Patient-Provider Communication Classification

Yingxue Pan (Introduction, Conclusion)
Heyuan Huang (Baseline Machine Learning Models)
Ji Qi (Improvement)
Gaoqianxue Liu (BERT models)

Table of Contents

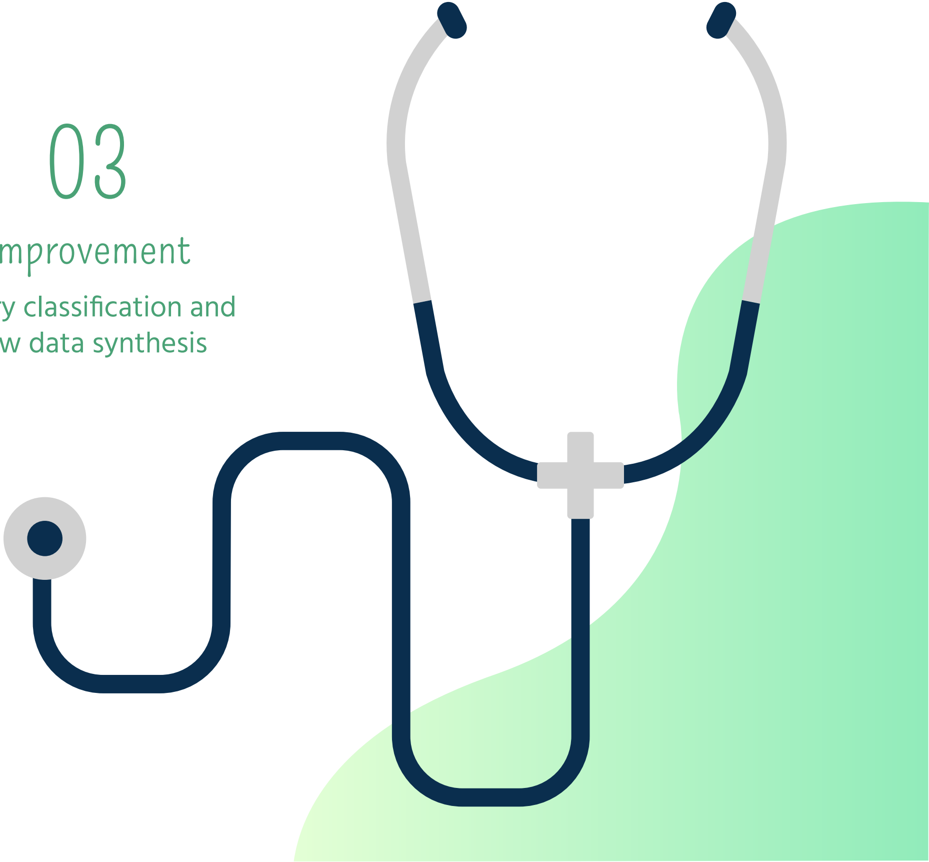
01
Introduction
Importance of our topic and data exploration

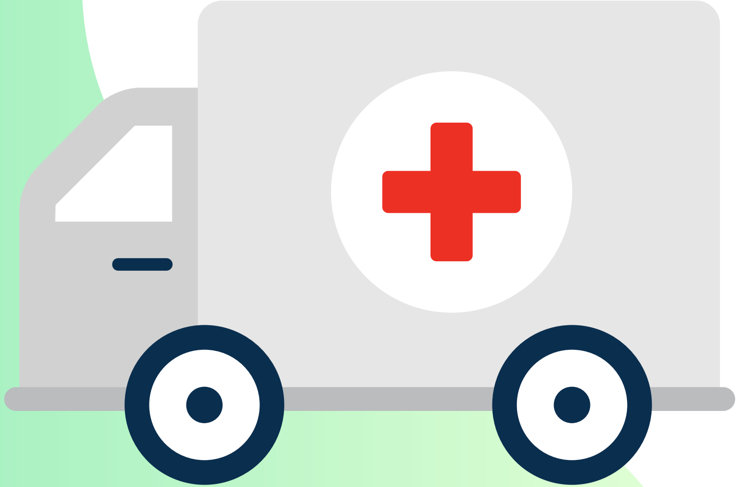
02
Baseline
Machine Learning models

03
Improvement
Binary classification and new data synthesis

04
BERT models

05
Conclusion & Discussion





01

Introduction

Importance of our topic and data exploration

Introduction

Definition

Patient-Provider Communication (PPC): analysis and interpretation of language used by patients and healthcare providers in the context of a clinical encounter

Why Classify Them

Effective and rapid PPC facilitates patient adherence monitoring and timely reaction to change in health, and as a result it improves patient-centered outcomes.

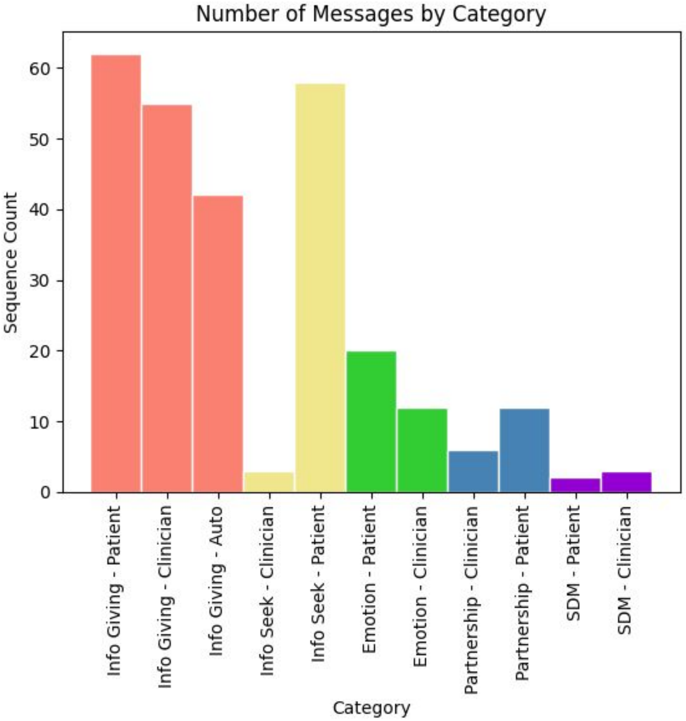
Manual Codes

By function: Information giving, Information seeking, Emotion, Partnership, Shared Decision-making (SDM)

