MED264_FinalProjReport_Group3.pdf

Submission Date: 12/07/23
Submission ID:
File name: MED264_FinalProjReport_Group3.pdf
Word Count: 2479 (Total) 226 (Abstract)
Character Count: 20306

Advancing Interpretability and Transparency of Large Language Models in Biomedical Informatics: A Focus on Fine-tuned ClinicalBERT for Predicting 30-Day Mortality

MED 264 Final Project Report Fall 2023

Abstract:

This report addresses the research problem of enhancing the interpretability and transparency of Large Language Models (LLMs) in Biomedical Informatics (BMI), particularly focusing on the prediction of 30-day mortality. The primary objective was to adapt the ClinicalBERT model to predict mortality with high accuracy and also to make the prediction process understandable for clinicians.

For this purpose, the MIMIC-III Dataset was utilized. A preprocessing approach was employed to format and clean the text data, concentrating on early clinical notes within 24 hours of admission. The ClinicalBERT model was fine-tuned and evaluated using this preprocessed data. The model exhibited good accuracy levels (0.76 for training, 0.67 for validation and 0.70 for testing), indicating its robustness in mortality prediction. However, when assessing the model's interpretability using Mean Reciprocal Rank (MRR), Precision @ K, and Mean Average Precision (MAP), the results were mixed. While the model showed effectiveness in identifying relevant terms (MAP score of 0.832), it fell short in aligning its top predictions with the terms identified as most critical by clinicians (Precision @ K score of 0.034).

Our project demonstrates the potential of transformer-based models for predictive tasks. Our findings also highlight the need for further advancements in model interpretability, crucial for their practical application in clinical settings. This research contributes to the evolving landscape of LLM applications in healthcare, offering insights for future developments in the field.

Introduction:

With the advent of Transformer architecture and its applications, Large Language Models (LLMs) have emerged as powerful tools across various domains, including text, image, video understanding, and more (Vaswani et al., 2017; Arnab et al., 2021; Devlin et al., 2019; Dosovitskiy et al., 2021; Jumper et al., 2021). In BMI, the significance of textual data from clinical notes in electronic health records (EHR) has been underscored by several systematic reviews (Ford et al., 2016; Jensen et al., 2012; Pivovarov & Elhadad, 2015), paving the way for the application of LLMs in this field.

Recently, BMI researchers have been focusing on adapting LLMs for biomedical contexts, leading to the development of models like clinicalBERT and BioMedRoBERTa (Gururangan et al., 2020; Huang et al., 2020). Unlike in other fields, BMI demands that these models not only perform exceptionally but also maintain high levels of explainability to earn the trust of clinicians and patients. Traditional approaches in NLP and LLMs, while effective in performance, often fall short in terms of explainability (Choi et al., 2016; Huang et al., 2020).

Although LLMs are inherently complex, recent advancements have made strides toward demystifying these models. Techniques like gradient-based, attention-based, and local relevance-based methods have been developed to enhance their transparency (Chefer et al., 2021; Hao et al., 2021; Jain & Wallace, 2019; Yosinski et al., 2015).

Our project, recognizing the novel nature of Transformer architectures in the BMI field, aims to contribute to this evolving landscape by enhancing the interpretability and transparency of a fine-tuned clinicalBERT model. By applying a local relevance-based transformer interpretability method, our project seeks to provide insights for clinicians in predicting 30-day mortality. This work represents an initial step towards understanding and leveraging LLMs in clinical settings, potentially offering suggestions for how clinicians assess and manage patient care.

Methodology:

Data Set:

Our data source is the MIMIC-III Clinical Database (version 1.4), a comprehensive repository featuring deidentified patient data, including laboratory results, medication data, and vital signs, along with clinical notes (Johnson et al., 2016). This database is highly relevant to our study's objective due to its extensive range of clinical notes, which are crucial for understanding patterns associated with 30-day mortality. We focused on the "TEXT" field in the "NOTEEVENTS" and linked it with the "ADMISSION" CSV file, where we kept the patients who both have admission information and at least one clinical note record within 24 hours of admission. We excluded lower-quality data and the discharge summaries to prevent overfitting, as they contain death information for each patient. All the text files went through preprocessing to transform the text into a consistent format for further analysis. Preprocessing of the clinical notes follows the workflow provided on the ClinicalBERT Github Page and we adjusted their code to fit our cases (Huang et al., 2020).

