MoM (Review Meeting 4)

Date and Time: **10th April 2024, 8 PM to 9.30 PM IST.**

**Meeting Attendees**:

* David de Hilster, Hugo Watanuki, Dr Shobha G, Dr Jyothi Shetty, Dr. Vishalakshi Prabhu, Nihar Mandahas, Skanda P R, Pratheek Rao, Manvith L B, Arya Hariharan, Eshan Mathur, Prashant Ronad, Nikhil Vasu.

**Agenda 1**: Present the findings and work done in the past few weeks related to the Legal Assistance Tool and understand future steps.

*Discussion/Demonstration Done:*

1. The working of the web interface with updated features which included the feature to selectively add or delete necessary queries on the user’s discretion.
2. Issues regarding inconsistent query outputs, incomplete case statement output and case sensitivity of keywords searched.
3. The issue of having a static dataset and the possibility of updating it with new cases using the Indian Kanoon API.

*Recommendations:*

1. Pass an index object rather than a Dataset object while querying to try and fix errors.
2. Dataset should ideally be standardized and the casing should be fixed to allow for predicted working in scenarios with negligible difference (e.g. when the keywords are the same but the letter case may be different).
3. Proceed with a static dataset, however continue searching for a free source for database updates.
4. Clearly document the cause of inability to currently update the database with new cases, including sources discovered, reasons for rejection of use and other supplementary details.
5. Consider consulting other sources like the HPCC Stack Overflow in case of errors or doubt.

*Actions to be Taken:*

1. Investigate on the methods suggested to fix the querying errors.
2. Standardize the database letter casing (i.e. lower-case or upper-case) to prevent ambiguity.
3. Document the cause for using a static dataset.

**Agenda 2**: Present the work related to “Building NLP pipeline for Electronic Health Records” and understand the future steps.

*Discussion/Demonstration Done:*

1. Tokenization and preprocessing done on entire dataset. Preprocessed in such a way that age, gender, body part, duration of symptom are labelled.
2. Two models namely Comprehend and BILSTM-CRF are tried extensively on PubMed dataset (1800 patient records so far). During March, tried on 250 records and issues rectified.
3. There is an issue in manual labelling. Same model namely **biomedical\_ner\_all** is used for labelling in both Comprehend and BILSTM-CRF
4. In Comprehend model, extract character embedding, pass it through LSTM encoder, get word embeddings. BIOBERT is used for extracting embeddings. Tested on NVIDIA A400 GPUs in college. Model not working fine on non-zero labels which is actual target. Working fine on zero labels
5. In BILSTM-CRF, get word embedding and concatenate. In CRF, joint probability of all embeddings is computed. When used 150 epochs, loss is reduced. BILSTM-CRF works better than comprehend in terms of accuracy.
6. For MRI, Nikhil has tried inception model on OASIS dataset. There were 1311 images of 4 classes and 89% accuracy. He used transfer learning models (VGG16, inception model). Used regularization to minimize overfitting.

*Recommendations:*

1. Hugo had to drop in between this meeting to attend another meeting. There are no new instructions.
2. Students want to spray the data on HPCC cluster.
3. David said to leverage other channels in case Roxie is not working. Take support from Roxie assigned person.
4. David asked who is the end user and what are the interface types?
5. David said, “Continue the good work exploring NLP models.”

*Actions to be Taken:*

1. Train the two models (Comprehend and BILSTM-CRF) on more dataset and check performance (accuracy, loss)
2. Finetune both NLP models for EHR.
3. Minimize overfitting, train on other MRI’s.
4. GPU ran out of memory when tried for more than 1500 patient records.