IBA SMCS Department of Computer Science

MS Project End-Semester Progress Report

Name of Student: Syed Bilal Rizwan

ERP: 23943

Title of Project: Navigating ICD-10 Coding Complexity: A Comparative Evaluation of Pre-trained Language Models and MEDCAT/SNOMED-CT for Clinical Note Assignments

Supervisor's Name: Dr. Sajjad Haider

(Digital) Signature of Supervisor:

Date: 07/01/2024

What are the core functionalities of the product/solution that you have developed?

The project aims to do a comparative study between pretrained language models and MEDCAT/SNOMED in mapping ICD-10 codes from clinical notes. The major features achieved in this project are mentioned below:

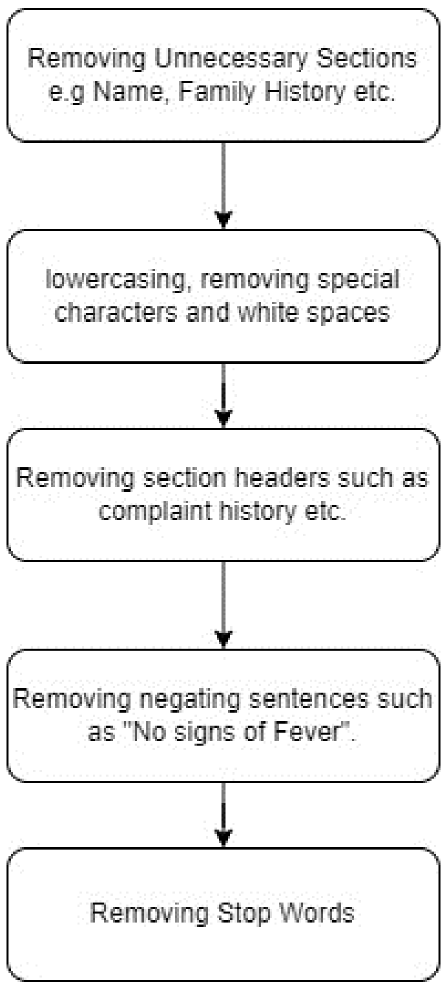
- **In depth Analysis**: A thorough analysis has been performed using a benchmark dataset MIMIC IV. Only 1000 shortest notes were used as a testing set to compute metrics.
- **Extensive Preprocessing**: The project devised a preprocessing technique that preprocesses the dirty clinical note itself by removing unnecessary sections, stop-words, and advanced natural language techniques to remove unnecessary information from the text and make it classification-ready.
- **Modern Methods**: The project tried different modern Transformer-based architectures such as pre-trained BERT to achieve the multi-label classification.
- **Handling long and dirty clinical texts**: The devised approaches can handle varying sizes of input clinical text since each healthcare professional may have different writing styles and each patient may have longer or shorter record.
- **Multi-label Classification:** The project aimed to test methodologies that can assign multiple ICD-10 codes to a single clinical note.

Illustrate your final design methodology diagram and describe each component/module in detail:

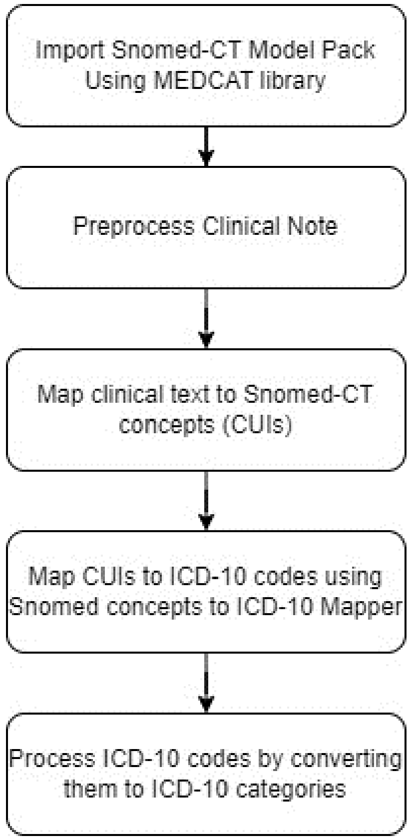
Since this was a research-based project, there was no final design that was developed as a solution. However, there were several methodologies tested using pretrained language models against MEDCAT library to assess the efficacy of each method. The methodologies are explained in simple flowcharts below starting with preprocessing techniques:

1. Preprocessing

The preprocessing of this project was a crucial step because free text can vary widely, and there are fixed preprocessing techniques for all clinical text types, with additional techniques focused on clinical text from the MIMIC IV dataset. Let's review a few steps taken to ensure correct preprocessing:

