MEDICAL NOTES

Classifying and Grading Medical Examinations using Transformers

Kamaljeet Ghotra, Nicholas Brathwaite, Dylan Jin

SLIDESMANIA.COM

Importance of Medical Notes

- 80% of diagnoses are made on medical history alone.[1]
- < 10 minutes for history, examination and explanation
- Success depends on asking the right questions and interpreting answers

SLIDESMANIA.COM

[1] Summerton N; The medical history as a diagnostic technology. Br J Gen Pract. 2008 Apr58(549):273-6. doi: 10.3399/bjgp08X279779.

USMLE℠

Step 2:
Clinical Skills Exam

SLIDESMANIA.COM

Step 2 - Clinical Skills Cancelled

- COVID-19 pandemic
- High cost to examinee
- Costly to grade
- Manual process

SLIDESMANIA.COM

Opportunities and Challenges

- Automated grading
    - Reduce costs and increase availability
    - Potentially reduce (or propagate) bias

- Matching expressions
    - Medical terminology and jargon
    - Combination of symptoms

- People
    - Inconsistent shorthand
    - Spelling mistakes

SLIDESMANIA.COM

[1] Summerton N; The medical history as a diagnostic technology. Br J Gen Pract. 2008 Apr58(549):273-6. doi: 10.3399/bjgp08X279779.

Data

Clinical Cases	10 unique “patients”
Patient Notes	42,146 submissions
Features	9 to 18 features per case (14.3 average)
Training Data	1000 annotations

SLIDESMANIA.COM

Methodology

Labeling	In order to improve the accuracy of finding patient notes and annotations that contained targeted features we used offset mapping and text extraction techniques to create labels for our models
Scoring	Similar to labeling, we used distance similarity functions on encoded text entries to calculate a score amongst annotations, patient history, and our targeted text. This helped provide a threshold for our decoding model for each case and patient observed
Pre-Trained Models	In hopes of improving our predictive and accuracy results we chose to use pre-trained models that have been trained with medical vocabulary like Roberta, Deberta, ChatGPT, BioBert
EDA	To simplify the encoding, labeling, and scoring we combined the separate csv files into 1 dataframe using the unique case and patient numbers provided

SLIDESMANIA.COM

01

Model 1: Bio-BERT

Bio-BERT: Bio-BERT is another Bidirectional transformer trained on thousands of text entries from PubMed. It’s basis is the standard BERT model and was made for use in medical notations and more.

- Pre-trained medical model capable of both encoding and decoding architectures
- Additional attention masking to discern importance of each sequence.
- Pre-trained weights for prediction and classification

Running the Bio-BERT model required converting the text sequence and pre-processed labels into tensor datasets for training, validation, and testing. Next, we created the model based on the input ids, token types, and attention masks from the dataset to find the features within the patient notes.

Learning Rate	Batch Size	Epochs	Dropout Rate	Val Loss
0.01	50	2	0.3	0.619
0.001	32	2	0.1	0.622

SLIDESMANIA.COM

02

Model 2: Biomed-Roberta

Biomed-Roberta: The Biomed-Roberta model uses the input ids and attention masks from its tokenizer as inputs to the model for feature detection.

- An altered self attention mask focused on the dynamic change of medical text.
- Trained using over 7 billion tokenized inputs from 46GB of data.
- Removed next sentence prediction with intent of higher classification and recognition performances.

The Biomed-Roberta model like the BERT model before required pre-processing of the medical text, however the tokenization output was input ids for the relevant text and dynamic self attention masks.

Learning Rate	Batch Size	Epochs	Dropout Rate	Val Loss
0.01	32	2	0.3	0.6215

SLIDESMANIA.COM

03

Model 3: Roberta

RoBERTa (Robustly Optimized Bert) has following differentiation as compared to Bert Models

- More **training data (16G vs 160G)**
- Uses **dynamic masking pattern** instead of static masking pattern.
- Replacing the **next sentence prediction** objective with full sentences without NSP.
- Training on **Longer Sequences.**

The model was fine-tuned for text classification using the labeled fragments. We added a layer on top of the Roberta-base model to output binary labels for input_ids and offset mapping.

Learning Rate	Batch Size	F1
9.00E-05	8	0.533
1.00E-05	4	0.579
5.00E-05	4	0.547
1.00E-05	8	0.720

SLIDESMANIA.COM

04

Model 4: DeBERTa

DeBERTa (Decoding-enhanced BERT) with disentangled attention that improves the BERT and RoBERTa models using two techniques:

- **Disentangled attention mechanism,** where each word is represented using two vectors that encode its content and position and the attention weights among words are computed using disentangled matrices on their contents and relative positions
- An **enhanced mask decoder** is used to incorporate absolute positions in the decoding layer to predict the masked tokens in model pre-training. In addition, a new virtual adversarial training method is used for fine-tuning to improve models' generalization.