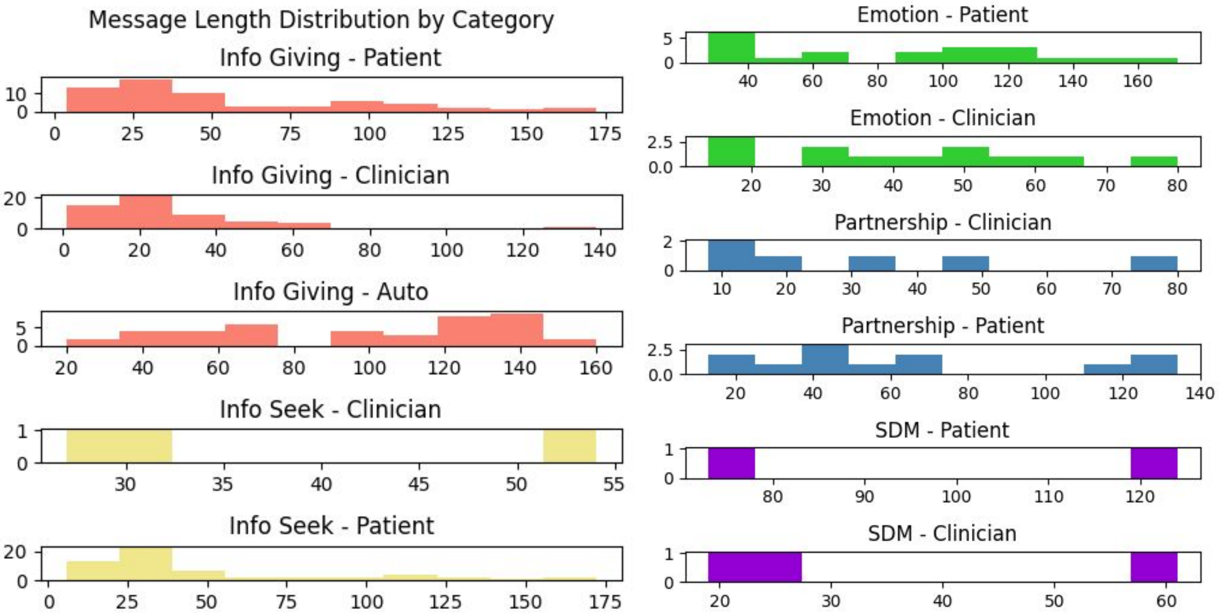
By sender: Clinician, Patient, Auto

Data Visualization

Message Count



Message Length

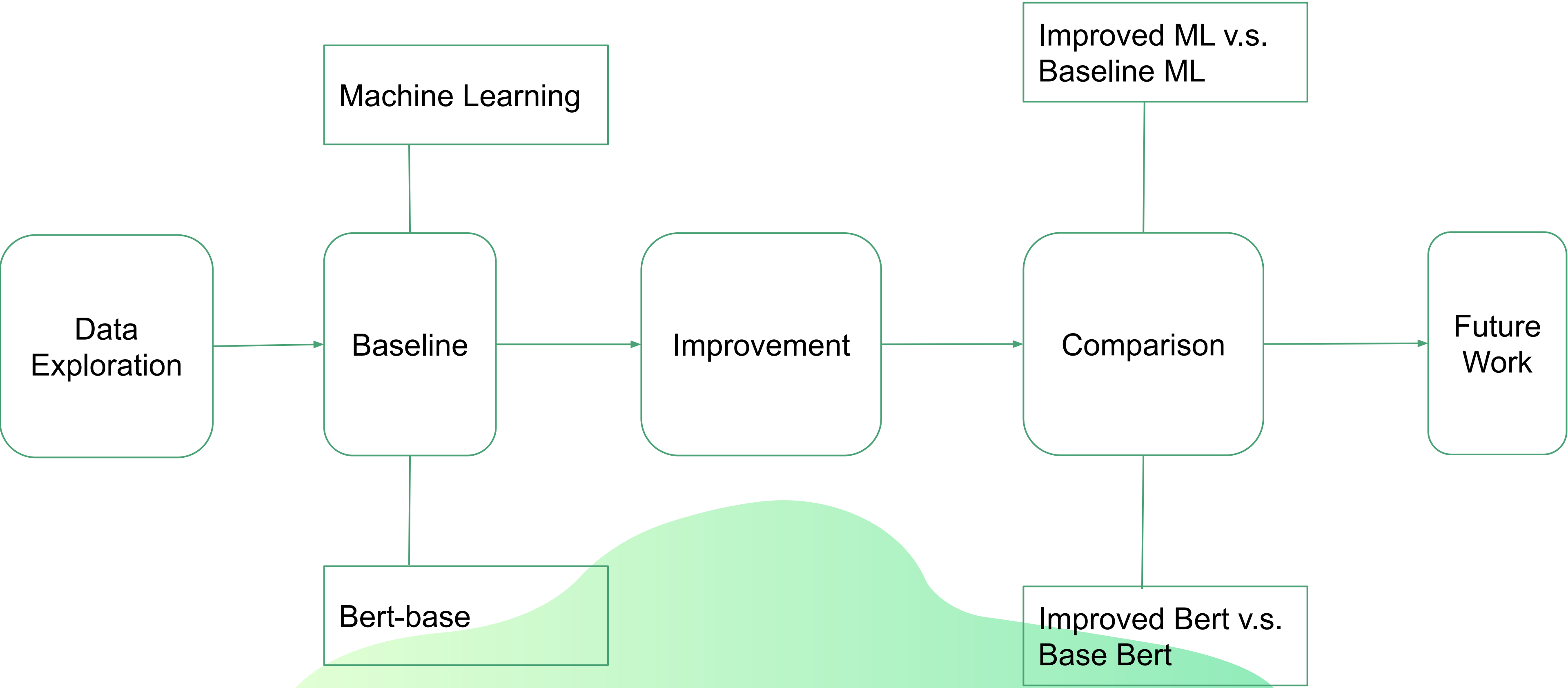


Instances of Original Data

	Clinician	Patient
Info-giving	Hi I just sent a copy via mychart If you didn't get it you can pick up a hard copy at the front desk on NP 8 Have a great weekend and say Hi to your wife	Dr I noticed I have not been scheduled for any blood work before my next visit May 15 Do I not need to have it done
Info-seeking	Hi are they planning on using port for procedure if so it will be flushed then-please let me know what works best for you His port was last flushed on xxx	I spoke to Today is my Mom's last dose of Tarceva is tonight Please let me know if she going to continue at this dose
Emotion	Excellent-hope you feel better-take care we are here if you need us for anything	Hi Can you please have Dr call me on my cell phone We are quite anxious to speak with him relative to my Dad's increase Thank you :)
Partnership	Mr I'm afraid we are still waiting for the insurance authorization Will keep you posted	Dr has asked if it is ok for me to stop the Xeloda 5 days prior to the ct myelogram that they are trying to schedule If this is ok I'll confirm with dr's staff Thank you
SDM	Sounds good- I just put in prescription for-numbingsoothing medication as well Do n't need to pick it up if you feel better with Pepto	Dr I have not scheduled the colonoscopy as yet because in light of some additional spots that have been seen from my recent CTScan and PETScan I'm using time off to resolve these issues Can we possibly worry about the colonoscopy once these other potential issues are resolved I'm meeting with Dr this afternoon to discuss my next steps in treatments of the new spots that were found in my lungs Thank you

Info-giving Auto	Appointment Information: Visit Type: Phone Consult Date: xxx2020 Dept: at Smilow Fairfield Provider: Neal A xxxx Time: 10:45 AM Appt Status: Scheduled

Workflow



Machine Learning

Improved ML v.s. Baseline ML

Data Exploration

Baseline

Improvement

Comparison

Future Work

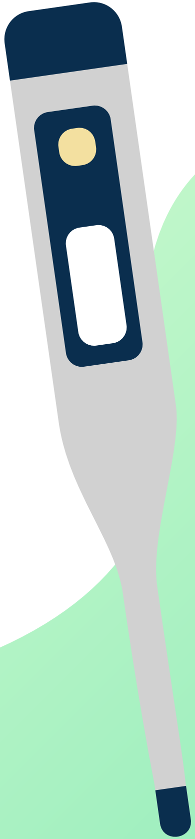
Bert-base

Improved Bert v.s. Base Bert

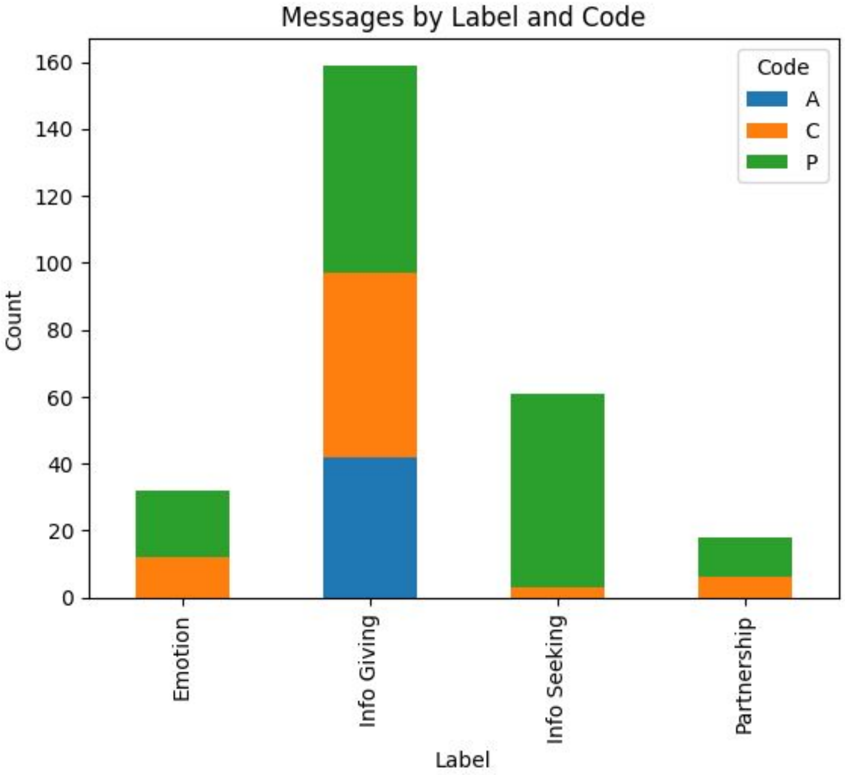
02

Baseline

Machine learning models



Text Data Preprocessing



Original Data: 5 classes including Information Giving, Information Seeking, Emotional Support, Partnership, and Shared Decision-Making(=5)

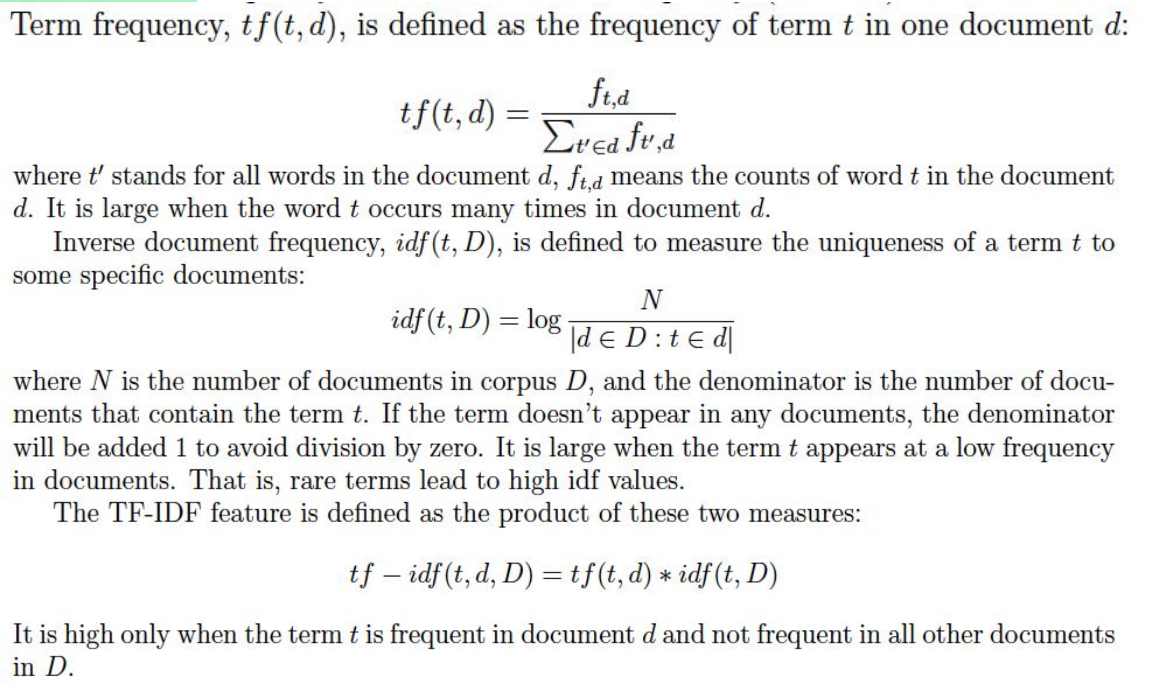
Processed Data:4 classes, 270 messages in total, vocabulary size = 1537

Code A stands for system Automatic message, C stands for Clinician and P stands for Patient.

NLTK.stem.WordNetLemmatizer() to restore different inflected forms of the same word.
E.g. ‘corpora’ → ‘corpus’

Feature Extraction

Term Frequency-Inverse Document Frequency (TF-IDF)



Term frequency, $tf(t,d)$, is defined as the frequency of term t in one document d:

$$tf(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

where t' stands for all words in the document d, $f_{t,d}$ means the counts of word t in the document d. It is large when the word t occurs many times in document d.

Inverse document frequency, $idf(t,D)$, is defined to measure the uniqueness of a term t to some specific documents:

$$idf(t,D) = \log \frac{N}{|d \in D : t \in d|}$$

where N is the number of documents in corpus D, and the denominator is the number of documents that contain the term t. If the term doesn't appear in any documents, the denominator will be added 1 to avoid division by zero. It is large when the term t appears at a low frequency in documents. That is, rare terms lead to high idf values.