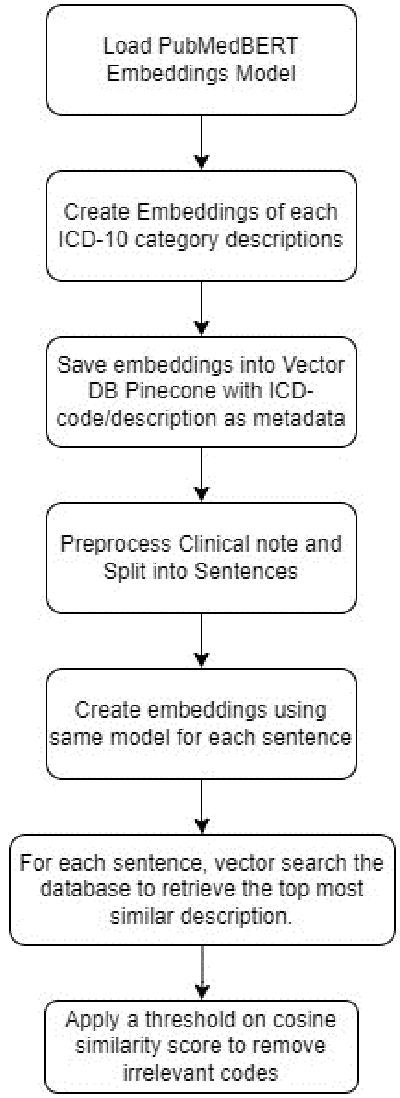


2. Using MEDCAT and SNOMED-CT



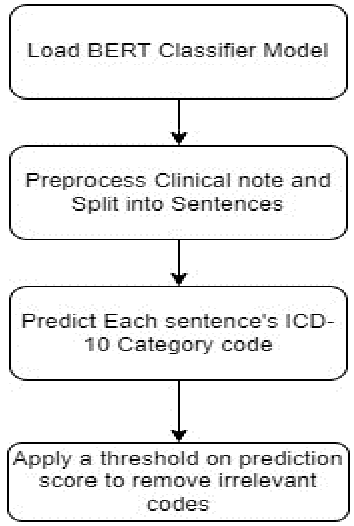
In this phase, the MEDCAT library and its SNOMED-CT model pack were imported and utilized to map processed clinical notes to ICD-10 categories.

3. Using Embedding Model (PubMedBERT)



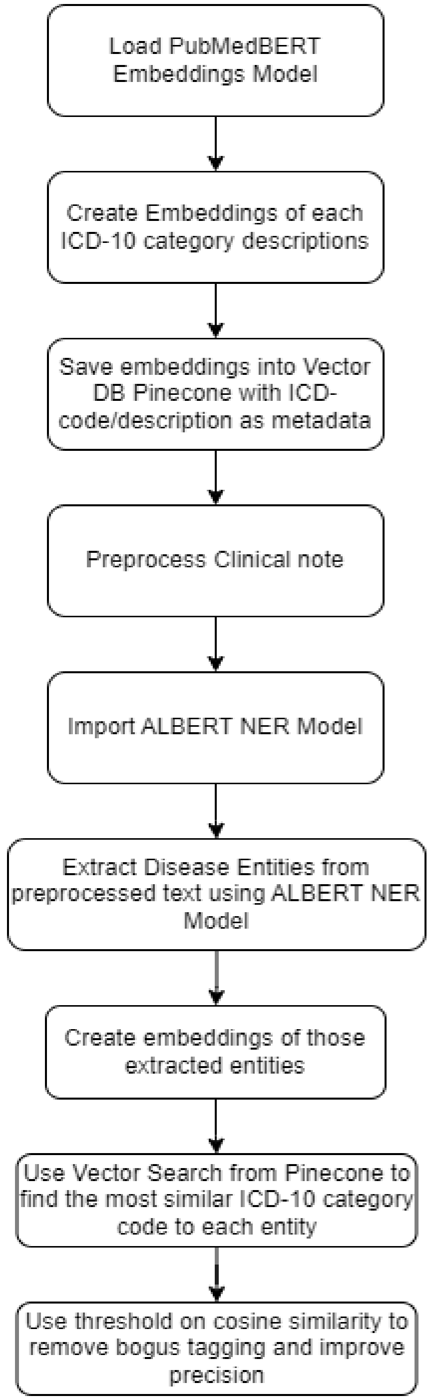
This approach involved creating embeddings for each ICD-10 category description using a pre-trained BERT model on the MIMIC dataset, specifically PubMedBERT (Yu Gu and Co, 2020 [10]). The processed clinical note was segmented into sentences, and the embedding for each sentence was generated using the same model. A vector search identified the top ICD-10 category description, which was then labeled accordingly. Thresholding was applied to the similarity score in the final step to enhance the precision of the overall process.

4. Using Pretrained BERT Classifier (Emran's)



This approach involved using pre-trained BERT classifiers to multi-label classify ICD-10 categories for clinical notes. The model from the research paper by Emran and colleagues (Al-Bashabsheh and co [2]) was imported from Hugging Face and applied to the processed clinical note, which had been divided into sentences. Predicted categories from the BERT classifier model underwent thresholding on scores to improve the precision of the overall process.