Preprocessing:

Our analysis began with the preprocessing of the MIMIC-III dataset to prepare it for the application of the ClinicalBERT model. We imported the ADMISSIONS and NOTEEVENTS datasets using the Python Pandas library, ensuring the date and time fields (ADMITTIME, DISCHTIME, and DEATHTIME) were formatted consistently. We then merged the admissions data (df_adm) with the notes data (df_notes), focusing on key identifiers, and kept the patients who both have admission information. The dataset underwent cleaning, where entries marked as errors were removed (n = 869), and discharge summaries were excluded. We defined 30-day mortality status by subtracting patients' death time from patients' admission time. Patients who died within 30 days are labeled as "1," while the rest is labeled as "0". For patients who don't have a death event occur (no death time), we assigned them to be 30-day mortality status = 0 since they didn't die within 30 days after their admission. Finally, there are 4852 patients who died within 30 days and 47,702 who didn't in our final analysis.

To emphasize early clinical indicators, we extracted notes from the first 24 hours of admission, concatenating them to form complete textual data for each patient. Our preprocessing steps included the removal of de-identified brackets, normalization of abbreviations, and special characters. The text was segmented into 318-word parts, aligning with the ClinicalBERT standard for detailed textual analysis. The dataset was then divided into training (80%), validation (10%), and testing (10%) sets, ensuring a balanced representation of patient outcomes. Rigorous checks were performed to prevent any overlap between training and testing datasets. Finally, these preprocessed sets were exported as CSV files, laying the groundwork for the subsequent training and evaluation phases using the ClinicalBERT model. Given the initial imbalance in our dataset, we applied downsampling to the training, validation, and test subsets to achieve a 50/50 ratio of patients who died within 30 days to patients who did not die within 30 days) cases. This approach was essential to prevent the model from achieving deceptively good accuracy by predominantly predicting the more frequent non-death outcome.

Fine-tuning Pre-trained Model (ClinicalBert):

In the fine-tuning stage of our project, the ClinicalBERT model was adapted to predict 30-day mortality, beginning with the setup of our computational environment. The choice of ClinicalBERT was based on its compatibility with our dataset and its proven effectiveness in similar tasks. We employed PyTorch for computational optimization, utilizing the NVIDIA RTX A6000 GPU to enhance processing efficiency. When GPU resources were unavailable, CPU-based computations were executed. This foundational setup enabled us to load and record sentence counts from the training, validation, and testing datasets into Pandas dataframes.

The adaptation of the ClinicalBERT model commenced with the application of the BERT tokenizer from the Hugging Face transformers library, chosen for its compatibility with the 'ClinicalBERT' model. This tokenizer played a pivotal role in converting textual data into a format suitable for BERT. To improve efficiency, we utilized pre-existing encoded data when available and processed new data as needed. Due to the variability in input sequence lengths, we standardized all sequences to a uniform length of 512 tokens through padding or truncation, and generated attention masks for each sequence to enable the model to differentiate between relevant data and padding.

For data preparation, we converted the inputs, labels, and masks from our datasets into torch tensors, ensuring compatibility with the BERT model framework. DataLoaders were configured to process the data in batches of 64, optimizing the training and evaluation efficiency.

The model was augmented with an additional classification layer for binary classification. Training involved using the AdamW optimizer with specific learning rate and epsilon settings, alongside a learning rate scheduler. This phase included dataset training, backpropagation, optimization, and continuous performance evaluation.

The model's training accuracy was 0.76, with validation and testing accuracies at 0.67 and 0.70, at the second epoch. These results highlight the model's effectiveness and stability in predicting 30-day mortality. Post-training, the model was validated with a separate dataset to

ascertain its generalizability. For future analysis, each epoch's model was saved, and loss values were recorded.

Evaluation of the model

The interpretability of our fine-tuned BERT model was achieved using a local relevance-based method, employing Deep Taylor Decomposition to assign relevance scores to tokens in relation to the prediction (Chefer et al., 2021). Each entry in the dataset was analyzed to determine the importance of individual tokens in the context of 30-day mortality prediction.

