Automated Classification of Medical Documents

John Savage

Indiana University

**Title: Automated Classification of Medical Documents**

**Team: John Savage**

**Project Description:**

**Objective:**

This project will classify textual medical documents that are typically been transcribed by transcriptionists or voice recognition systems. The goal is to classify the documents as office visits (e.g., consult, history and physicals, discharge summaries), procedures (e.g., surgeries, endoscopies, epidural injections), or tests (e.g., imaging, stress tests, electromyograms). This will solve the problems of using office staff time to sort incoming documents and prioritizing medical providers time utilization in reviewing documents

**Usefulness:**

This classification application would eliminate need for medical office clerical workers to assign a class to medical documents arriving at the office for entry into the proper partition within the electronic medical record (EMR). This application also allows the medical provider to prioritize the review of incoming patient documents. The classification allows automated scanning of the EMR for medical procedures or tests that are performed at specific intervals for disease surveillance as a part of preventive care.

Most classification. is used for billing and coding of medical tasks and procedures. Hyland Healthcare describes a system for classifying documents similar to the goals of the project. Hyland’s description of their system implies it handles the classification and attaches patient demographic information but doesn’t identify prioritizing document review by class.[A System For Health Document Classification Using Machine Learning](https://samphina.com.ng/system-health-document-classification-machine-learning/) is described on the Saphina.com.ng website that classifies documents on the internet by subject. (A System for Health Document Classification Using Machine Learning, n.d.) Shrutha Kashyap, Sushma M G, Varsha Rajaram, Vibha S, 2015, in Medical Document Classification describe their system as a classifier of medical papers by system such as cardiology or pulmonology. (Shrutha Kashyap, 2018) This project is different in that the documents will be classified for the purposes of expediting healthcare provider workflows. The target users are primary care providers which number approximately 295,000 in the US as of 2010. (Quality, 2018)Primary care providers are targeted because of the high flow of medical documents through primary care offices. Effective use of NLP could save more than $3 billion in the US annually (my estimate based on: 2 hours weekly savings, $100/hr, 295,000 providers) and an incalculable benefit to the quality of patient care.

**Data:**

The dataset is Medical Transcriptions found on the Kaggle website taken from mtsamples.csv which can be found elsewhere on the web. (BOYLE, 2018) This data set was created by MTSAMPLES.COM which is intended for learning by transcriptionists and likely medical residents. The documents have been submitted by various users of MTSAMPLES .COM and are representative of actual physician transcriptions. (MTSAMPLES.COM Home, 2022)

The data is in CSV format. It has not been cleaned for use by NLP applications. There is a disclaimer that there may be unusual characters included in some of the dictations. The dataset, mtsamples.csv, has 4998samples consisting of 5 fields (‘description’, ‘medical\_specialty’, ‘sample\_name’, ‘transcription’, ‘keywords’). The fields are a brief description, label of medical specialty, sample name, transcription, and keywords. The data is labeled by specialty and sample name, but these are not useful for labels for the purposes of this project. Therefore, the data will be the unlabeled transcription and unsupervised learning will be used to group the document in 3 classes. Labels will be created for validation (10%) and test data (5%) for purposes of evaluation and assigning classes. The data is transcription data so symbols, numbers, and punctation will need to be removed. I’m unable to think of any special stop words; however, numbers may be indicative of test data.

**Functionalities:**

The main NLP objective will be to classify the documents as office visit, procedure, or test. An extension the classification would be to classify the test data as urgent (eg cardiac tamponade, pneumothorax), critical (eg, cancer diagnosis, follow-up is needed)), or unremarkable but this may not be achievable as the data is not labeled; however, perhaps unsupervised sentiment analysis techniques with a custom lexicon could be utilized as the importance of test results is somewhat analogous to sentiment. The main objective will be to view the documents as topic models.

User interaction will consist of loading documents and sorting these documents by topic. Retrieving the documents by topic. This will simulate the operation of office staff scanning various documents. Retrieving the documents by topic will simulate the provider reviewing the individual documents. An extension could be to sort procedures by type. However, the main objective is to allow the provider to review the tests, procedures, and office notes in that order which is there order of their importance to decision making.

**Communication and Sharing:**

Key information and data will be posted on git hub at the following repository link: <https://github.com/jbsavage57/DSC590_project.git>. This is an individual project so no need for other communication.

**Personal Contribution Statement:**

I have created a schedule of project milestones to keep me on track. There is a bit of slack built in to refine development stages or get caught up if problems arise. The timeline is on GitHub.

# References

*A System for Health Document Classification Using Machine Learning*. (n.d.). Retrieved from https://samphina.com.ng/system-health-document-classification-machine-learning/

BOYLE, T. (2018). *Medical Transcriptions*. Retrieved from kaggle: https://www.kaggle.com/datasets/tboyle10/medicaltranscriptions?select=mtsamples.csv

*MTSAMPLES.COM Home*. (2022, 2 1). Retrieved from MTSAMPLES.COM: https://mtsamples.com/index.asp

Quality, A. f. (2018, July). *The Distribution of the U.S. Primary Care Workforce* . Retrieved from Agency for Healthcare Research and Quality,: https://www.ahrq.gov/research/findings/factsheets/primary/pcwork3/inde

Shrutha Kashyap, S. M. (2018, 7 30). *Medical Document Classification*. Retrieved from INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT): https://www.ijert.org/medical-document-classification