Mar 3rd, 2024

Approaches tried from last interaction:

1. Wiki BM25 Lexical Model + Logistic Regression
2. Wiki BM25 Lexical Model + Probabilistic Linear Discriminant Analysis (PLDA)
3. Glove + 1D Convolution + Softmax
4. Glove + Fully Connected Layer
5. T5 Encoder + PLDA

Motivation behind each approach:

1. **Wiki BM25 Lexical Model Approach:** Search engines employ the ranking function in the domain of information retrieval, known as Okapi BM25, or best matching (BM) abbreviated, to determine the relevancy of content to a particular search query. It is based on the probabilistic retrieval paradigm. As legal documents are concerned with the legal phrases used, these phrases are the keywords that act as deciding factors for a sentence being as- signed to a particular class for classification. In TF- IDF (term frequency-inverse document frequency) method, TF is a measure of how often a phrase appears in a document, and IDF is about how important that phrase is. Hence, I tried out the improved TF-IDF method, with refinements of TF and IDF components using the BM25 function.
2. **Glove:** GloVe stands for Global Vectors for word representation. The Stanford University academics that created it wanted to create word embeddings by combining global word co-occurrence matrices from a specific corpus. The GloVe word embedding's fundamental premise is to infer the link between words using statistical data. Glove Embedding requires very little preprocessing and fine-tuning. It does not take into consideration the context of the text. So this could be inferior as compared to the transformer-based approaches. But this is a lightweight approach, hence, to find out, how much f1 gain occurs when trying a simpler model as compared to a heavyweight model, this approach was considered.

3. **T5:** T5 or Text-to-Text Transfer Transformer, is a Transformer-based architecture that employs a text-to-text method. The model text is fed as input into each task, such as translation, question answering, and classification, and trained to produce some target text. This enables us to reuse the same model, loss function, hyper parameters, and so on across our wide range of workloads. This reason to choose this model is the extensive pre-training data (C4 corpus) with a strong foundation for handling diverse classification problems, potentially requiring less fine-tuning compared to smaller models.

Results:

Approach	**Macro F1**	**Precision**
Wiki BM25 Lexical Model + Logistic Regression	0.38	0.51
Wiki BM25 Lexical Model + Probabilistic Linear Discriminant Analysis (PLDA)	0.33	0.38
Glove + 1D Convolution + Softmax	0.31	0.35
Glove + Fully Connected Layer	0.26	0.31
T5 Encoder + PLDA	0.43	0.5

Analysis: In the earlier 3 weeks, the older statistical models were okay but not great. Now, we are starting to see improvement in metrics as we move towards transformers based models. I hope we will see more improvement in the next 3 weeks.

Approaches for next 3 weeks:

1. T5 Encoder + FAISS (Facebook AI Similarity Search)
2. T5 Base + LSTM
3. paraphrase_xlm_r_multilingual_v1+FAISS
4. paraphrase_xlm_r_multilingual_v1+PLDA