To benchmark our model against human clinical reasoning, we utilized GPT-3's capabilities in simulating a clinical environment, drawing inspiration from its performance in USMLE Soft Skill evaluations (Brin et al., 2023; Brown et al., 2020). In this simulation, GPT-3 functioned as a clinician, tasked with identifying key terms in medical notes indicative of patient mortality risk. The top 10 terms identified by GPT-3 for each patient note formed our ground truth for comparing the interpretability results of our model. We employed three metrics for this assessment: Mean Reciprocal Rank (MRR), Precision @ K, and Mean Average Precision (MAP), each offering insights into the model's effectiveness in mirroring clinical decision-making processes.

Mean Reciprocal Rank (MRR):

We utilize MRR to measure the rank effectiveness of our interpretability model. MRR is calculated by taking the average of the reciprocal ranks of the first relevant term from the ground truth (GPT-3 list) in our prediction list. For each ground truth term, we find its first occurrence in the prediction list and calculate its reciprocal rank (1/rank). If the term doesn't appear in the prediction list, its reciprocal rank is 0. The MRR is the mean of these values. This metric helps us understand at what position the first relevant term (as identified by GPT-3) appears in our model's prediction list. If the first relevant term from the GPT-3 list is found early in our prediction list, this implies a higher MRR and indicates a stronger alignment with the GPT-3 model's reasoning.

Precision @ K:

Precision @ K assesses the proportion of relevant terms identified by our model in the top K predictions. Precision @ K is calculated by dividing the number of relevant terms found in the top K predictions of our model by K. In our case, K is 10, so we look at the top 10 terms predicted by our model for each patient note and count how many of these are in the ground truth list provided by GPT-3. This metric provides insight into how many of the top 10 terms identified by our model are actually considered important by the GPT-3 clinician model. A higher Precision @ K score signifies greater accuracy of our model in pinpointing critical terms related to mortality risk.

Mean Average Precision (MAP):

MAP is used to evaluate the average precision across all ranks of the prediction list. MAP is the mean of the average precision scores for each query. Average Precision (AP) for a single query is calculated by taking the mean of the precision scores at each rank where a relevant term (as per GPT-3) is found. It considers the entire ranked prediction list and not just the top K. This metric is crucial for understanding the overall effectiveness of our model in identifying mortality-linked terms throughout its entire set of predictions. A higher MAP score indicates that our model not only identifies relevant terms but also ranks them effectively, closely mirroring the ground truth provided by GPT-3.

Results:

The results of the evaluation using Mean Reciprocal Rank (MRR), Precision @ K, and Mean Average Precision (MAP) provide a comprehensive insight into the performance of our interpretability model in the clinical setting. The MRR score of 0.449 suggests that, on average, the first relevant term identified by the GPT-3 model appears relatively high in our model's prediction list, indicating a moderate level of alignment in prioritizing key terms. However, the Precision @ K score of 0.034 shows that only a small fraction of the top 10 terms predicted by our model align with the terms identified by the GPT-3 clinician model, pointing to a gap in identifying the most critical terms. In contrast, the high MAP score of 0.832 indicates that across the entire set of predictions, our model is quite effective in identifying and ranking relevant terms. This suggests that while our model may not always align with the GPT-3 model in terms of the most critical terms, it is generally successful in identifying relevant terms throughout its predictions. These results imply that our interpretability model demonstrates a significant degree of effectiveness in mimicking clinical reasoning, with specific areas for improvement in precision at the top ranks.

Discussion:

The evaluation of the ClinicalBERT model using Mean Reciprocal Rank (MRR), Precision @ K, and Mean Average Precision (MAP) provided insightful outcomes. The MRR score of 0.449 indicates a moderate alignment with the initial relevant terms as identified by the GPT-3 model, suggesting the model's capability in identifying pertinent terms. However, the low Precision @ K score of 0.034 highlights a significant gap in the model's ability to prioritize the most critical terms in line with clinical judgment. This divergence is crucial in the context of clinical decision-making, where the accuracy of prioritizing vital information is imperative.

In the process of data preprocessing, the focus was placed on the clinical notes within the first 24 hours of admission, and discharge summaries were excluded to maintain the integrity of the data and prevent overfitting. These methodological decisions were essential to ensure the reliability of the study's outcomes and the interpretability of the model. The choices reflect the critical balance between data comprehensiveness and quality, a common consideration in machine learning applications within healthcare.