The TF-IDF feature is defined as the product of these two measures:

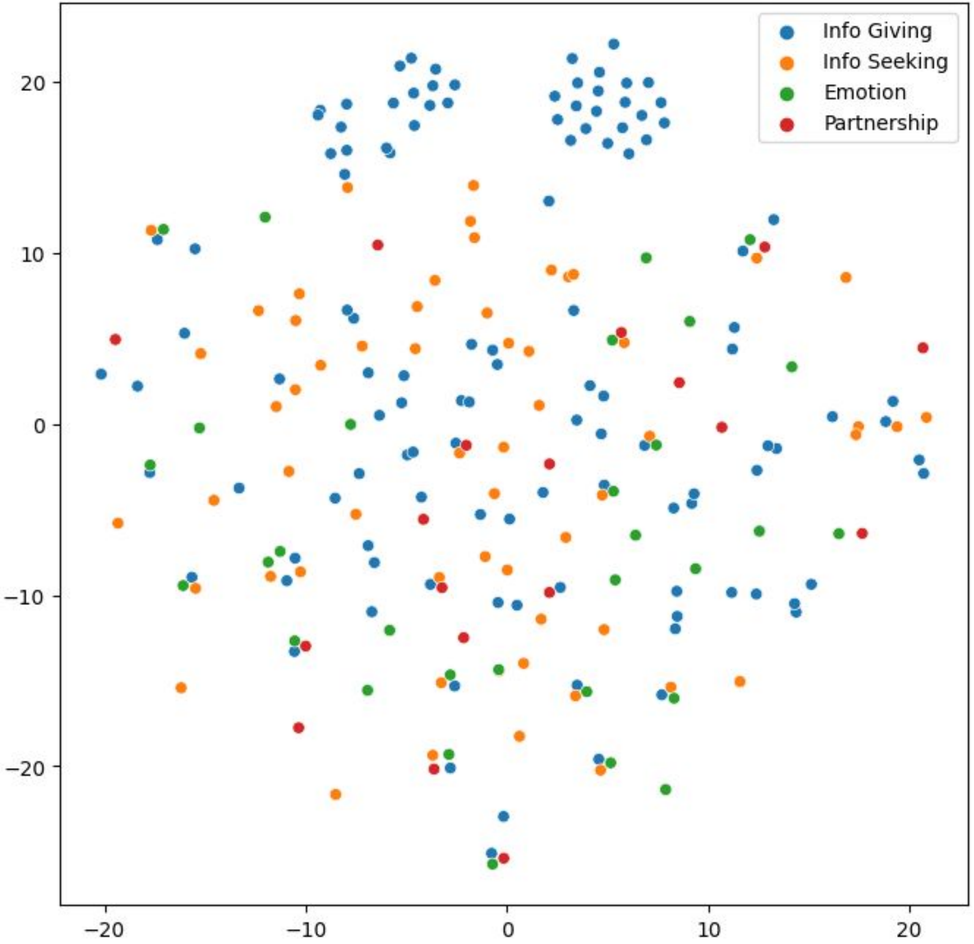
$$tf - idf(t,d,D) = tf(t,d) * idf(t,D)$$

It is high only when the term t is frequent in document d and not frequent in all other documents in D.

Sklearn.TfidfVectorizer(ngram_range=(1,3), max_df=0.75, max_features=1000)

Feature matrix shape = (270,1000)

TSNE 2-d Visualization



Feature matrix shape = (270,1000)

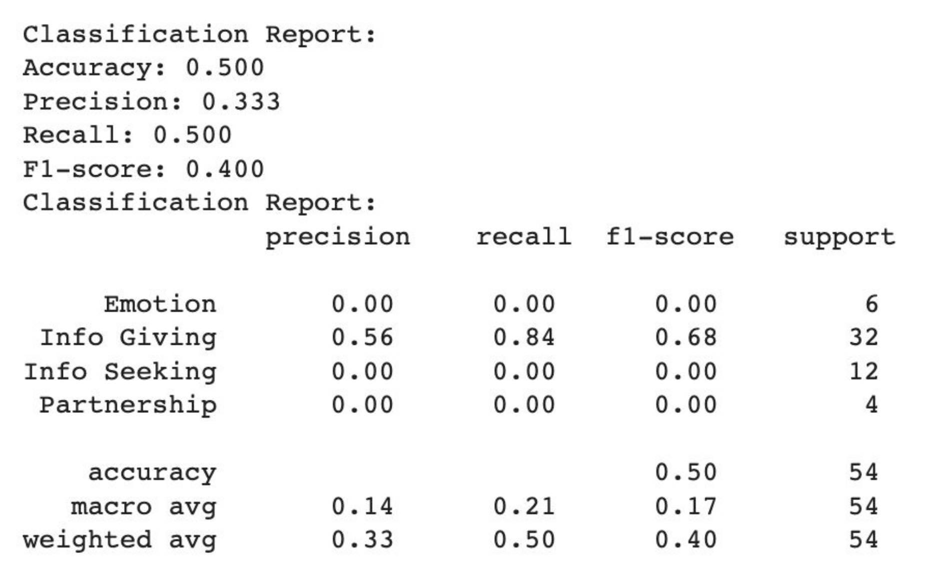
Project 1000-dimension features to 2-dimensional figure.

Machine Learning Models

Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, XGBoost

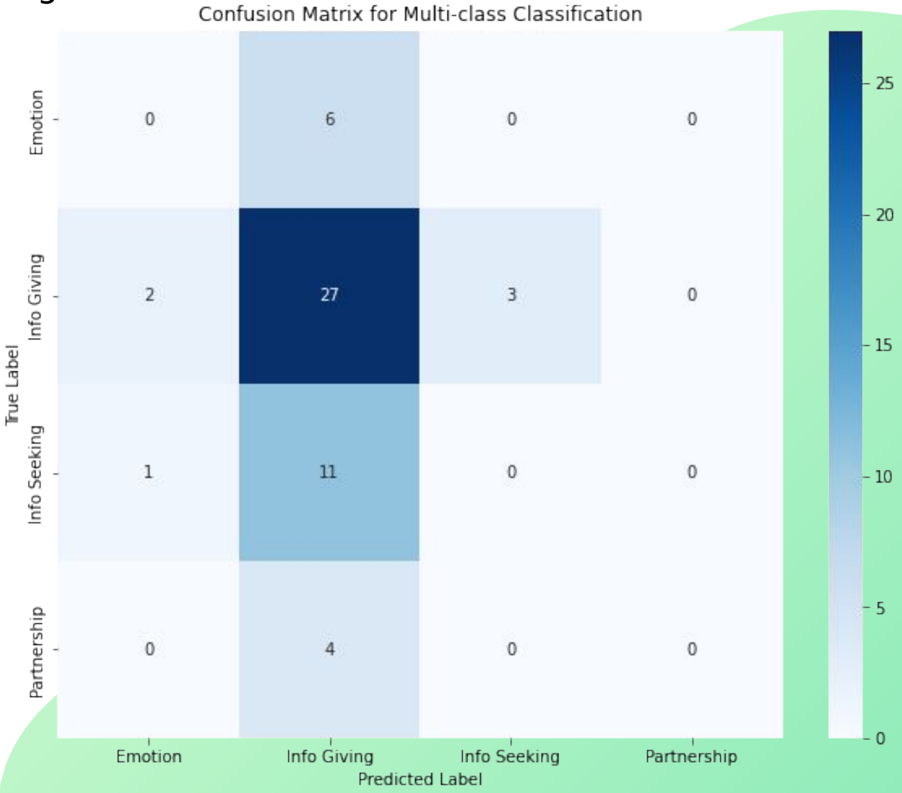
Grid Search & cross validation(split n=5) - hyperparameter tuning

Accuracy: 50% to 57%

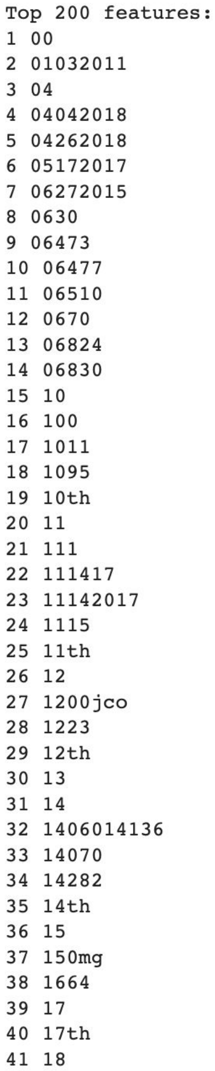


```
Classification Report:
Accuracy: 0.500
Precision: 0.333
Recall: 0.500
F1-score: 0.400
Classification Report:
              precision    recall  f1-score   support

     Emotion       0.00      0.00      0.00         6
 Info Giving       0.56      0.84      0.68        32
Info Seeking       0.00      0.00      0.00        12
 Partnership       0.00      0.00      0.00         4

    accuracy                           0.50        54
   macro avg       0.14      0.21      0.17        54
weighted avg       0.33      0.50      0.40        54
```

Random Forest's Confusion matrix



Problem Found



```
Top 200 features:
1 00
2 01032011
3 04
4 04042018
5 04262018
6 05172017
7 06272015
8 0630
9 06473
10 06477
11 06510
12 0670
13 06824
14 06830
15 10
16 100
17 1011
18 1095
19 10th
20 11
21 111
22 111417
23 11142017
24 1115
25 11th
26 12
27 1200jco
28 1223
29 12th
30 13
31 14
32 1406014136
33 14070
34 14282
35 14th
36 15
37 150mg
38 1664
39 17
40 17th
41 18
```

Meaningless but document-specific numeric terms have high tf-idf values

Solution:

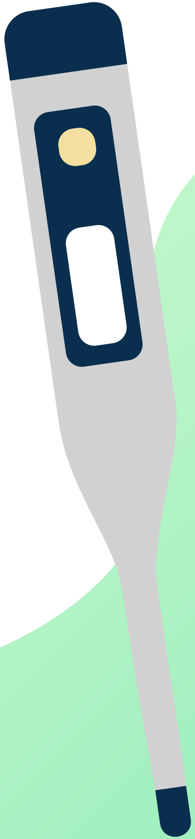
1. Replace some meaningless terms with special token, such as <NUMBER>

2. Try another feature selection method: chi-square test for binary classification

03

Improvement

Re-extract features using binary classification
Mitigate data insufficiency and imbalance



Question1: How to improve the features quality?

Extract **Top Features** from multiple Binary Classification Tasks

For each **MODEL** among {**logistic regression, decision tree, random forest, gradient boosting, XGBoost**}:

Perform binary classification using the **MODEL** for binary **TASKs** (such as Info-giving vs. Non-Info-giving, Info-seeking vs. Non-info-seeking, Emotion vs. non-emotion, and Partnership vs. non-partnership).

For each **TASK**:

1. Extracts the features from the text messages using the TF-IDF vectorizer.
2. Performs feature selection using chi-square test to select the **top K features**.
3. Splits the dataset into training and testing sets.
4. Trains a **MODEL** classifier using cross-validation.
5. Evaluates the trained classifier on the testing set.

After performing binary classification tasks, performs a multi-class classification using the **top(4*K) selected features** from all binary classification **TASKs** combined.

Features Interpretation Improved!

Top 25 features in XGBoost **Baseline** Model

```
1 00
2 01032011
3 04
4 04042018
5 04262018
6 05172017
7 06272015
8 0630
9 06473
10 06477
11 06510
12 0670
13 06824
14 06830
15 10
16 100
17 1011
18 1095
19 10th
20 11
21 111
22 111417
23 11142017
24 1115
25 11th
```

Top 25 features in XGBoost **4-fold feature extraction** Model

```
1 challenging
2 happen
3 hear
4 pattern
5 http
6 option
7 lovanox
8 bit
9 food
10 situation
11 overlook
12 manage
13 telemedicine
14 scheduling
15 explore
16 mild
17 generic
18 seek
19 parker
20 street
21 advice
22 assistance
23 received
24 provide
25 info
```

Question2: How to address the issues of small sample size and class imbalance?

1. **Resampling techniques**: oversample the minority class (partnership and emotional support), undersample the majority class (info-seeking and info-giving), or combine both methods to balance the class distribution. Techniques like **SMOTE** (Synthetic Minority Over-sampling Technique).