DeBERTa-v3 model was chosen as the base model for this project. It was fine-tuned on the data using the MLM technique. The pre-trained model was then used for downstream tasks.

Batch Size	Learning Rate	Epochs	Dropout Rate	Val Loss	F1 Score
8	5e-4	5	0.1	0.0816	0.001
4	9e-4	3	0.1	0.005	0.021
4	1e-5	10	0.1	0.094	0.001

SLIDESMANIA.COM

05

MODEL 5: GPT-3

GPT-3 (Generative Pre-trained Transformer)

- Pre-trained neural network using the transformer architecture
- Autoregressive transformer decoder model.
  - Unidirectional
- Generates contextually relevant responses to various prompts

SLIDESMANIA.COM

05

MODEL 5: GPT-3

Methodology

- Traditional
  - Parse text and count included features

- Generative
  - Generate an “ideal” patient note
  - Compare with exam submission

```
['lack of other thyroid symptoms', 'anxious or nervous', 'stress due to caring for elderly parents', 'heavy caffeine use',
'no depressed mood', 'weight stable', 'insomnia', 'female', 'decreased appetite', '45 year']
```

Patient is a 45 year old female who presents with insomnia, decreased appetite, lack of other thyroid symptoms, no depressed mood, and weight stable. Patient reports feeling anxious or nervous and states that stress due to caring for elderly parents is a contributing factor. Patient also reports heavy caffeine use.

