NLP Pipeline with Unstructured Data: Design Document

Ayoub Benchaita
abenchaita3@gatech.edu

1 PROJECT DESIGN

1.1 Project Summary

The purpose of this project is turn unstructured clinical data into a meaningful encoding. The project will focus on ingesting and preprocessing clinical notes, then applying an NLP transformation layer to predict ICD-10 diagnosis code for patients from each clinical note. The final web application will provide an interface to upload either a singular clinical note in a text field or upload a CSV file of multiple clinical notes. The output will be either the predicted ICD-10 diagnosis code or a CSV file of all the predictions in the case of multiple patient clinical notes were provided.

1.2 Tools and Technology

- Python
- Pandas
    - To assist with data ingestion and preprocessing
- Natural Language Toolkit
    - For tokenization, parsing, chunking
- TensorFlow
- Flask
- HTML

1.3 Data Sources

The data that will be used to create the NLP pipeline is from MIMIC (Medical Information Mart for Intensive Care) database. The MIMIC is a large database of deidentified health-related data from patients admitted to Beth Israel Deaconess Medical Center. The specific dataset that I intend to use is the MIMIC-III dataset, which contains data from 2001-2012. The table within the MIMIC-III dataset that will be useful in this project is the noteevents table. This table contains notes for patients, organized by category. The text field within the table contains the clinical notes that will be used to create and test the NLP pipeline.

1.4 Diagrams

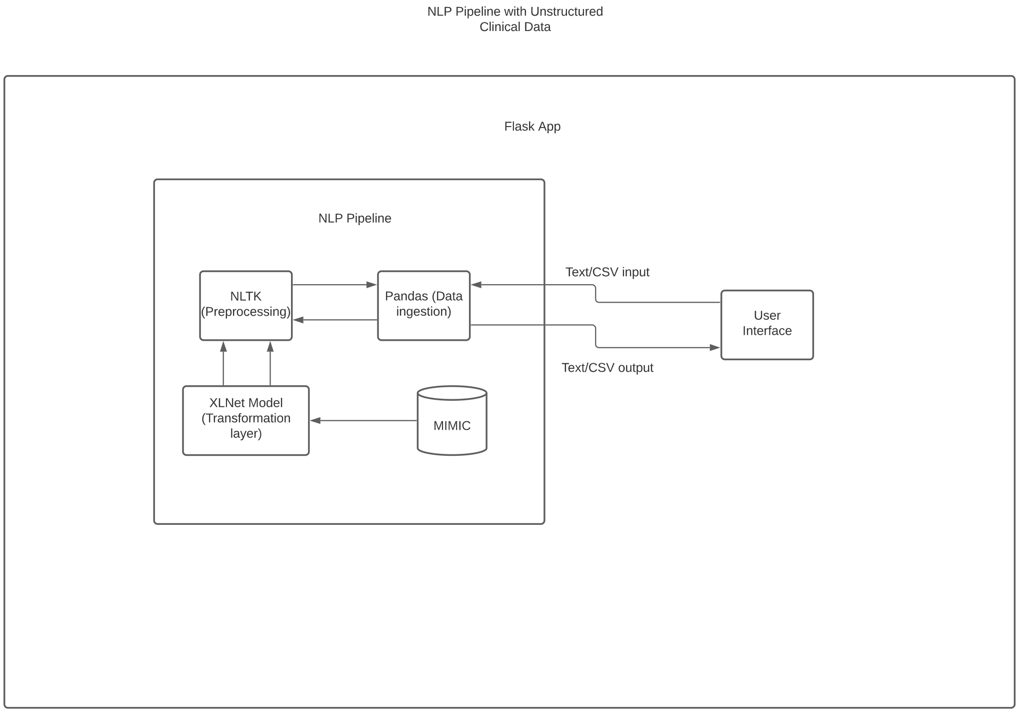
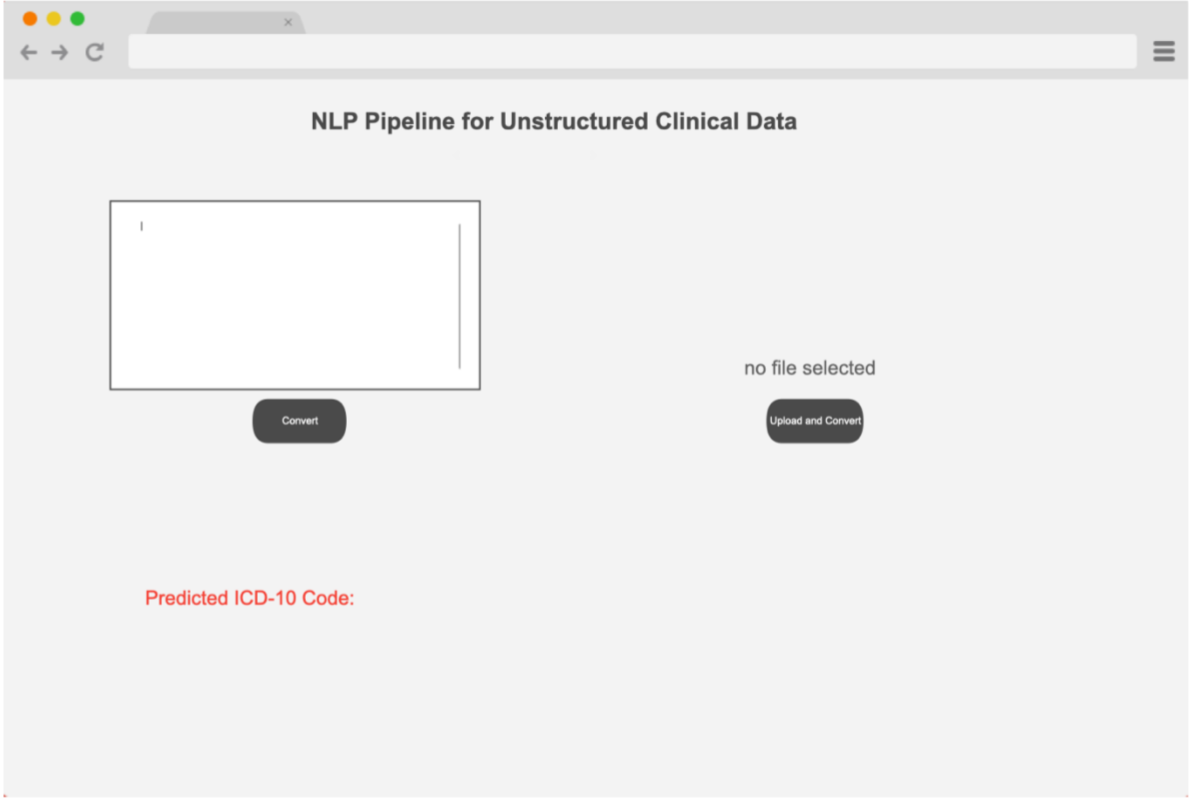