2. **GPT paraphrasing and generation**: use GPT-4 to learn the samples, then generate and paraphrase more samples for each class. Appropriate number of generated samples can be helpful to improve the power and generalization. However, it can be biased due to the GPT model property.

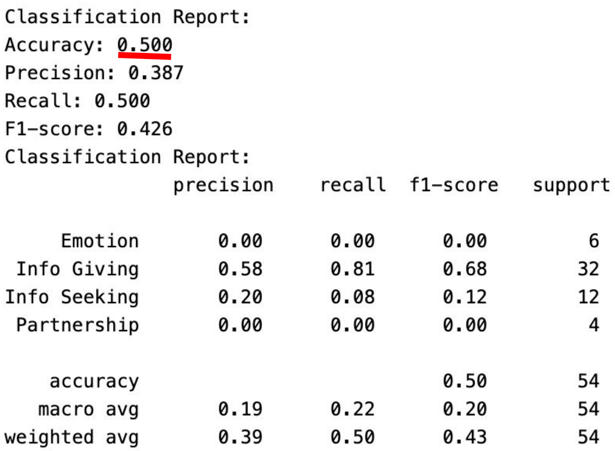
3. **Easy data augmentation (EDA)**: generates new samples from the existing data, data augmentation can improve model performance and generalization. (Wei & Zou, 2019)
    a. Synonym replacement (Replace n words in the sentence with synonyms from wordnet)
    b. Random Deletion (Randomly delete words from the sentence with probability p)
    c. Random Swap (Randomly swap two words in the sentence n times)
    d. Random Insertion (Randomly insert n words into the sentence)

Instances of Generated Data

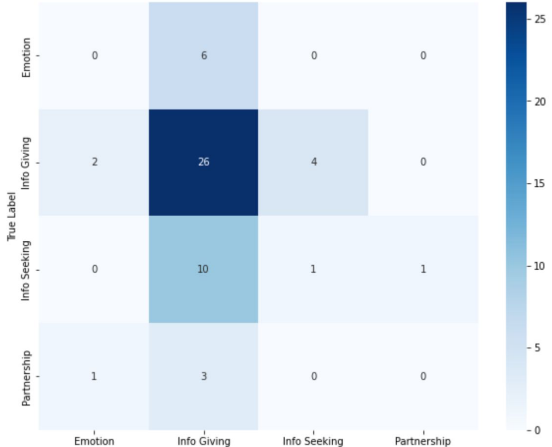
	GPT-4	EDA
Info-giving	I've noticed that my seasonal allergies seem to be worse this year. I've tried over-the-counter antihistamines, but they don't seem to be providing much relief.	
Info-seeking	Dr. Anderson, my child has been experiencing recurrent ear infections. Can we schedule an appointment to discuss potential causes and treatments to prevent future infections?	
Emotion	Hi Dr. Smith, my father recently underwent surgery, and his recovery is taking longer than expected. He is feeling down and discouraged. Can you please provide some words of encouragement or advice to help him stay positive during this challenging time? We appreciate your care and expertise.	i understand your concern but dr is away for a few days i have you message and will give it to him tomorrow if he is reachable if not for sure on friday when he is back in the office
Partnership	Dr. Smith, I just started my new medication and I'm experiencing some side effects. Can we discuss some potential adjustments or alternatives?	dr has asked if it is ok for me to stop the xeloda days prior to the ct myelogram that they are trying to docket if this is ok ill corroborate with drs staff give thanks you
SDM		sounds good i just set in prescription for numbingsoothing medication as well do nt need to pick it up if you spirit better with pepto

GPT helps!

Without GPT, XGBoost

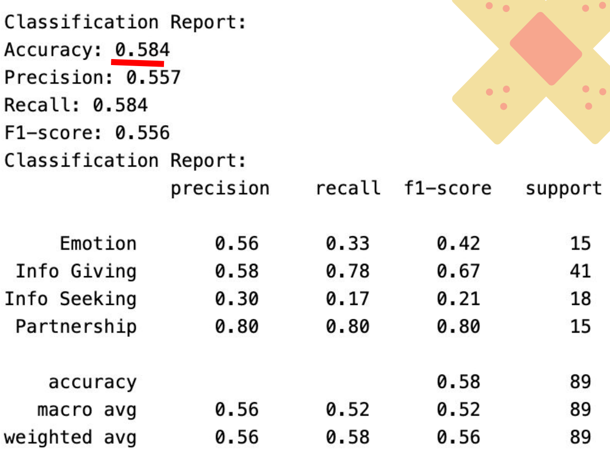


```
Classification Report:
Accuracy: 0.500
Precision: 0.387
Recall: 0.500
F1-score: 0.426
Classification Report:
              precision    recall  f1-score   support

     Emotion       0.00      0.00      0.00         6
 Info Giving       0.58      0.81      0.68        32
Info Seeking       0.20      0.08      0.12        12
 Partnership       0.00      0.00      0.00         4

    accuracy                           0.50        54
   macro avg       0.19      0.22      0.20        54
weighted avg       0.39      0.50      0.43        54
```

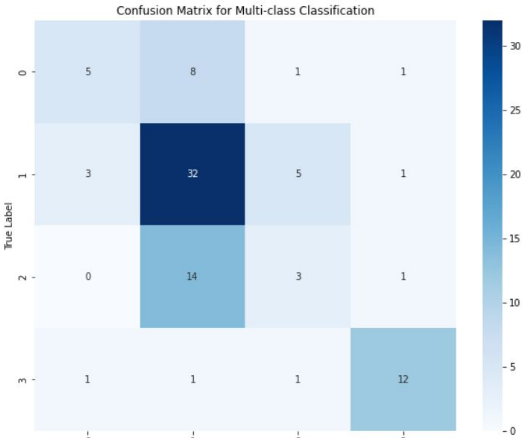


	Emotion	Info Giving	Info Seeking	Partnership
Emotion	0	6	0	0
Info Giving	2	26	4	0
Info Seeking	0	10	1	1
Partnership	1	3	0	0

With GPT, XGBoost



```
Classification Report:
Accuracy: 0.584
Precision: 0.557
Recall: 0.584
F1-score: 0.556
Classification Report:
              precision    recall  f1-score   support

     Emotion       0.56      0.33      0.42        15
 Info Giving       0.58      0.78      0.67        41
Info Seeking       0.30      0.17      0.21        18
 Partnership       0.80      0.80      0.80        15

    accuracy                           0.58        89
   macro avg       0.56      0.52      0.52        89
weighted avg       0.56      0.58      0.56        89
```



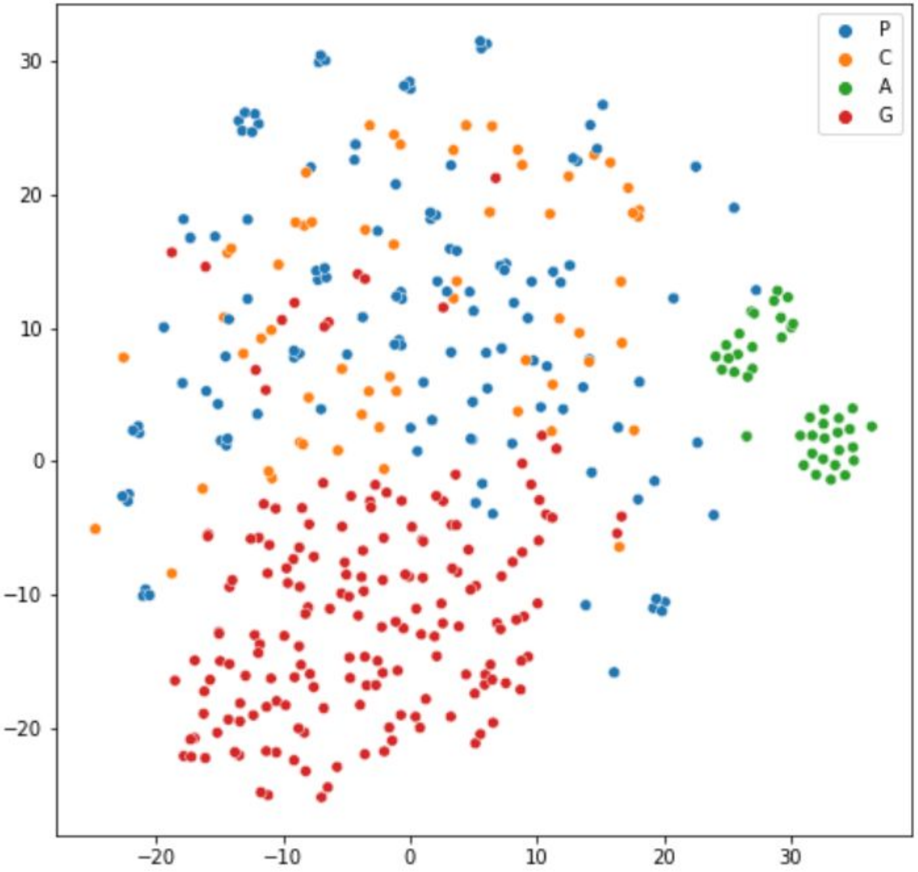
Confusion Matrix for Multi-class Classification

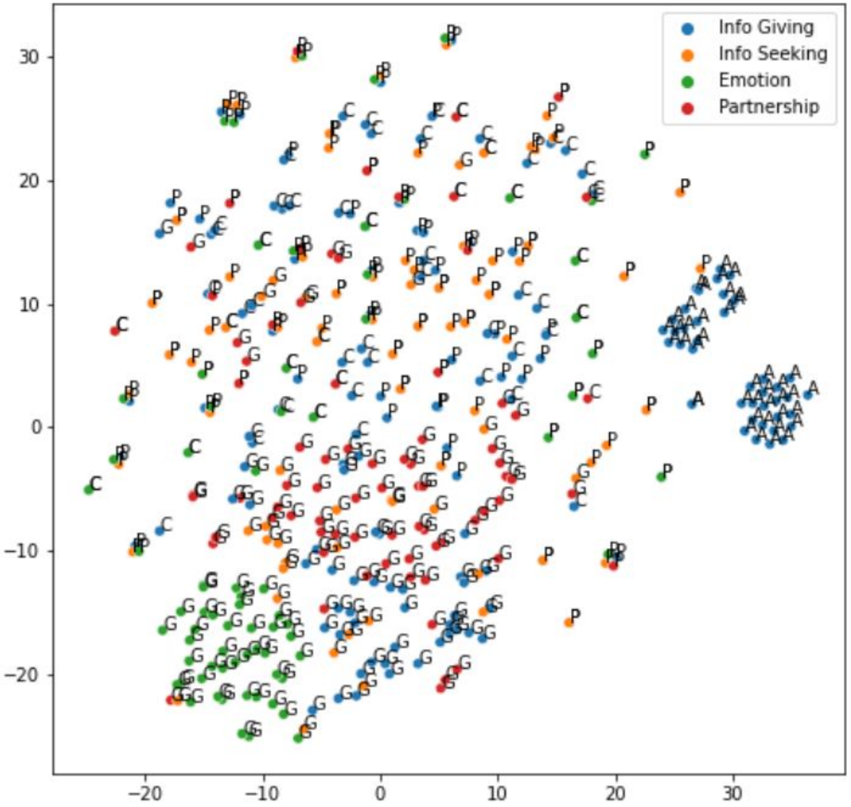
0	5	8	1	1
1	3	32	5	1
2	0	14	3	1
3	1	1	1	12