SLIDESMANIA.COM

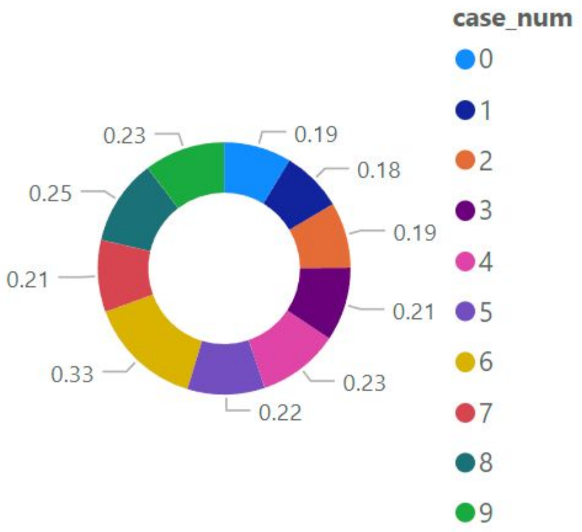
05

MODEL 5: GPT-3

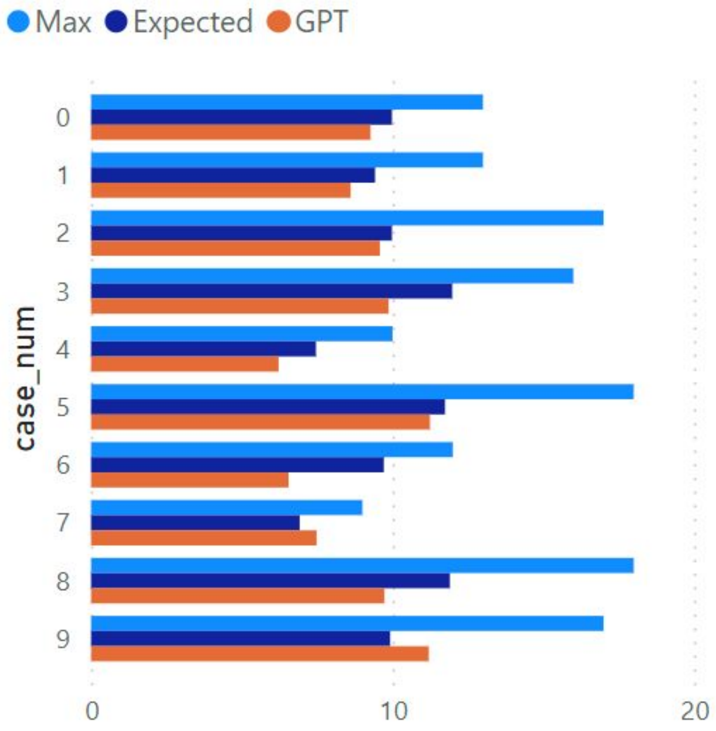
Davinci

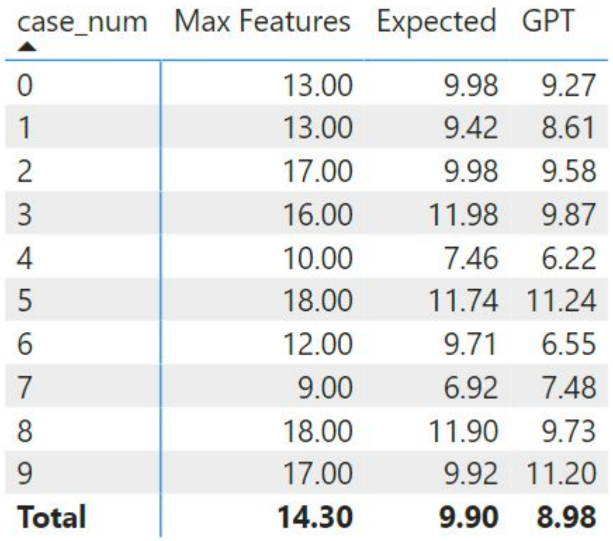
- Baseline
  - Average error: 0.36
- After tuning
  - Average error: 0.22