The study's findings prompt further questions regarding the optimization of machine learning models to mirror the complex nature of clinical reasoning. The disparity observed between the model's overall ability to identify relevant terms and its precision in ranking the most critical terms necessitates further investigation and refinement.

A limitation of this study is its reliance on a singular dataset and the exclusion of certain clinical note types. This may impact the generalizability of the findings across different clinical contexts. Future research should aim to enhance the model's accuracy in prioritizing terms and consider integrating diverse clinical data sources. Investigating the impact of various types of clinical notes on the model's predictions could also yield valuable insights into its broader applicability in medical practice.

Conclusion:

Our research with the ClinicalBERT model has successfully showcased its capability to predict 30-day mortality from clinical notes. The model achieved good accuracy levels (0.76 for training, and 0.67 for validation and 0.70 for testing), underscoring its robustness and potential applicability in clinical settings. These results were derived from rigorous training and evaluation processes using a comprehensive dataset, which provided a realistic simulation of the model's performance in a real-world clinical environment.

However, the interpretability metrics, particularly the Precision @ K score, indicated a discrepancy between the model's identification of relevant terms and its alignment with clinical priorities. This finding was drawn from comparing the model's predictions against a clinician-like benchmark set by the GPT-3 model. It suggests that while the model is effective in recognizing pertinent terms within the clinical notes, it may not always prioritize them in a manner consistent with clinical judgment.

To address these shortcomings, future work should concentrate on refining the model's precision in predicting top-tier terms. This could entail the integration of more sophisticated interpretability methods or further optimization of the current model to more closely reflect clinical reasoning processes. Additionally, diversifying the dataset to encompass a wider array of clinical notes and scenarios would likely enhance the model's generalizability and relevance across different clinical contexts.

Our study was inherently limited by the nature of the dataset used and the exclusion of certain types of clinical notes, which might have impacted the depth and breadth of our analysis. A more nuanced approach to processing and interpreting clinical notes could offer deeper insights into the complex dynamics of clinical data.

Code Availability

The code is publicly available at the following Github link. The fine-tuned BERT model can be downloaded from the following Google Drive link.

References:

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., & Schmid, C. (2021). *ViViT: A Video Vision Transformer* (arXiv:2103.15691). arXiv. https://doi.org/10.48550/arXiv.2103.15691

Brin, D., Sorin, V., Vaid, A., Soroush, A., Glicksberg, B. S., Charney, A. W., Nadkarni, G., & Klang, E. (2023). Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. *Scientific Reports*, *13*(1), Article 1. https://doi.org/10.1038/s41598-023-43436-9

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language Models are Few-Shot Learners* (arXiv:2005.14165). arXiv. https://doi.org/10.48550/arXiv.2005.14165

Chefer, H., Gur, S., & Wolf, L. (2021). *Transformer Interpretability Beyond Attention Visualization* (arXiv:2012.09838). arXiv. https://doi.org/10.48550/arXiv.2012.09838

Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., & Sun, J. (2016). *Doctor AI: Predicting Clinical Events via Recurrent Neural Networks* (arXiv:1511.05942). arXiv. https://doi.org/10.48550/arXiv.1511.05942

Damhuis, R. A. M., Wijnhoven, B. P. L., Plaisier, P. W., Kirkels, W. J., Kranse, R., & van Lanschot, J. J. (2012). Comparison of 30-day, 90-day and in-hospital postoperative mortality for eight different cancer types. *British Journal of Surgery*, *99*(8), 1149–1154. https://doi.org/10.1002/bjs.8813

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. https://doi.org/10.48550/arXiv.1810.04805

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale* (arXiv:2010.11929). arXiv. https://doi.org/10.48550/arXiv.2010.11929

Ford, E., Carroll, J. A., Smith, H. E., Scott, D., & Cassell, J. A. (2016). Extracting information from the text of electronic medical records to improve case detection: A systematic review. *Journal of the American Medical Informatics Association: JAMIA*, *23*(5), 1007–1015. https://doi.org/10.1093/jamia/ocv180

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). *Don't Stop Pretraining: Adapt Language Models to Domains and Tasks* (arXiv:2004.10964). arXiv. https://doi.org/10.48550/arXiv.2004.10964