1. Generated samples improve the model performance metric.

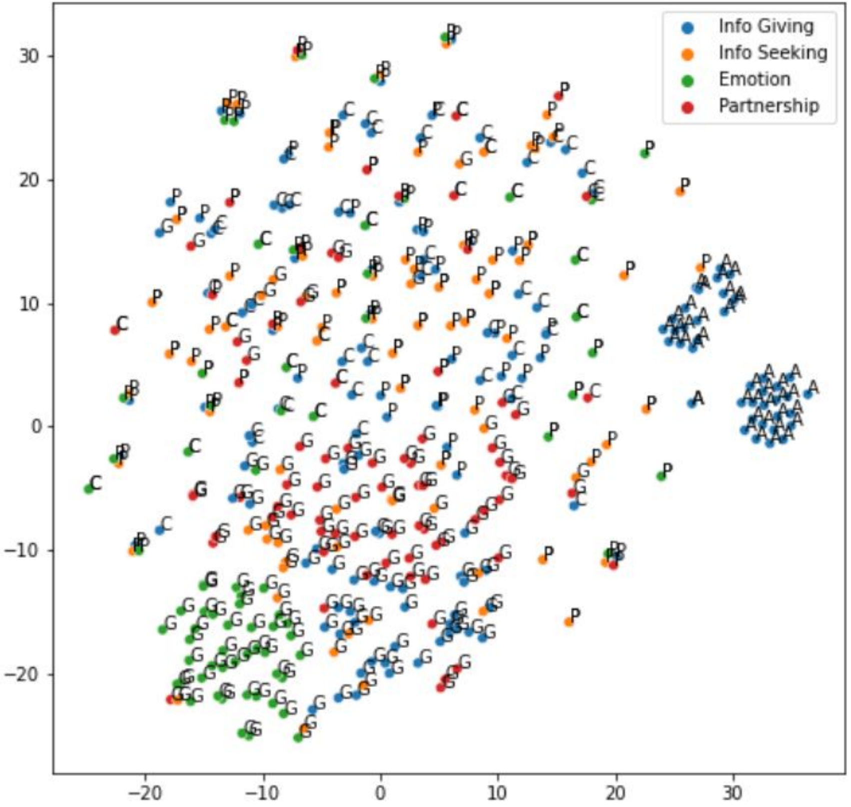
2. Generated samples make the model prediction less concentrated on one class.

GPT helps! But what is the cost?



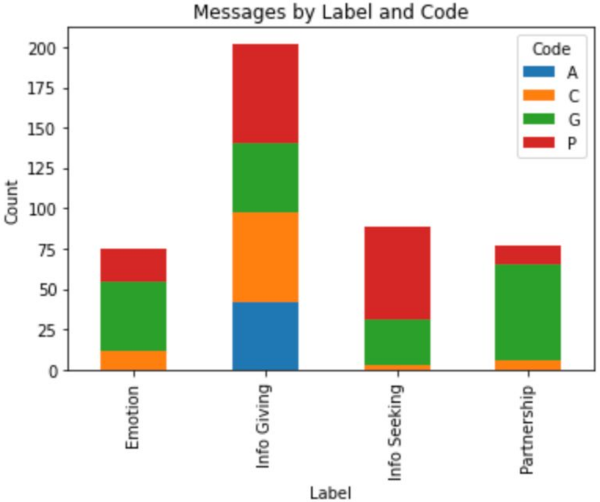


How to make GPT helpful?



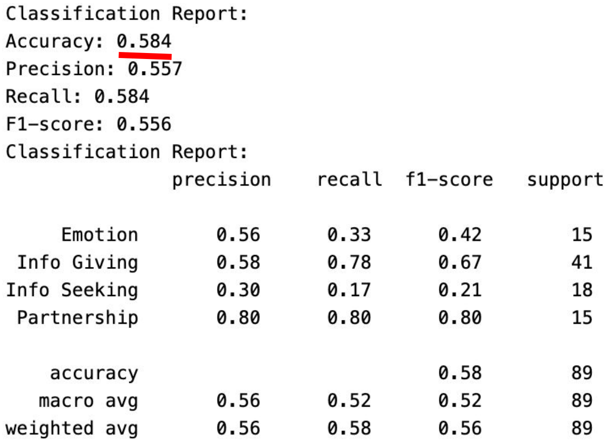
Pro: In the embedding space, more samples of the same class are generated near the embeddings of each class, increasing the power of the classification model.
Con: Bias introduced by the GPT model.

Control the generated number to trade off!

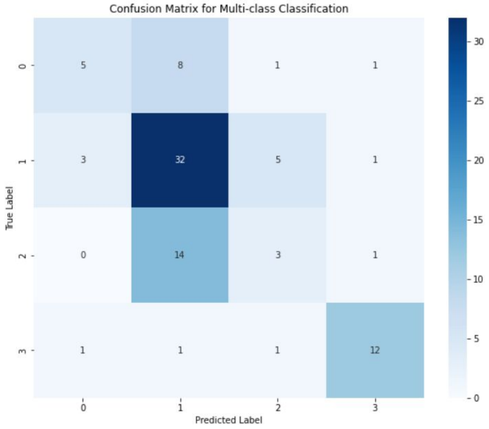


Data augmentation also helps...

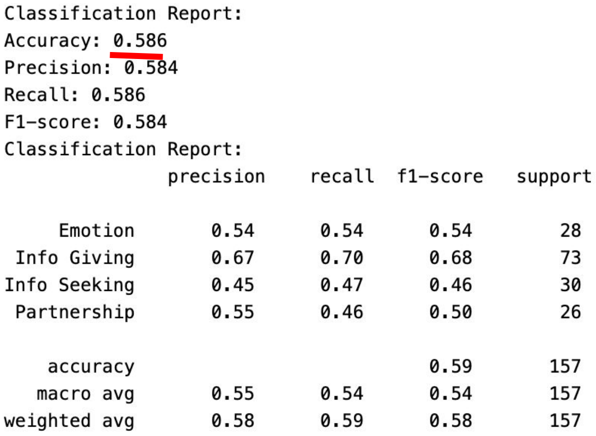
With GPT, without EDA, XGBoost



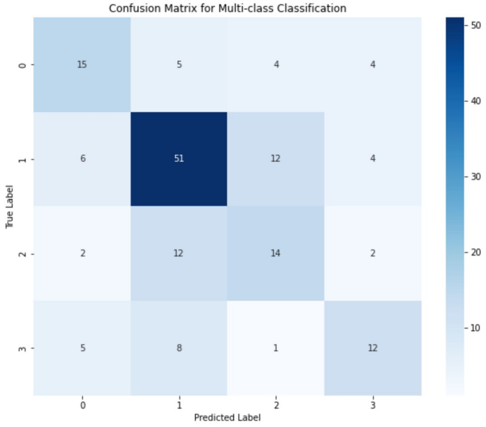
```
Classification Report:
Accuracy: 0.584
Precision: 0.557
Recall: 0.584
F1-score: 0.556
Classification Report:
              precision    recall  f1-score   support

     Emotion       0.56      0.33      0.42        15
 Info Giving       0.58      0.78      0.67        41
Info Seeking       0.30      0.17      0.21        18
 Partnership       0.80      0.80      0.80        15

    accuracy                           0.58        89
   macro avg       0.56      0.52      0.52        89
weighted avg       0.56      0.58      0.56        89
```