Figure 1—Simple Architecture Diagram

1.5 Screen Mock-ups



NLP Pipeline for Unstructured Clinical Data

Figure 2—Screen Mock-up for User Interface

2 IMPLEMENTATION PLAN

2.1 Project Tasks

1. Complete training and request access for dataset (Sprint 3)
2. Build the user interface (Sprint 4)
3. Create data ingestion and preprocessing module (Sprint 5)
4. Research/Decide best NLP model(s) for given data (Sprint 6)
5. Build NLP transformation layer to predict code (Sprint 7 & 8)
6. Return correct output to user, whether text or CSV file (Sprint 9)
7. Project presentation (Sprint 10)

2.2 Project Timeline

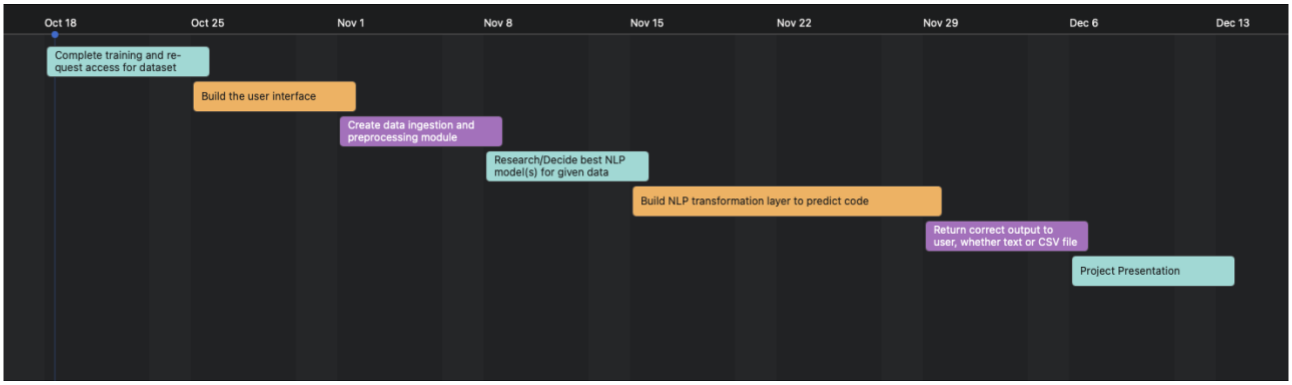


Figure 3 — Project timeline Gantt Chart

2.3 Needs/Risks

The main risk I may run into in this project is needing to find a new dataset to work with if my access request is denied. From my research most of the people that take the mandatory course and submit a certificate of completion receive access. I tried to cope with this risk by having the first task after requesting access be independent of the dataset (creating the user interface).

I have not used Flask and I am not fully aware if it is the best technology to use given the scope of my project.

Another potential risk I may run into is customizing my data ingestion and preprocessing part of my pipeline too much for my dataset. The pipeline should be built in a way where any clinical note can be consumed if it is in text format. It may be a good idea to use another dataset to prevent this from happening.

3 REFERENCES

8. Goyal, K. (2021, January 11). *Top 7 python NLP libraries [and their applications in 2021]*. upGrad blog. Retrieved from https://www.upgrad.com/blog/python-nlp-libraries-and-applications/.
9. Johnson, A., Pollard, T., & Mark, R. (2019, April 24). *Mimic-III clinical database demo*. MIMIC-III Clinical Database Demo v1.4. Retrieved from https://physionet.org/content/mimiciii-demo/1.4/.
10. *Mimic documentation*. MIMIC. (n.d.). Retrieved from https://mimic.mit.edu/docs/.