TrainAssist

Final Report

Total Words(Excluding References,Permissions and descriptions): 1939
Penalty : 0

Permissions

Permission by Karan Kapur:

1. Permission to post video: Yes
2. Permission to post final report: Yes
3. Permission to post source code: No

Permission by Sai Praveen:

1. Permission to post video: Yes
2. Permission to post final report: Yes
3. Permission to post source code: No

Introduction

TrainAssist is a text classifier tool developed in collaboration with the Centre for Addiction and Mental Health, CAMH. In a recent effort to improve their operations, CAMH conducts Trauma-Informed De-escalation Education for Safety and Self-Protection or TIDES training for all the CAMH Staff. The TIDES training deals with various ways to handle patients with mental disorders which include techniques to restrain patients, control the mental breakdown by having debriefing sessions etc.

The tool was developed having two goals in mind:

1. Survey response classification using numerical and textual data collected from TIDES(Trauma-Informed De-escalation Education for Safety and Self-Protection) pre and post training surveys.
2. Perform Topic modeling on useful feedback responses to extract themes/topics to identify areas of improvements to assist CAMH to improve on their training materials and methodologies.

With TrainAssist, we aim to polish this process for the team using new state of the art Natural Language Processing models to classify future training data and execute topic modeling.We believe developing this classifier tool will help CAMH in improving their training methodologies and in future have a single tool for all training programs.

