captured elsewhere. This information can be further utilized to develop intelligent models to improve healthcare process. However, traditional model building requires quantifiable, tangible information. Text mining converts text into numeric form that allows it to be used for analysis. There are several text mining algorithms available that are suitable for a variety of problem domains. This technique has been widely used in areas of 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 that are supplemented by text analysis. The text mining process begins will collection of the documents that are to be analyzed. Domain knowledge plays a vital role in extraction of 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 terms in the analysis or perform analysis on 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 in the document while performing the analysis, e.g. commonly occurring terms e.g. “of, on, the” that do not carry much value are ignored. This provides a more reliable analysis of the term occurrences. Once an initial run is performed, the results present all the terms that occur in the document set along with their frequency and relationship. This relationship signifies the co-occurrence of terms, e.g. if the term “shortness of breath” always appear in the document with the term “heart attack”, it’s indicative that the patients have these symptoms in common.

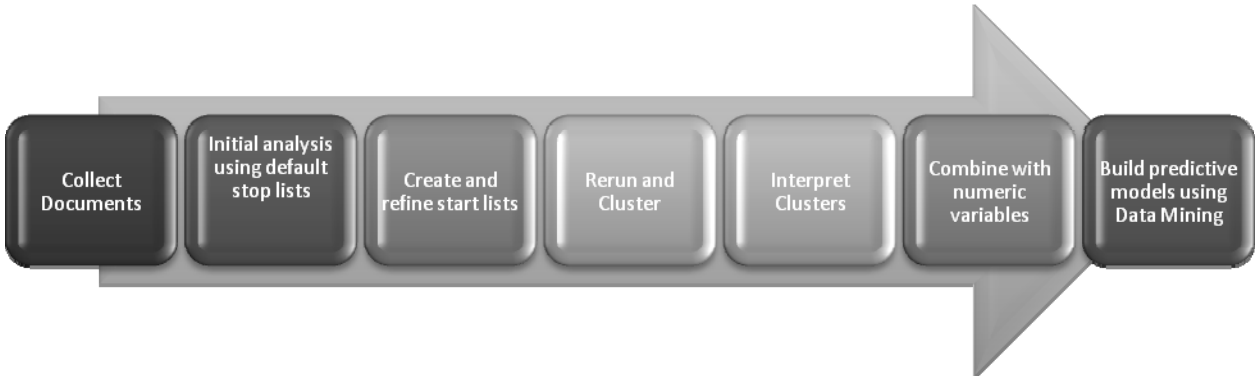

Figure 1: A general strategy for model building using text mining

An effective way of performing the analysis is the use of start list. The start list restricts the analysis to a defined list of terms, e.g. if we are only exploring the relationship between smoking and cancer, we can create a list of smoking related terms and cancer related terms and only perform analysis on these terms.

Once the analysis is complete, it provides the ability of clustering the documents based on the similarity (or difference) of the terms that occur in them. 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 traditional model building process.

TEXT MINING IN HEALTH CARE

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

One notable research initiative in was performed at the Venderbilt Clinic, New York [1]. The objective was to determine if a natural language processing program (NLP) 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; however, it is also a very tedious and time-consuming process. If this could be accomplished accurately with technology it would save the medical facilities a substantial amount of resources. The researchers extended the existing NLP 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 similar results to the human coders. This is a promising finding for research into automated coding for ICD-9 codes, which are the main basis for reimbursement, in majority of healthcare settings.

A study, conducted at the University of Utah [2], used a modified version of MedLEE as well as a phrase-matching algorithm to extract data for research initiatives. Most electronic records are dictated in a narrative form and 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 be things such as infections, complications from misplacement, and pneumothorax (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 and 6.2% respectively. However, when used together the results were promising. They yielded a 72.0% sensitivity and an 80.1% specificity which are acceptable values. This study shows potential for using NLP systems to automate research data extrcation.

Event detection is another significant area of research. Hazlehurst et. al [3] preformed a study to identify vaccine reactions for the Vaccine Safety Datalink Project (VSD). The VSD is a partnership between the CDC and eight large HMO's to investigate adverse events following immunization. They are attempting to do this by analyzing medical care databases and patient medical records. In this study, a modified version of the NLP system MediClass which had been trained with the knowledge necessary to detect possible vaccination reactions was used. It achieved both a high sensitivity and specificity percentage. Compared with methods that are 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, e.g. Cerrito and Cerrito [4] analyzed the electronic medical records from the emergency department of a hospital over a six month period, using text mining. They found that similar complaints were treated differently depending on the physician on call. Such differences can affect care quality and costs. Therefore text mining of prior expert treatment can provide physicians on call with an optimized treatment plan. It can also lead to development of protocols to alleviate disparity in treatment.