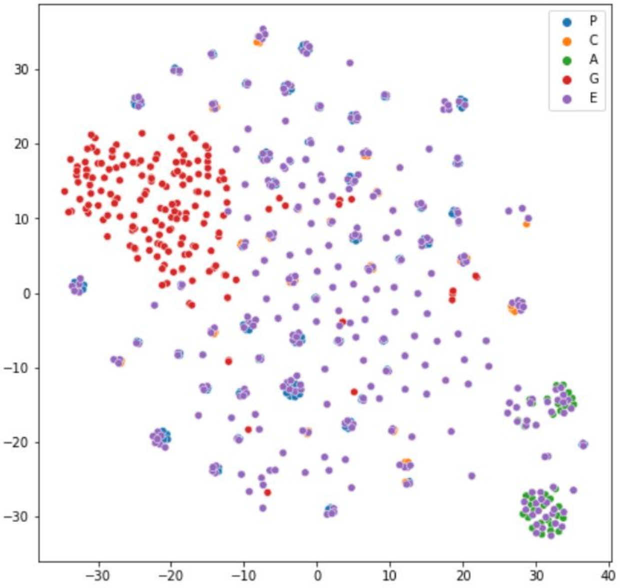
With GPT and EDA, XGBoost

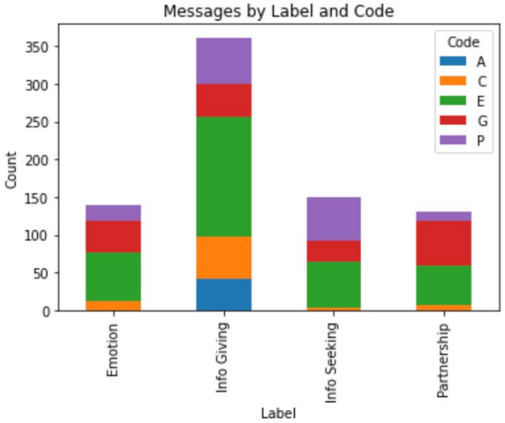


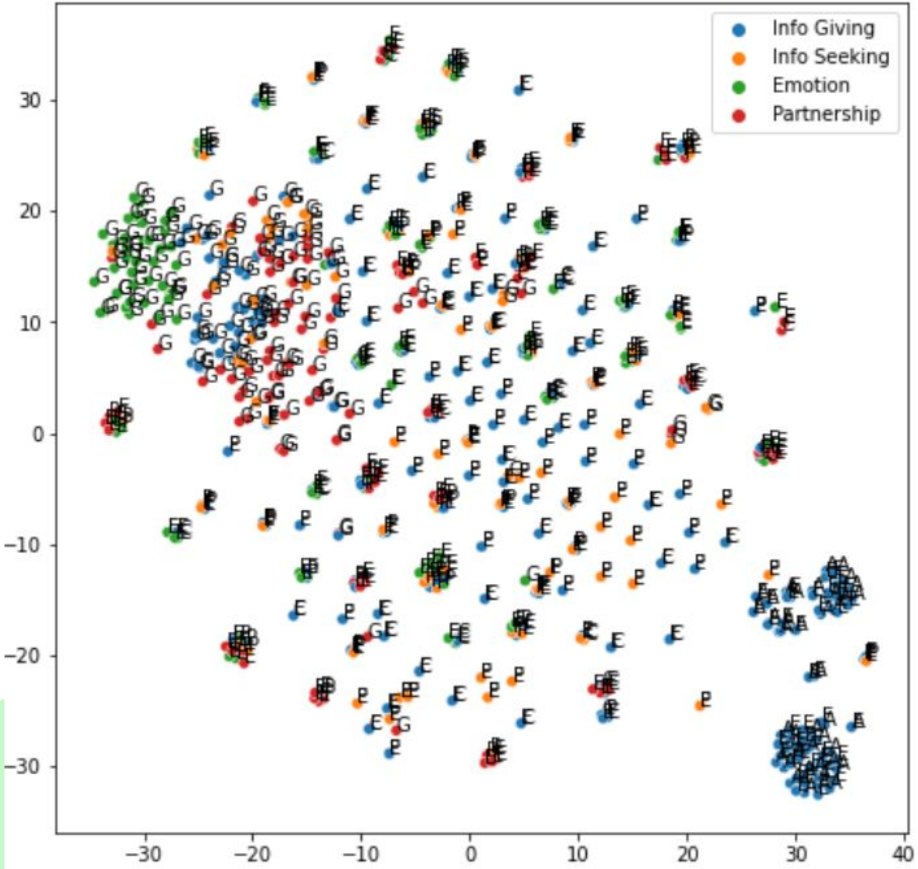
```
Classification Report:
Accuracy: 0.586
Precision: 0.584
Recall: 0.586
F1-score: 0.584
Classification Report:
              precision    recall  f1-score   support

     Emotion       0.54      0.54      0.54        28
 Info Giving       0.67      0.70      0.68        73
Info Seeking       0.45      0.47      0.46        30
 Partnership       0.55      0.46      0.50        26

    accuracy                           0.59       157
   macro avg       0.55      0.54      0.54       157
weighted avg       0.58      0.59      0.58       157
```



What does the EDA embeddings look like?

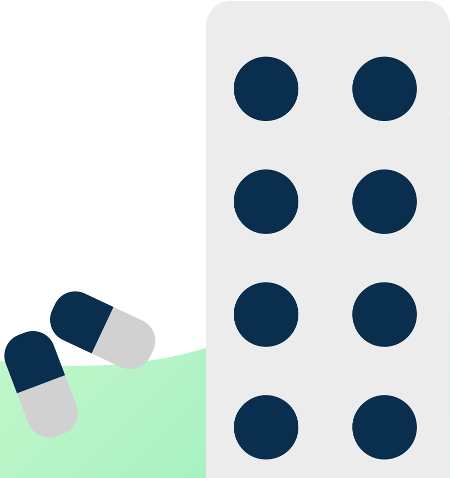




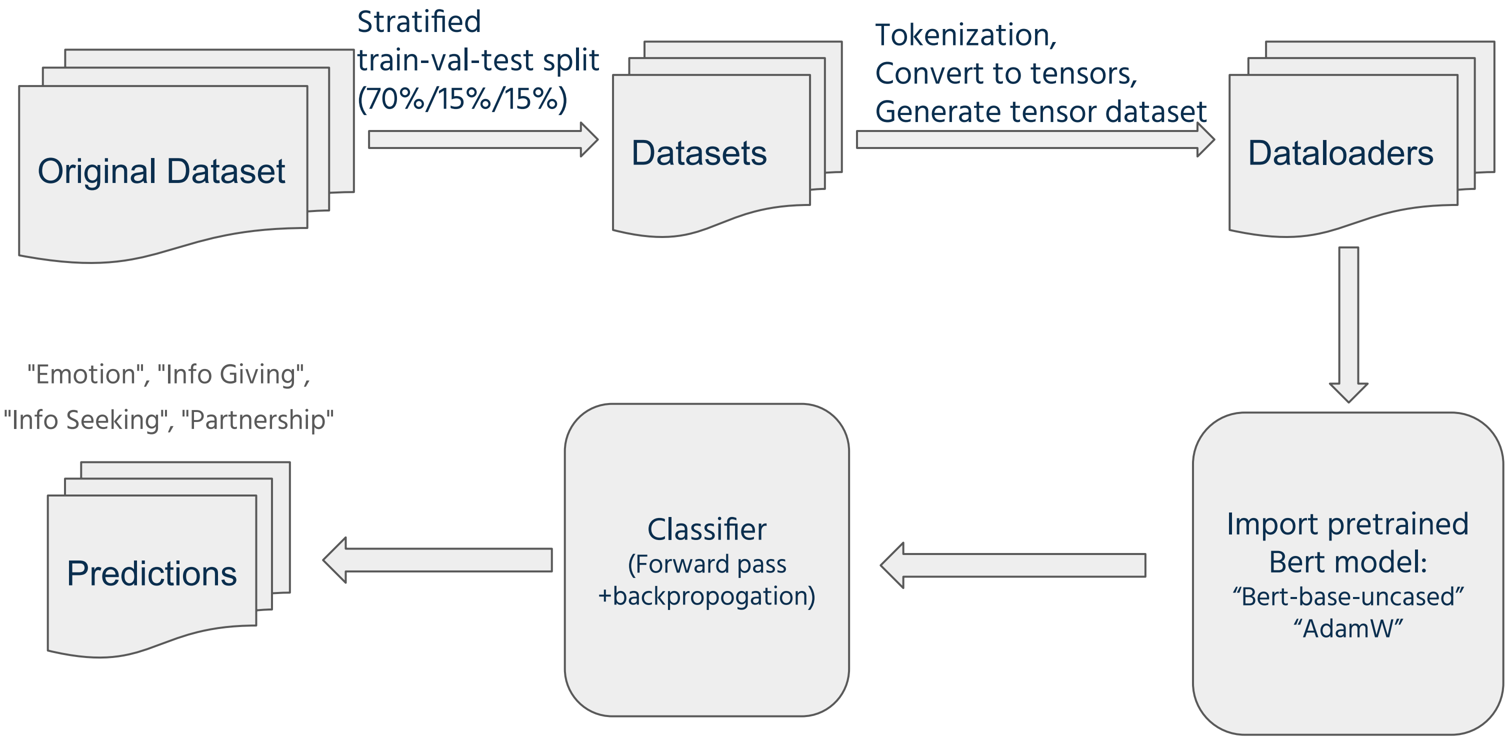


04

Bert Models



Baseline Bert Model (with Original Dataset)



Original Dataset

Stratified train-val-test split (70%/15%/15%)

Datasets

Tokenization, Convert to tensors, Generate tensor dataset

Dataloaders

Import pretrained Bert model: “Bert-base-uncased” “AdamW”

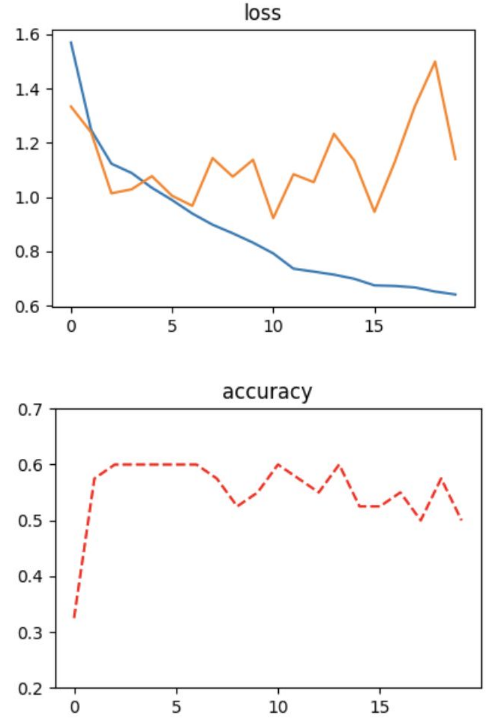
Classifier (Forward pass +backpropogation)

Predictions

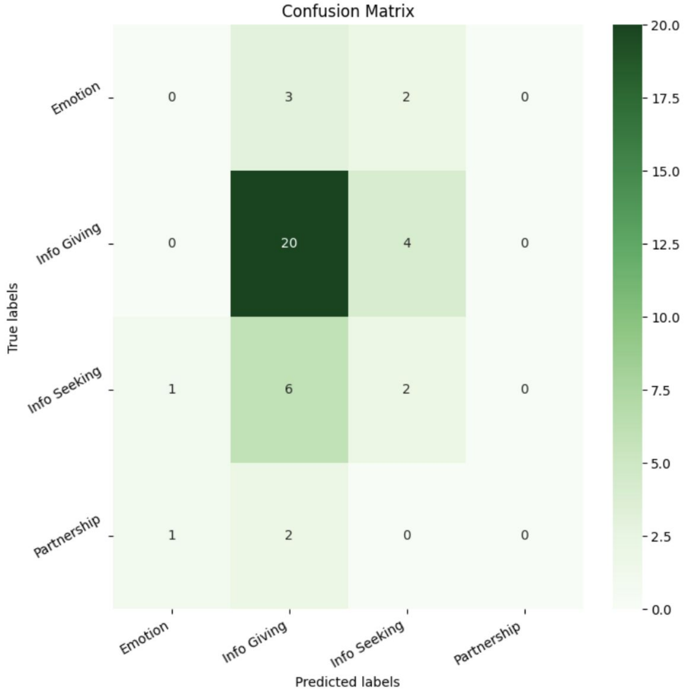
"Emotion", "Info Giving", "Info Seeking", "Partnership"

Baseline Bert Model (with Original Dataset)

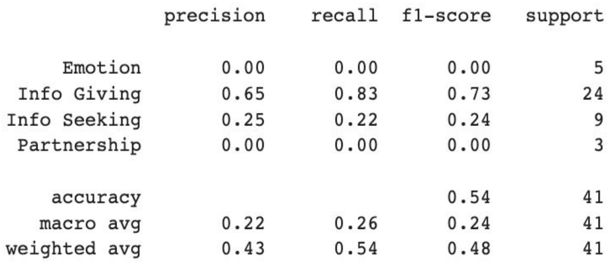
Training and Validation Performance:



Testing Performance:

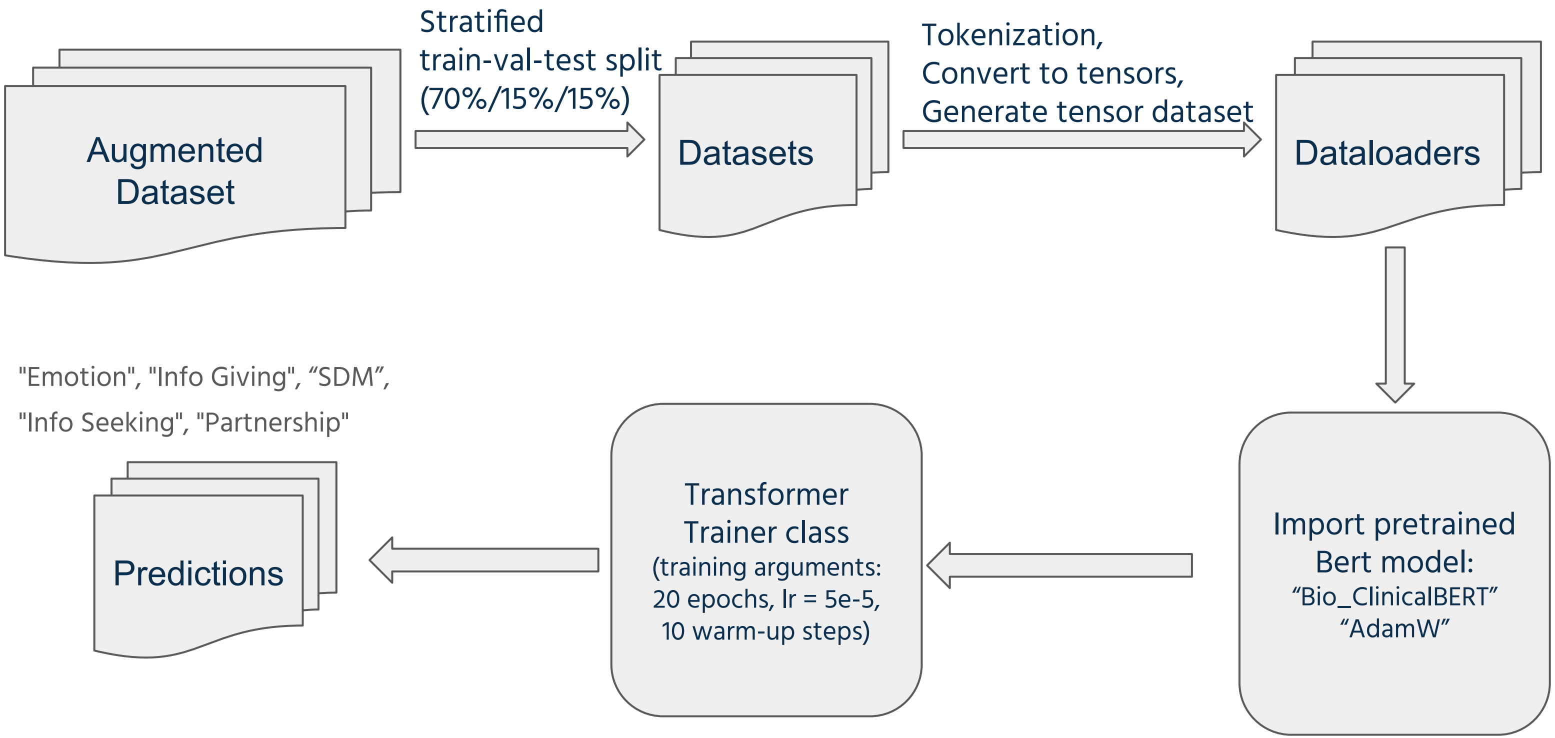


True labels \ Predicted labels	Emotion	Info Giving	Info Seeking	Partnership
Emotion	0	3	2	0
Info Giving	0	20	4	0
Info Seeking	1	6	2	0
Partnership	1	2	0	0



```
              precision    recall  f1-score   support

     Emotion       0.00      0.00      0.00         5
 Info Giving       0.65      0.83      0.73        24
Info Seeking       0.25      0.22      0.24         9
 Partnership       0.00      0.00      0.00         3

    accuracy                           0.54        41
   macro avg       0.22      0.26      0.24        41
weighted avg       0.43      0.54      0.48        41
```

Improved Bert (Clinical Bert with Augmented Data)



Augmented Dataset

Stratified train-val-test split (70%/15%/15%)

Datasets

Tokenization,
Convert to tensors,
Generate tensor dataset

Dataloaders

Import pretrained Bert model:
“Bio_ClinicalBERT”
“AdamW”

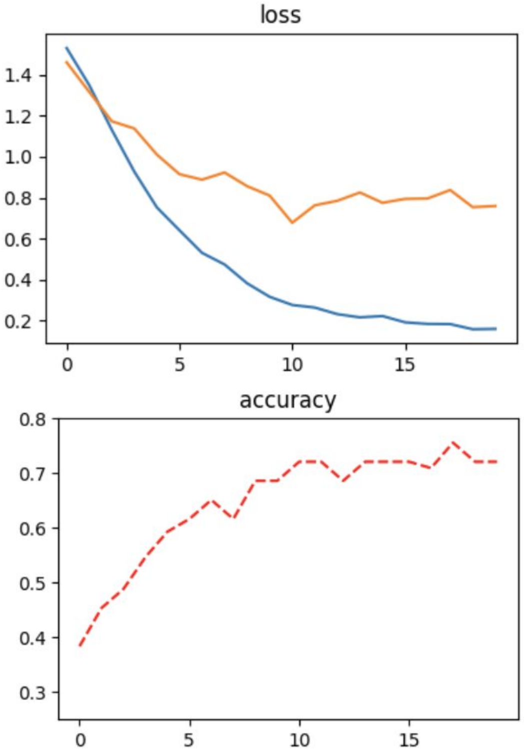
Transformer Trainer class
(training arguments:
20 epochs, lr = 5e-5,
10 warm-up steps)

Predictions

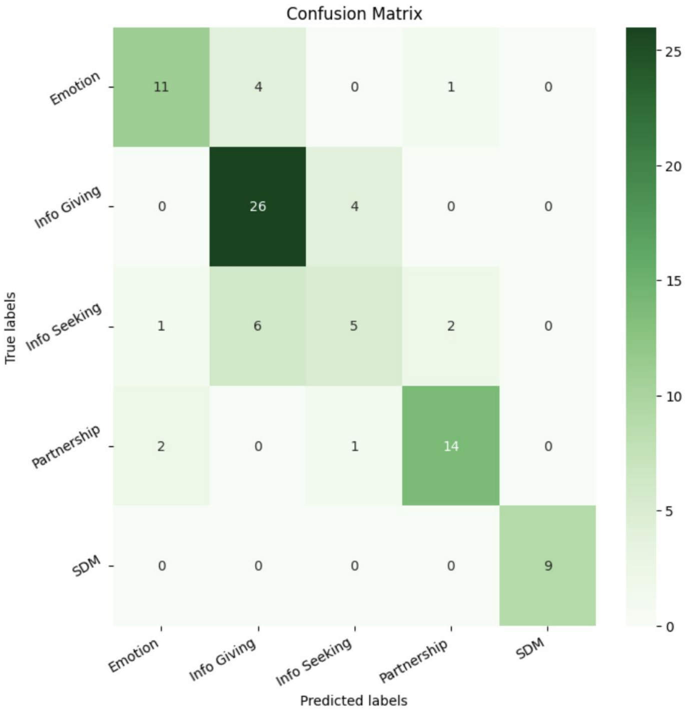
"Emotion", "Info Giving", “SDM”,
"Info Seeking", "Partnership"

Improved Bert (Clinical Bert with Augmented Data)

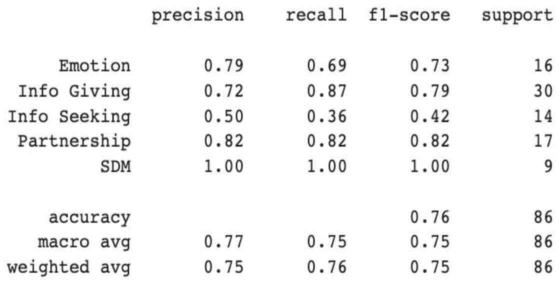
Training and Validation Performance:



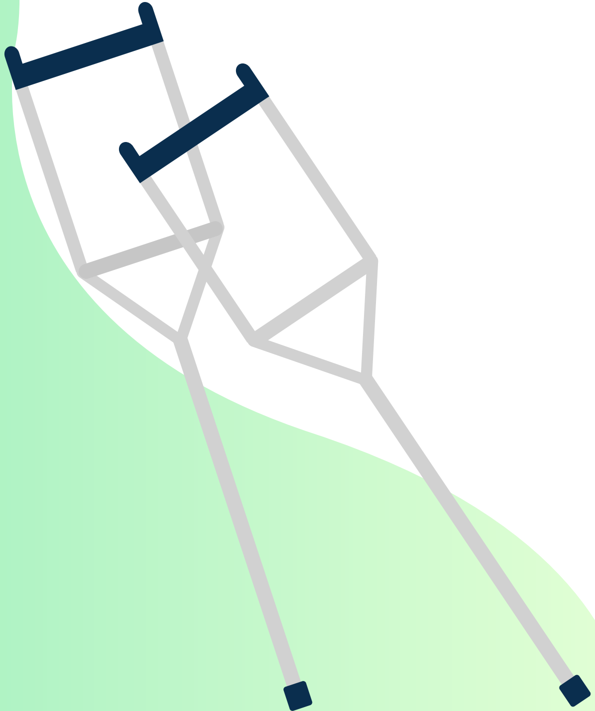
Testing Performance:

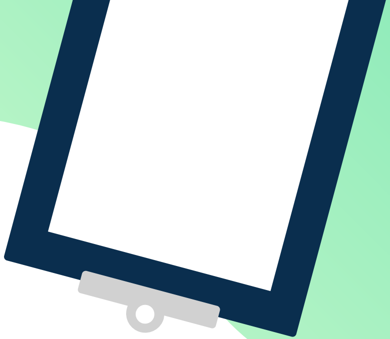


	Emotion	Info Giving	Info Seeking	Partnership	SDM
Emotion	11	4	0	1	0
Info Giving	0	26	4	0	0
Info Seeking	1	6	5	2	0
Partnership	2	0	1	14	0
SDM	0	0	0	0	9



```
              precision    recall  f1-score   support

     Emotion       0.79      0.69      0.73        16
 Info Giving       0.72      0.87      0.79        30
Info Seeking       0.50      0.36      0.42        14
 Partnership       0.82      0.82      0.82        17
         SDM       1.00      1.00      1.00         9

    accuracy                           0.76        86
   macro avg       0.77      0.75      0.75        86
weighted avg       0.75      0.76      0.75        86
```