5. Using AlBERT NER tagger with PubMedBERT embedding model.



The last phase of experiment involved a unique approach to extract disease entities using a transformer-based NER tagger named albert-medical-ner-proj from huggingface website. Meanwhile, category description embeddings are already saved in pinecone vector DB. After extracting entities, embeddings of those entities are created using PubMedBERT embedding model and a vector search is used to find the closest ICD-10 category for each entity. Lastly, a thresholding is applied on cosine similarity to weed out bogus matches to improve precision.

Implementation Details: [mention as a bulleted list]

- Programming Language: Python (3.10)
- Programming Software: Jupyter Notebook for Experimentation/VSCode

There are several libraries/APIs explored and experimented with in this project:

Dataset and Preprocessing:

- Pandas == 1.5.3
- NumPy
- Regex == 2.2.1
- Spacy
- Sci-kit Learn.
- Matplotlib
- NLTK ==3.7

SNOMED Methodology:

- MedCAT
- SNOMED-CT Model Pack
- Sci-Spacy

GPT Methodology:

- OpenAI (GPT 3.5)
- Pinecone (Vector DB)
- Cohere (Text Embeddings)

LLM Methodology:

- Transformers == 4.21.3
- Hugging Face Custom Pre-Trained Models (BERT, PubMedBERT, AlBERT)
- PLM Model

Datasets Used:

- MIMIC III
- MIMIC IV

Computer-Specifications Used:

- Processor: 11th Gen Intel(R) Core (TM) i7-11800H

- Ram: 32GB
- GPU: NVIDIA GEFORCE RTX 3050Ti (4GB, not used)

The implementation time taken for this project was 5 months based on extensive experimentation, methodology development and methodology creation.

Demo of Results:

Demo Link:

Potential impact of your product in the industry/society:

There was no demo given to any industry during the project however, a demo is planned to be given to a large hospital in Karachi, Pakistan and their feedback will be taken for further enhancements and improvements in the methodologies. The major impact of this project was to explore different pretrained language models and compare their performance with MEDCAT and it was seen that MEDCAT's Snomed-CT model pack trumps all other advanced methodologies in terms of metrics. Furthermore, the research also shows that it is possible to do a multi-label classification on clinical texts to extract multiple ICD-codes from the same text of large sizes.

Upload code to GitHub:

Zip file link:

IBA SMCS Department of Computer Science

References (if applicable, make a bulleted list):

[1] Nguyen, Anthony N., Donna L. Truran, Madonna Kemp, Bevan Koopman, David Conlan, John O'Dwyer, Ming Zhang, Sarvnaz Karimi, Hamed Hassanzadeh, Michael Lawley and Damian J. Green. “Computer-Assisted Diagnostic Coding: Effectiveness of an NLP-based approach using SNOMED CT to ICD-10 mappings.” AMIA ... Annual Symposium proceedings. AMIA Symposium 2018 (2018): 807-816.

[2] Al-Bashabsheh, Emran, Ahmad Alaiad, Mahmoud Al-Ayyoub, Othman Beni-Yonis, Raed Abu Zitar, and Laith Abualigah. "Improving clinical documentation: automatic inference of ICD-10 codes from patient notes using BERT model." The Journal of Supercomputing (2023): 1-25.

[3] Amin, Saadullah, Günter Neumann, Katherine Dunfield, Anna Vechkaeva, Kathryn Annette Chapman, and Morgan Kelly Wixted. "MLT-DFKI at CLEF eHealth 2019: Multi-label Classification of ICD-10 Codes with BERT." In CLEF (Working Notes), pp. 1-15. 2019.

[4] Huang, Chao-Wei, Shang-Chi Tsai, and Yun-Nung Chen. "PLM-ICD: automatic ICD coding with pretrained language models." arXiv preprint arXiv:2207.05289 (2022).

[5] “International Classification of Diseases.” Encyclopædia Britannica. Accessed December 29, 2023. https://www.britannica.com/topic/International-Classification-of-Diseases.

[6] Ahmed, Joinal. “Automatic ICD-10 Code Assignment to Consultations Using Deep Learning.” Halodoc Blog, June 29, 2023. https://blogs.halodoc.io/automatic-icd-10-code-assignment-to-consultations-using-deep-learning/.

[7] Kraljevic, Zeljko, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio et al. "multi-domain clinical natural language processing with medcat: the medical concept annotation toolkit." Artificial intelligence in medicine 117 (2021): 102083.

[8] Zeljko. “MedCAT: Introduction - Analyzing Electronic Health Records.” Medium, April 23, 2020. https://medium.com/@w_is_h/medcat-introduction-analyzing-electronic-health-records-e1c420afa13a.

[9] Johnson, Alistair, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. "Mimic-iv." PhysioNet. Available online at: https://physionet. org/content/mimiciv/1.0/ (accessed August 23, 2021) (2020)

[10] Gu, Yu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. "Domain-specific language model pretraining for biomedical natural language processing." ACM Transactions on Computing for Healthcare (HEALTH) 3, no. 1 (2021): 1-23.