Average Feature Errors



Average Feature Counts



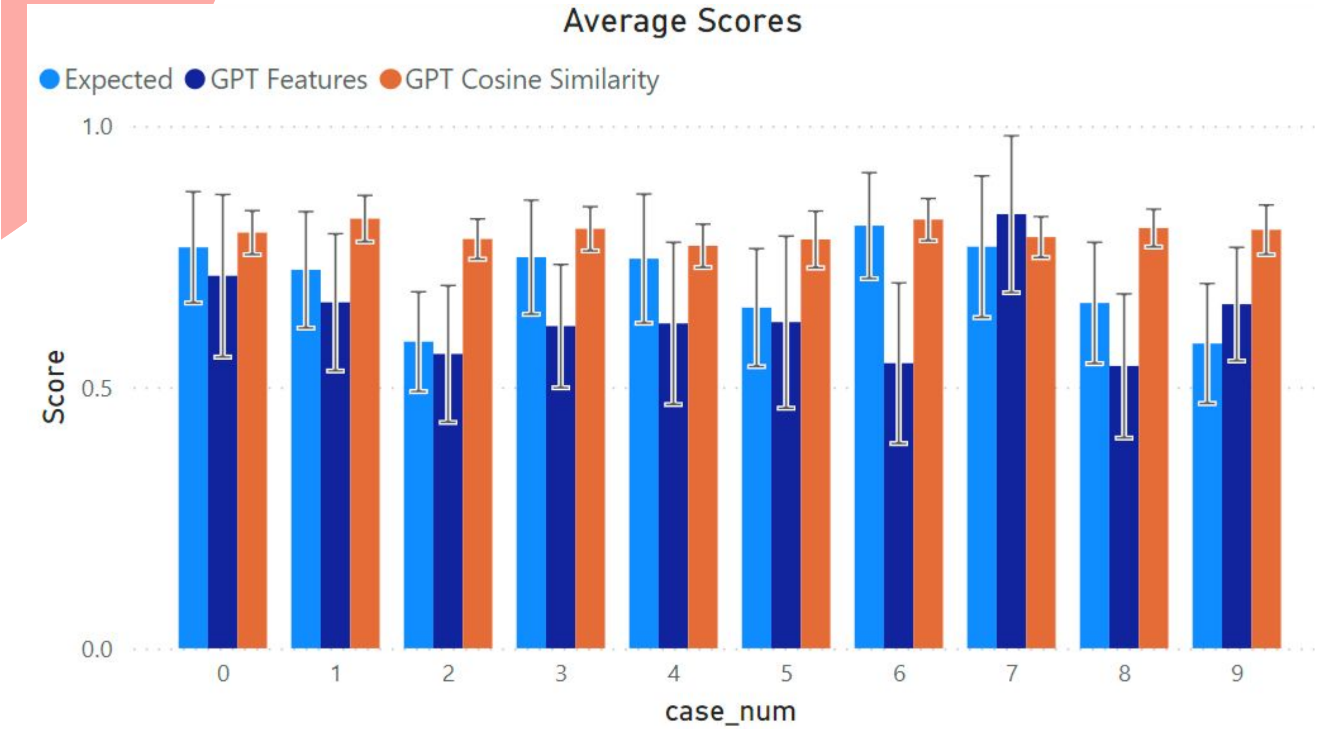


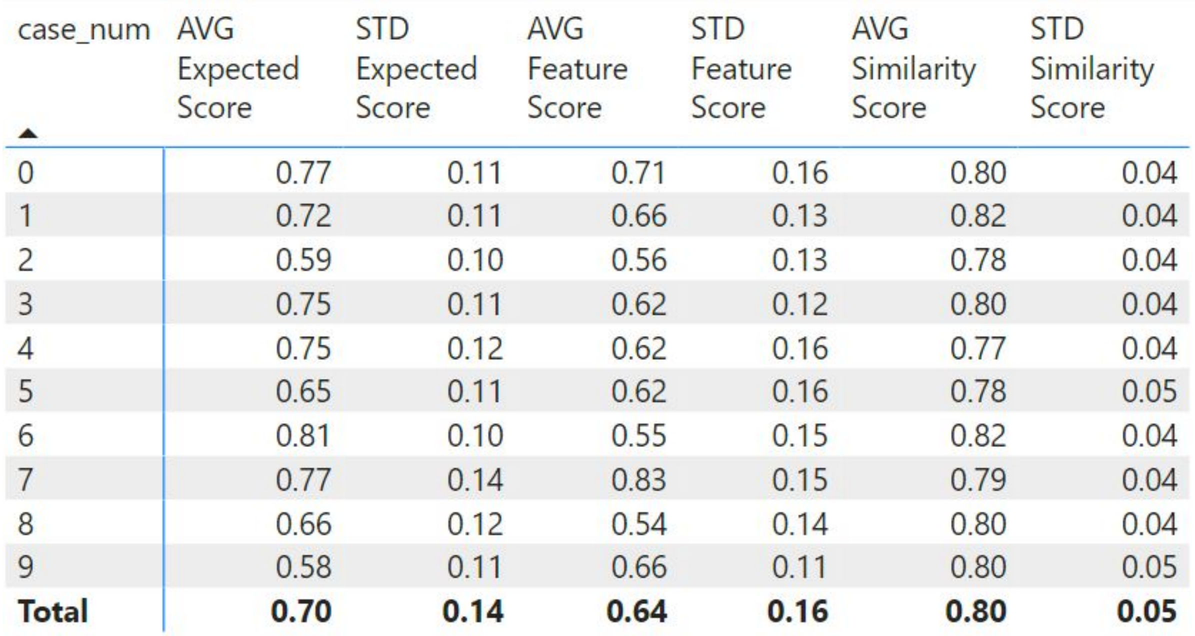
case_num ▲	Max Features	Expected	GPT
0	13.00	9.98	9.27
1	13.00	9.42	8.61
2	17.00	9.98	9.58
3	16.00	11.98	9.87
4	10.00	7.46	6.22
5	18.00	11.74	11.24
6	12.00	9.71	6.55
7	9.00	6.92	7.48
8	18.00	11.90	9.73
9	17.00	9.92	11.20
Total	**14.30**	**9.90**	**8.98**

SLIDESMANIA.COM

05

Model 5: GPT-3





case_num	AVG Expected Score	STD Expected Score	AVG Feature Score	STD Feature Score	AVG Similarity Score	STD Similarity Score
0	0.77	0.11	0.71	0.16	0.80	0.04
1	0.72	0.11	0.66	0.13	0.82	0.04
2	0.59	0.10	0.56	0.13	0.78	0.04
3	0.75	0.11	0.62	0.12	0.80	0.04
4	0.75	0.12	0.62	0.16	0.77	0.04
5	0.65	0.11	0.62	0.16	0.78	0.05
6	0.81	0.10	0.55	0.15	0.82	0.04
7	0.77	0.14	0.83	0.15	0.79	0.04
8	0.66	0.12	0.54	0.14	0.80	0.04
9	0.58	0.11	0.66	0.11	0.80	0.05
Total	**0.70**	**0.14**	**0.64**	**0.16**	**0.80**	**0.05**

SLIDESMANIA.COM

conclusion

Medical Scoring

Further development may be practical for future exam scoring

Patient Review

Feature detection can help patients with irregular conditions and tracking family medical history

Medical Diagnosis

Predictions using patient notes and annotations can help doctors with their diagnosis of patients

BERT

- BioMed-Roberta
- Bio-BERT

GPT

- GPT-3.5 DaVinci

Additional Work

- Two-tower
- ScaNN

SLIDESMANIA.COM

THank YOU!

Do you have any questions?

SLIDESMANIA.COM