Text Mining And Analysis in Healthcare

Introduction

The goal of this paper is to review existing text data applications utilized in the healthcare domain to improve quality of care and decrease costs. A few of the examples of text analysis applications are analyzing patient profiles to automate or assist with diagnosis coding while discovering trends of data across a population of patients can generate knowledge of epidemics.

NLP in Healthcare

There are vast amounts of research of using Natural Language Programs (NLP) to to create healthcare centric applications. Kukafa, Bales, Burkhardt, and Friedman [1] adapted MedLEE [2] an existing medical NLP program to encode patient information from discharge summaries. The conclusion was that the NLP program still needs to be improved but the performance was similar to their human counterparts.

Melton and Hripcsak [3] were able to successfully detect adverse events [4] in discharge Summaries using an adapted version of the MedLEE system. The system could not solely be relied on automating due to its low sensitivity, however once it presented a result it could be trusted due to its high sensitivity.

Hazlehurst, Frost, Sittig, and Stevens created MediClass [5] a knowledge system for detecting and classifying clinical events from Electronic Medical Records. The system was designed to identify medical concepts using NLP algorithms in both free and structured data and generate classifications based on the identified concepts. The dictionary of concepts used were from the Unified Medical Language System (UMLS) [6]. The classification system used was based on rules derived from the medical concepts generated. The MediClass system was used for detecting possible vaccination reactions, disease survelliance and care quality. It was determined that the MediClass had similar performance coding encounters when compared to human experts.

A common challenge with using NLP that was noted throughout all the papers was that sentence structure is not always grammatically correct which causes more ambiguity and makes it more difficult for NLP to be accurate. Another common challenge is that some of the content in the free text sections is actually structured /canned response that can be easily entered using a macros workflow. In addition to that, the language in the medical domain is highly specific, this means that expert domain knowledge is needed to create much more effective NLP systems that can identify special medical phrases or terms.

Text Clustering Analysis Approaches

Raja, Mitchell, Day and Hardin [7] focused on converting the unstructured text to numerical data that can be utilized to create predictive models. To accomplish this task their strategy was to create clusters of words from 1500 pathology reports which contained clinical summaries. In the first attempt, the reports were categorized into three clusters using term weight entropy. The clusters were deemed meaningless and contained a lot of noise. The solution to this problem was to use a

[1] Human and Automated Coding of Rehabilitation Discharge Summaries According to the International Classification of Functioning, Disability, and Health, https://academic.oup.com/jamia/article/13/5/508/734020

[2] MedLEE, https://academic.oup.com/jamia/article/13/5/508/734020

[3] Melton and Hripcsak, https://europepmc.org/article/med/15802475

[4] Preventing Adverse Events, https://www2.health.vic.gov.au/hospitals-and-health-services/patient-care/older-people/resources/improving-access/ia-adverse

[5] MediClass, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1205600/

[6] UMLS, https://www.nlm.nih.gov/research/umls/index.html

[7] Text Mining in Healthcare: Applications and Opportunities, https://www.researchgate.net/publication/24182770_Text_mining_in_healt

start list to form the clusters, the formation of this start list required an expert to manually review the unique terms throughout the corpus. The second iteration then focused on creating clusters based on the words in the defined start list, the outcome were 8 clusters that included themes of bone marrow biopsies, kidney pathologies, tumors and thinprep cytologies.

The authors also attempted to investigate if there was a relation between hospital re-admittance and complications after a patient was subscribed to certain medications. A start list was created using the RxNorm database [8] which contains an exhaustive vocabulary of clinical drug terms. Cluster analysis was then performed which resulted in 18 clusters which contained similar groups of medications associated with discharge summaries. No further analysis of the data was provided, however one point that the researchers failed to mention is how would they classify medications that the patient was already taking versus medications that were prescribed due to complications. I'm assuming that they would have to use historical recorded data to account for this. I also believe that their findings can also be compared and or combined against coded diagnosis and medication data that already exists in the patient chart.

Vector Space Model for Automated Diagnosis

Pendyala, Fang, Holliday, Zalzala [9] attempt to automate diagnosis detection based on symptoms reported from a patient. The motivation for the automation is to improve healthcare access for underprivileged populations. To ensure that this was possible, the authors performed cluster analysis of discharge sheets to see if there was a standard association between the clusters generated and the actually diagnoses for the patients. Once the clusters generated matched the grouped of diagnoses, the authors decided to investigate if they can match a patient symptom document against a discharge document to detect the appropriate diagnoses. The strategy they used to accomplish this was to use a vector space model where each word in the corpus represented a dimension. All of the discharge documents were computed as vectors and a query vector was also formed for the patient symptoms. The goal was then to figure out which document vector was most similar to the query symptom vector, this was determined using a k-nearest neighbor algorithm.

It was determined that this approach produced promising results but was not specific or sensitive enough to automate diagnosis detection. I believe that a factor that decreased the value of the solution produced is the difference in jargon used by medical professionals versus patients to describe the same problem.

Conclusion

Although text applications are not ready to be utilized in the day to day healthcare domain, research over the years continues to make that reality seem closer. Common themes with creating successful applications require an expert knowledge of the area, this trait illustrates how complex the medical domain is. The most popular applications for text mining are resolved around diagnosis detection and adverse event detection. Both of those categories can provide information that can lead to increased quality of care and decreased costs for patients and healthcare businesses such as hospitals, clinics and insurances.

Common challenges that plague healthcare text data are grammatical errors and highly specific sentence structures that are unique to the field. Another common challenge is adhering to HIPPA Compliance [10] rules for data security and patient privacy. This acts as a big hindrance to innovation as it limits the amount of data that is actually accessible to researchers. The general solution

[8] RxNorm, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1205600/

[9] A Text Mining Approach to Automated Healthcare for the Masses, https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6970257

[10] Summary of HIPPA security rule, https://www.hhs.gov/hipaa/for-professionals/security/laws-regulations/index.html

to this problem is investing in De-Identification tools, however the risk associated with violating HIPPA makes organizations lean towards doing this process manually which is a mundane and tedious operation. The MediClass system avoided this issue by ensuring that the whole system designed ran locally on healthcare organizations servers and did not have any data leave their existing network/servers.