Hao, Y., Dong, L., Wei, F., & Xu, K. (2021). *Self-Attention Attribution: Interpreting Information Interactions Inside Transformer* (arXiv:2004.11207). arXiv. https://doi.org/10.48550/arXiv.2004.11207

Huang, K., Altosaar, J., & Ranganath, R. (2020). *ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission* (arXiv:1904.05342). arXiv. https://doi.org/10.48550/arXiv.1904.05342

Jain, S., & Wallace, B. C. (2019). *Attention is not Explanation* (arXiv:1902.10186). arXiv. https://doi.org/10.48550/arXiv.1902.10186

Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: Towards better research applications and clinical care. *Nature Reviews Genetics*, *13*(6), Article 6. https://doi.org/10.1038/nrg3208

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, *3*(1), Article 1. https://doi.org/10.1038/sdata.2016.35

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), Article 7873. https://doi.org/10.1038/s41586-021-03819-2

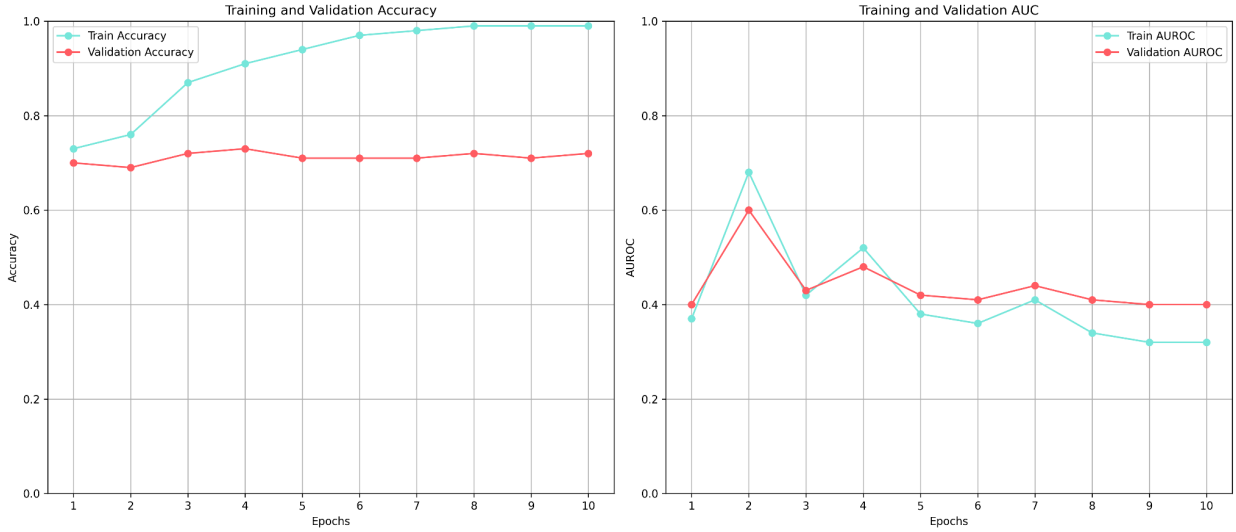
Pivovarov, R., & Elhadad, N. (2015). Automated methods for the summarization of electronic health records. *Journal of the American Medical Informatics Association*, *22*(5), 938–947. https://doi.org/10.1093/jamia/ocv032

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need* (arXiv:1706.03762). arXiv. https://doi.org/10.48550/arXiv.1706.03762

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., & Lipson, H. (2015). *Understanding Neural Networks Through Deep Visualization* (arXiv:1506.06579). arXiv. https://doi.org/10.48550/arXiv.1506.06579

