**Title: Automated Classification of Medical Documents**

**Team: John Savage**

**Project Description:**

**Objective:**

**This project will classify textual medical documents that have 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 injecitons), or tests (e.g., imaging, stress tests, electromyelograms)**

**Usefulness: This classification application would eliminate need for medical office clerical workers to assign an class to medical documents arriving at the office for entry into the proper partition within the electronic medical record (EMR). This applications 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.**

**This automated classification would eliminate the need for office staff to assign a class to each document entered in the electronic medical record. The classification would also allow healthcare providers to prioritize the documents that they review based on the classification. 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.** [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. This project is different in that the documents will be classified for the purposes of expediting healthcare provider work flows. The target users are primary care providers which number approximately 295,000 in the US. 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, 2 hours weekly, $100/hr, 295,000 providers)

Data:

* Dataset is Medical Transcriptions found on the Kaggle website taken from mtsamples.csv which can be found elsewhere on the web. This data set has 4998 samples consisting of 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 to be used for labels for the purposes of this project. Therefore, the data will be the unlabelled transcription and unsupervised learning will be used to group the document in 3 classes. Labels will be created for validation and test data for purposes of evaluation and assigning classes. The data is transcription data so symbols, numbers, and punctation will need to be removed.

NLP functions: The main 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 could be utilized as the importance of test results is somewhat analogous to sentiment

.User interaction will consist of loading documents and sorting these documents by class. Retrieving the documents by class.

Communication and Sharing: This is an individual effort so there is no plan for communication.

* which varies considerably and appears to be the reason for the transcription

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| description | medical\_specialty | sample\_name | transcription | keywords |

**1. Project Title (5pts)**

**2. Team (Names)**

**3. Project Description (80pts)**

* Objectives 10pts: State as clearly as possible what you want to do. What problem do you solve and etc?
* Usefulness 10pts: State as clearly as possible why your chosen application is useful. Make sure to answer the following questions: Are there any similar or equivalent applications out here?  If so, what are they and how is yours different? Which user group/stakeholders is your application targeting?
* Data 30pts:
  + Describe dataset origin (who collected, when, and for what purpose). Do not use data without any information provided. Do not use text-preprocessed clean data. What is data format: csv, json ...
  + Provide initial data description.
    - Besides a text field, what other information do you have (for example, review data -> location, stars, users; twitter data -> user, likes;...)
    - do you have labeled data?
    - #records, #fields (columns if available), #NA values
    - What type of cleaning does it require? For example, do you need to customize stop words, symbols (e.g. keeping symbols for sentiment)..
* Functionalities 20pts: Describe tentatively what tasks your application will perform. There are two types of functions you would need to offer:
  + - **NLP Functions**: specific to your NLP tasks
    - **User interaction:**For example, allowing users to select/filter/search ...
  + Note - your application should be focused on NLP methods
* Communication and Sharing 10pts: Set up your preferred communication methods (zoom, teams, discord ...) and provide a github repository link for the project:
  + create a Read.me file with the initial project description
  + upload dataset

**4. Personal Contribution Statement (10pts)**

* team project: What was your contribution during the Planning stage (Part 1)
* individual project: Reflect on how you will manage time/tasks to complete milestones

**Writing (5pts)**

* Make sure it is free from spelling mistakes - consider it as your development proposal for your company
* Format your submission according to graduate school standard: APA style.

**Be responsible for your assigned tasks and team deadlines**. The final grades will be based on each person's contribution.

Note - do not copy and reuse existing apps and projects (from kaggle, github etc). Penalty will apply (see students handbook on academic dishonesty).

Shrutha Kashyap, Sushma M G, Varsha Rajaram, Vibha S, 2015, Medical Document Classification, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) NCRTS – 2015 (Volume 3 – Issue 27),

[Medical Document Classification – IJERT](https://www.ijert.org/medical-document-classification)

Classifies document as Respiratory, Digestive, Cardiology, Neurology and seems to be based on term frequency clustering classification

Abdullah Muhammad Alghoson from Claremont Graduate University has proposed a method to classify the medical documents [1]. The classification was done using predefined terms from Medical Subject Headings (MeSH). It used a corpus of 50 full-text journal articles (N=50) from MEDLINE, which were already indexed by experts based on MeSH. Using natural language processing (NLP), the algorithm classifies the collected articles under MeSH subject headings. The algorithms outcome was evaluated by measuring

its precision and recall of resulting subject headings from the algorithm, comparing results to the actual documents subject headings. The algorithm classified the articles correctly under 45% to 60% of the actual subject headings and got 40% to 53% of the total subject headings correct. This holds promising solutions for the global health arena to index and classify medical documents expeditiously.