UA/UAB CASE STUDY

The analysis of clinical records requires two major steps: processing the large volume of textual data and developing useful predictions based on the data. The analysis of the textual data requires understanding of algorithms that can convert this information to useful numeric information. This conversion leads us to datasets that can be coupled with other quantitative information collected and analyzed traditionally (e.g. patient age, gender, vital statistics) to improve the predictions we can make about various outcomes.

Any text mining project required domain experts and technical experts. Thus a collaborative research project between University of Alabama at Birmingham (UAB) and the University of Alabama (UA) was initiated to explore the applications of text mining in electronic clinical records. SAS Inc provided the software and licensing support needed for the project. UAB dataset contains patient, diagnosis and prescription information. The facility also houses experts who can extract and interpret the data. UA experts in data mining and text mining provided the analytical support needed for the project.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.

UAB has a web-based electronic clinical record system. This system has enterprise-wide availability. The 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 of 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 different documents making a single complete record very large and cumbersome to work with. 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 was to ensure that issues like inconsistency in document styles and lack of standardization did not adversely affect the evaluation process.

In order 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 1500 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 the word occurrences are analyzed to provide an understanding of the dataset. Such studies are useful in new domains where we do not want to constrain the analysis and 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, though 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 word which were things such as pm, glass, specimen volume, and hair-bearing we realized that our results were being skewed by “noise,” words that were clinically irrelevant. We decided to try and 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 like “no significant histopathologic change” to incorporate into a synonym list. 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, legion, benign, polyp, and carcinoma. The results, as 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 are kidney pathologies, cluster 4 are

tumors, and cluster 5 are thinprep cytologies. This result was very encouraging and indicated the ability of the text miner to work with specialized vocabulary and complicated data sets.

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

Table1: Pathology Clusters Using a Dedicated Start List

Through this attempt 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 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. The hospital administration is interested to explore the drug reconciliation within the hospital and to 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 15000 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 gave promising results as well. The resulting clusters are shown in table 2. In cluster 1 there are no terms because these patients did not receive any medications and we can assume that the patients in cluster 2 all probably had some kind of infection because all the drugs are types of 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 there. This was important because we were able to group patients with similar conditions together.

Cluster #	Terms	Frequency
1	NONE	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

Table 2: Discharge Summary Clusters Using the RxNorm Start List.

While all the drugs that were of the same type did cluster together, we noticed certain terms such as acetaminophen and Tylenol that are essentially the same thing that we would like to have been considered as a single term. So to alleviate this problem we decided to create a synonym list. A synonym list identifies the synonyms of the terms and uses the information during text mining to correlate the occurance 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 tells the relationship between the two drugs; whether it is an ingredient in the other, one is a generic form of the other; one is a different dose form of the other, etc. We wanted to ensure that generic forms were seen as the same as the original drug. We therefore sorted the new *rxnrel* table and kept only those entries whose *RELA* was *tradename_of*. 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, yet considerable simplified the terms in them. These clusters can now be used with other numeric data to create predictive models.

CONCLUSIONS AND FURUTRE 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. Our work on creation land use of these libraries in terms of start lists and synonym lists can be beneficial for the healthcare community in general. 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 creates. 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 would be over comes.

It is estimated that between 44,000 and 98,000 people die every year due to medical errors, making this a hot topic for research. It is fully within our reach to create predictive models using patient records that will significantly reduce that. While technology can never replace doctors it can serve as a very capable double-check system that could greatly reduce medical error related deaths or complications. Some areas that could be greatly affected are minimizing adverse drug interactions and uncovering correlations between specific patient characteristic and adverse reactions to proposed treatments that may not be evident due to clinical research alone. This type of warning systems could save countless lives and large sums of money.

REFERENCES

[1] R. Kukafka, M. E. Bales, A. Burkhardt, and C. Friedman, "Human and automated coding of rehabilitation discharge summaries according to the international classification of functioning, disability, and health.," *J Am Med Inform Assoc.*, vol. 13, pp. 508-523, 2206.

[2] J. F. Penz, A. B. Wilcox, and J. F. Hurdle, "Automated identification of adverse events related to central venous catheters." *Journal of Biomed Inform*, vol. 40, pp. 174-182, 2007.

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

[4] P. Cerrito and J. C. Cerrito, "Data and text mining the electronic medical record to improve care and to lower costs," in *SUGI 31* San Francisco, CA, 2005.

[5] U. Raja and M. Tretter, "Model formulation, validation and testing, using SAS enterprise miner," in *SAS User Group International (SUGI) 31*, San Francisco, CA, 2006.

ABOUT THE AUTHORS

Uzma Raja, PhD (uraja@cba.ua.edu) is an Assistant Professor at University of Alabama.

Tara Mitchell (tara_dawn77@hotmail.com) is an undergraduate student at University of Alabama.

Timothy Day, PhD (tday@uabmc.edu) is a Senior System Analyst at the University of Alabama at Birmingham Health System.

J. Michael Hardin, PhD (mhardin@cba.ua.edu) is a professor at the University of Alabama.