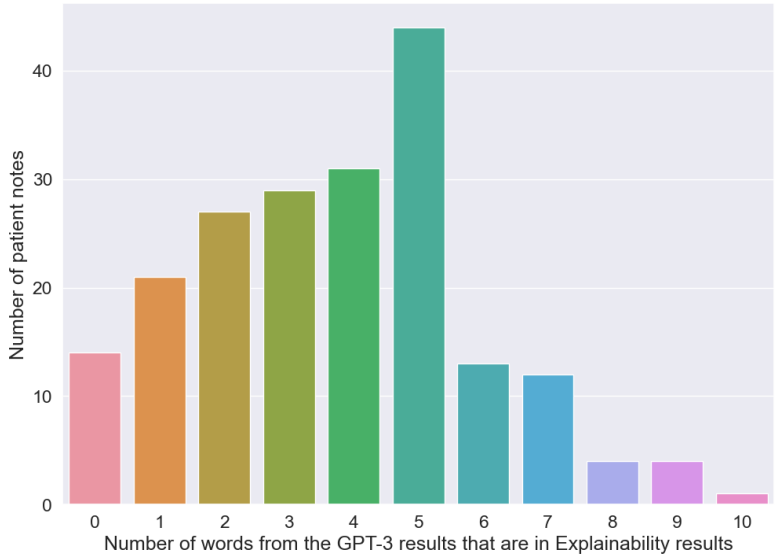
Figures and Tables:

Figure 1: Evaluation matrices of model for each training epoch.



This graph represents the accuracy and AUROC scores over the number of epochs.

Figure 2: Interpretability



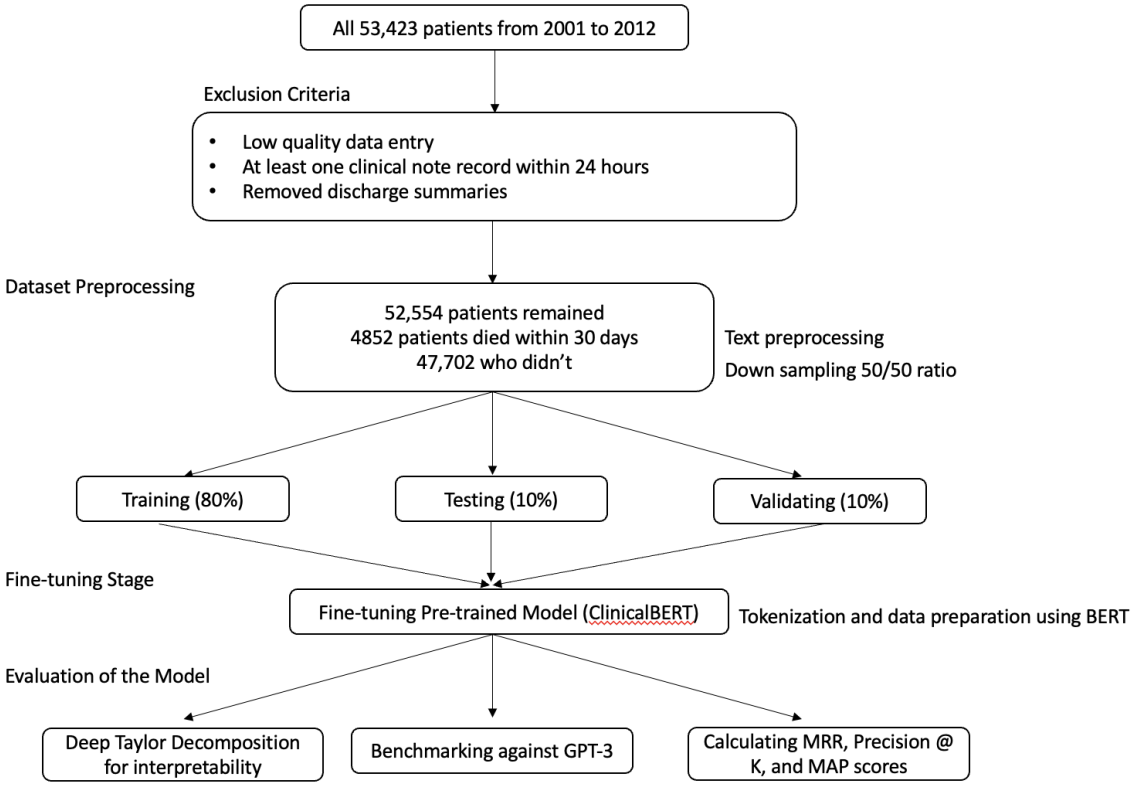
Shows a histogram plot based on the number of words from GPT-3 that are detected by the interpretability model.

Table 1: Results from the 3 Metrics

Metric	Score
Mean Reciprocal Rank	0.449
Precision @ K	0.034
Mean Average Precision	0.832

This table shows the score for each metric

Figure 3: Workflow Diagram



This diagram represents our workflow