Bodenreider[2], uses The Unified Medical Language System (UMLS) which contains semantic information about terms from various sources; each concept can be understood and located by its relationships to other concepts. According to Bodenreider, the UMLS concepts are used to classify condition terms in the Clinical Trials database into broad disease categories in the MeSH database in three major steps:

1. matching condition terms to the UMLS concepts using the exact match technique or normalization techniques such as inflection, punctuation, and case sensitivity.
2. Limiting the UMLS concepts to MeSH subject headings by passing through four consequent steps until the matching process succeeds. The four steps are: using MeSH term synonyms, choosing an associated expression as a translation, selecting a MeSH term using the MeSH hierarchy of concepts, and selecting the non-hierarchical related concept.
3. Assigning MeSH subject headings to the major categories of MeSH subject headings that represent the major disease categories using MeSH trees of hierarchy. The author uses "condition term" as an evaluation unit where the algorithm was applied to classify 12,612 condition terms in the Clinical Trials database, in which 1,823 terms were distinct. The evaluation of the algorithm outcomes was manually reviewed, and the evaluation metrics precision and recall were used. The results were astonishing: 96% of the 1,823 condition terms were successfully classified in MeSH.

Highland Healthcare

[Medical Records Classification | Healthcare | Hyland](https://www.hyland.com/en/healthcare/content-services/health-information-management/medical-records-classification)

Hyland Intelligent MedRecords utilizes Intelligent Classification technology and [automatically captures](https://www.hyland.com/en/platform/capabilities-and-features/capture/intelligent-capture-and-automated-data-extraction/data-capture), identifies and assigns document types – with the option to assign visit numbers and patient identifiers – and routes exceptions to staff for review. It improves the accuracy, speed and consistency of critical Health Information Management (HIM) tasks.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

[A System For Health Document Classification Using Machine Learning (samphina.com.ng)](https://samphina.com.ng/system-health-document-classification-machine-learning/)

© samphina.com.ng | All rights reserved  
  
[A System For Health Document Classification Using Machine Learning](https://samphina.com.ng/system-health-document-classification-machine-learning/)

The U.S. primary care workforce includes approximately 209,000 practicing primary care physicians, 56,000 nurse practitioners (NPs), and 30,000 physician assistants (PAs) practicing primary care, for a total of nearly**295,000 primary care professionals** ( Table 1 ).

[www.ahrq.gov/research/findings/factsheets/primary/pcwork3/index.html](http://www.ahrq.gov/research/findings/factsheets/primary/pcwork3/index.html)

https://www.bing.com/ck/a?!&&p=903e94b44df56545JmltdHM9MTY2NDkyODAwMCZpZ3VpZD0xZGJhYzU3Zi05ZDhiLTY5ZmItMmMzNS1jYTUyOWMyMzY4YjYmaW5zaWQ9NTQyMg&ptn=3&hsh=3&fclid=1dbac57f-9d8b-69fb-2c35-ca529c2368b6&psq=how+many+primary+care+providers+are+in+the+us&u=a1aHR0cHM6Ly93d3cuYWhycS5nb3YvcmVzZWFyY2gvZmluZGluZ3MvZmFjdHNoZWV0cy9wcmltYXJ5L3Bjd29yazMvaW5kZXguaHRtbCM6fjp0ZXh0PVRoZSUyMFUuUy4lMjBwcmltYXJ5JTIwY2FyZSUyMHdvcmtmb3JjZSUyMGluY2x1ZGVzJTIwYXBwcm94aW1hdGVseSUyMDIwOSUyQzAwMCwyOTUlMkMwMDAlMjBwcmltYXJ5JTIwY2FyZSUyMHByb2Zlc3Npb25hbHMlMjAlMjglMjBUYWJsZSUyMDElMjAlMjku&ntb=1