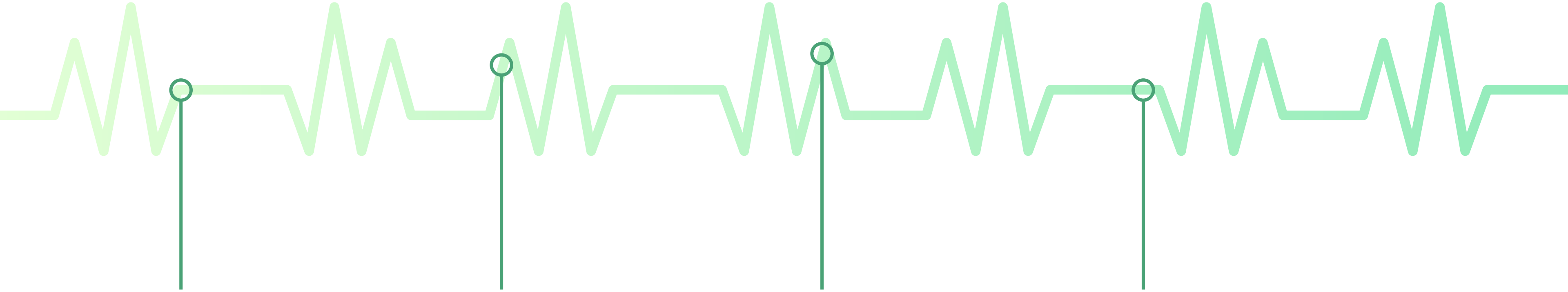




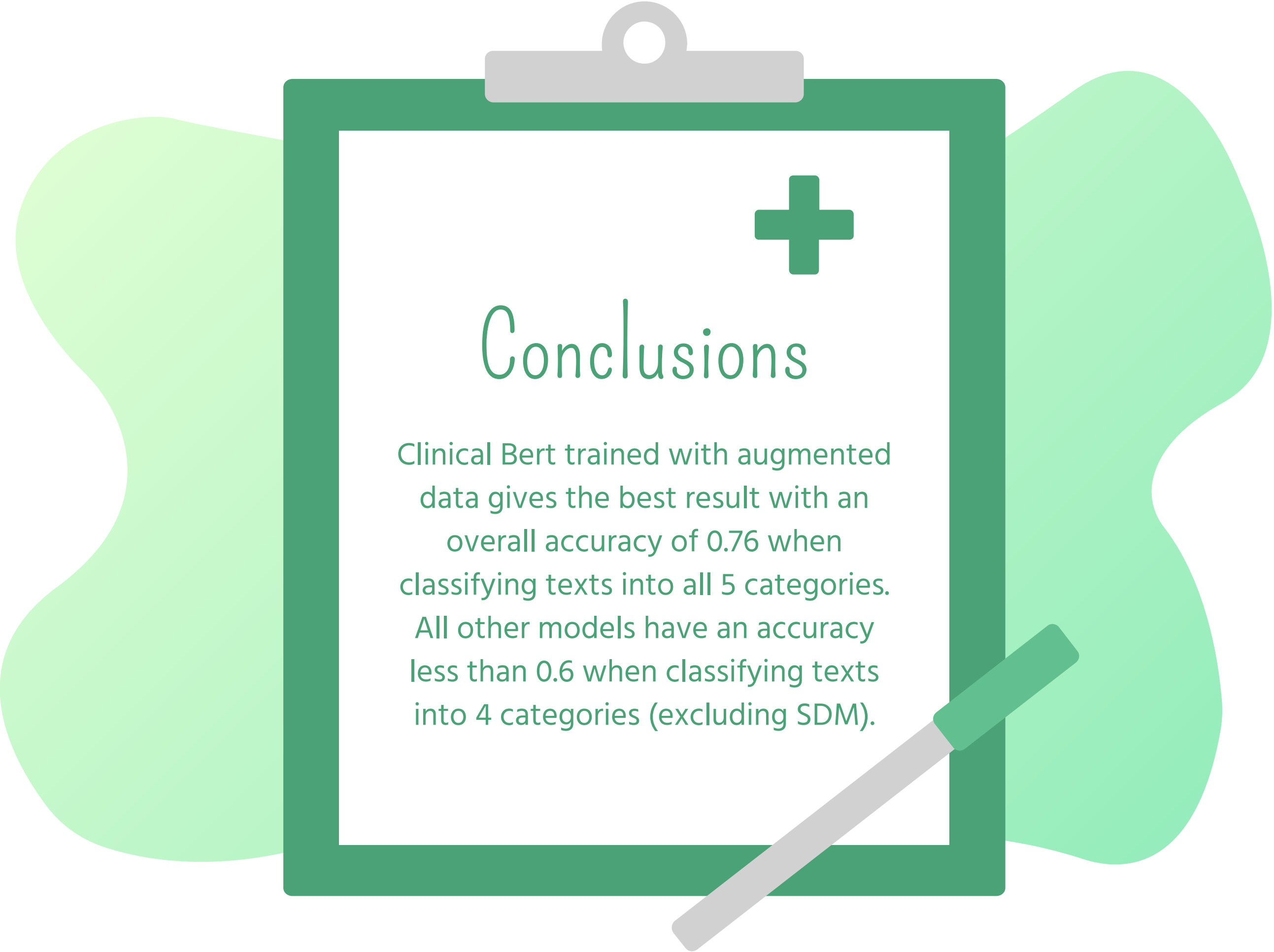
05

Conclusion & Discussion

Model Performance (Overall Accuracy)

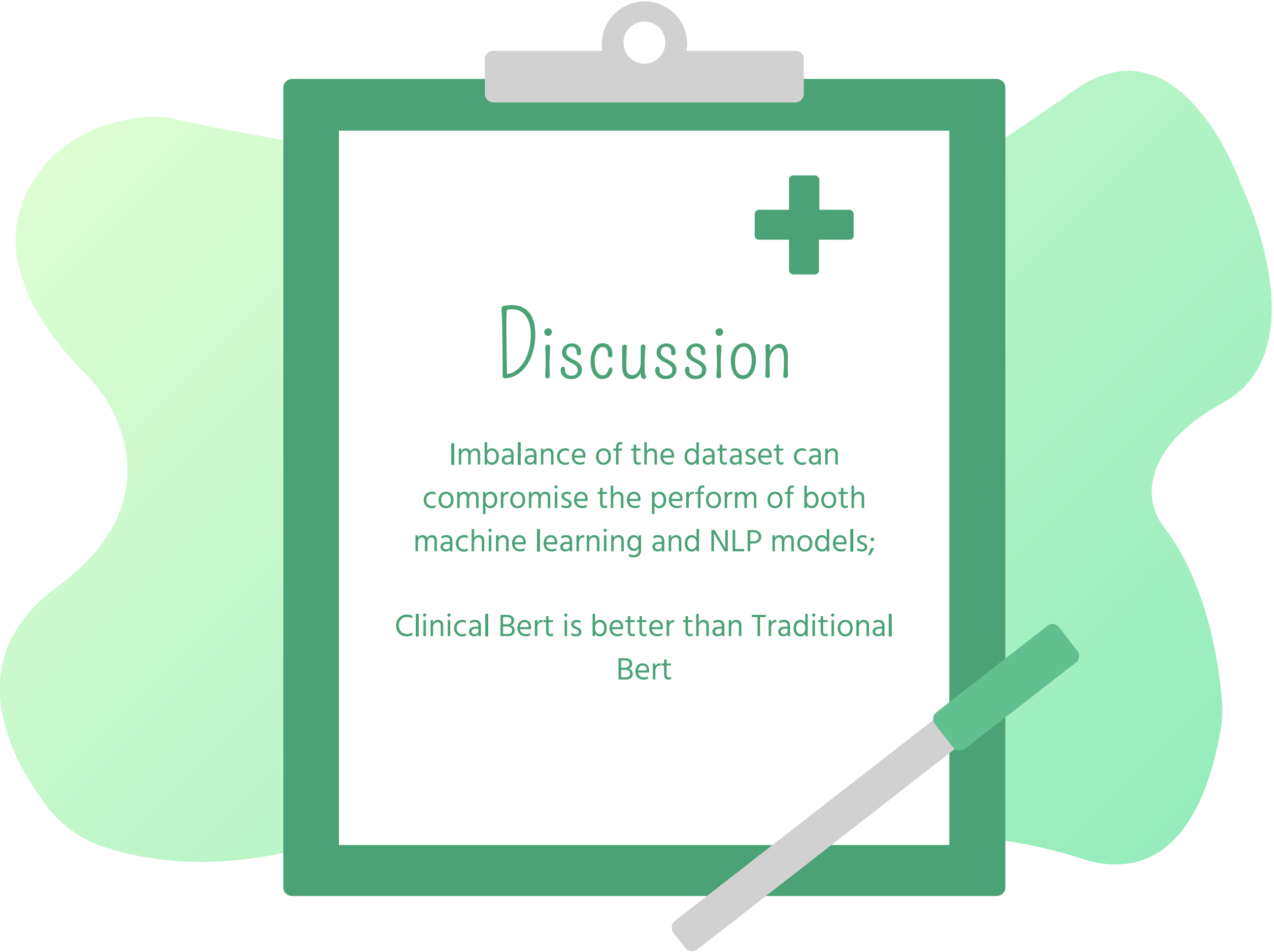


Baseline ML	Baseline Bert	Improved ML	Improved Bert
0.50~0.57	0.54	0.59	0.76



Conclusions

Clinical Bert trained with augmented data gives the best result with an overall accuracy of 0.76 when classifying texts into all 5 categories. All other models have an accuracy less than 0.6 when classifying texts into 4 categories (excluding SDM).



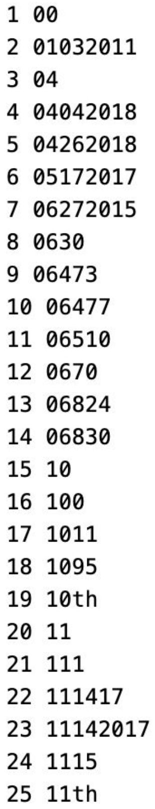
Discussion

Imbalance of the dataset can compromise the perform of both machine learning and NLP models;

Clinical Bert is better than Traditional Bert

Classify Sentence-by-Sentence may make more sense

- Some messages are coded into more than one category
  - E.g. “Hi That sounds reasonable I would do the colonoscopy more electively when no pressing issues are noted Best Vik” is coded as both “Emotion” and “Share Decision-Making”
- Some messages contain multiple sentences which should be coded into different categories
  - E.g. “Dr Is there any chance I could be given a summary of the procedure and what was done without having to wait until March 24th as the wait is driving me nuts with the thought of what could be It would just greatly put my mind at ease and thank you for a great procedure as there was absolutely no pain or discomfort once I awoke in recovery”
  -
  - Info-seeking Emotion Info-giving



```
1 00
2 01032011
3 04
4 04042018
5 04262018
6 05172017
7 06272015
8 0630
9 06473
10 06477
11 06510
12 0670
13 06824
14 06830
15 10
16 100
17 1011
18 1095
19 10th
20 11
21 111
22 111417
23 11142017
24 1115
25 11th
```

PPC data can be messy

- Numbers can break feature extraction
- Phone numbers, email addresses, zip codes, dates and times, dosage
- Match them using regular expression and replace them with special tokens

Further Exploration on Balancing the Data

- Resampling using SMOTE
- Try different EDA parameters
- Try different amount of balance

BERT hyperparameter search

- With more GPU access, we can conduct BERT hyperparameter search using trainer API
- This can potentially help further improve model performance

References

Packages

- Sklearn
- Matplotlib
- Seaborn
- NLTK
- Transformers
- re

Literature

- Wei & Zou (2019), *EDA: Easy Data Augmentation Techniques fo Boosting Performance on Text Classification Tasks,* https://doi.org/10.48550/arXiv.1901.11196

Data

- Fodeh (2023),