Illustration

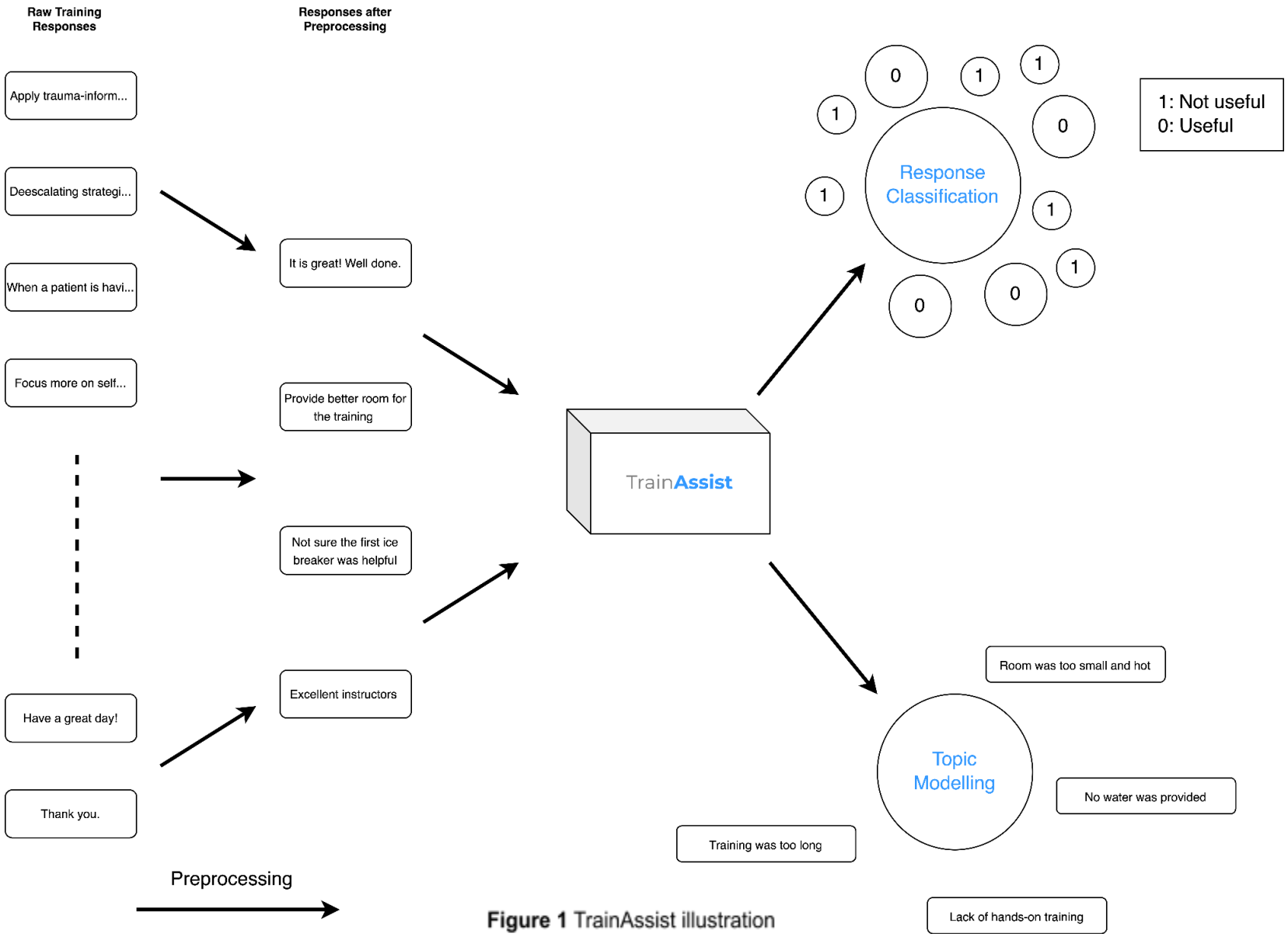


Figure 1 TrainAssist illustration

Background and related work

Some interesting papers and projects which have used multimodal architecture and their workings are:

1. Enriching BERT with Knowledge Graph Embeddings for Document Classification : Talks about how to combine text representations with metadata and knowledge graph embeddings on BERT, which encode author information.[1]

2. Ken Gu in his blogs mentions greatly the use of multimodal transformers with tabular data along with text with a good example from clothing dataset.[2]

3. TabTransformer: Tabular Data Modeling Using Contextual Embeddings talk about how the TabTransformer is built upon Transformers (Vaswani et al. 2017) to learn efficient contextual embeddings of categorical features. [3]

Data

Data was collected by the TIDES training staff by asking trainees to fill out a survey form. The survey has overall 4 parts :

1. User Details: Name, Gender, DOB, Department, Years of Experience
2. Pre-Evaluation: Set of questions which test the user knowledge on the techniques before the training
3. Post Evaluation: Same set of questions as Pre-Evaluation but the user answers this only post training which could give an idea of the impact of training.
4. Overall Feedback: Set of questions which asks users for feedback on the training program.

Data preprocessing

Cleaning of the subjective data which involved:

1. Removing all the blanks/NaNs
2. Removing responses having only one word
3. Other irrelevant responses

Labeling of all the subjective data into positive and negative responses was done as follows :

- Positive(label = 1) were assigned to responses that did not provide any useful information.
- Negative(label = 0) were assigned to responses that provided useful information.

Some examples of the labeled data are as follows:

Survey response	**Label**	**Useful/ Not useful**
ED special training would be appreciated. Same team control demos were not applicable/practice to ED setting	1	Useful
The space provided for training is very uncomfortable. Seating is not conducive to learning for hours	1	Useful
The trainers were well prepared	0	Not useful
It was a nice retreat to be to out from the unit and chat and learn from one another	0	Not useful

Table 1 Survey response examples

The following statistics describe about our pre and post questions (numerical data)

- A total of 40 columns
- Each column records the response on a scale from 1 to 5 (1 being least) for a user in their knowledge area.

A survey section is given below for reference.

Section 1: Knowledge

1. **Please rate your current level of understanding on the following:**

	Poor	Fair	Neutral	Good	Excellent
a. Understand the effects of moral conflict, distress, residue in the workplace	○	○	○	○	○
b. Understand trauma-informed strategies when utilizing restraints (e.g. seclusion, mechanical, chemical)	○	○	○	○	○
c. Understand what client behaviors warrant the initiation of seclusion, mechanical restraints, and/or chemical restraints as a last resort	○	○	○	○	○

Figure 2 Survey layout

The raw data provided by CAMH contained about 900 data points with NaN percentage of about 57%. The data has a class imbalance where about 70% of the responses are positive. We also observed that for some users the pre and post evaluation data i.e. numerical values were missing. We addressed this challenge by imputing the missing values with the average value of that particular response.Two different datasets were created from the processed data. One dataset has class imbalance and the other dataset has class balance.

Data Handling steps:

- Dataset of 384 samples was split into 80:20 stratified datasets for training and testing.
- To have a more balanced data, the class 1 labels were was downsampled to eliminate imbalance
- In addition, the dataset was augmented using the Parrot package and GPT-3s OpenAI playground to execute topic modeling.

In addition to the TIDES Dataset, CAMH supplied us with an additional dataset containing survey responses from the Integrated Care Upskilling carried out to train nurses. We utilized this dataset to test our baseline model.

Architecture and software

Response classification

The final architecture for survey response classification was chosen to be multimodal transformer architecture where a combining module is added which uses the outputs of transformers, and numerical features to generate enhanced multimodal features for classification further down the line.

We plan to combine textual data applied in the baseline model and combine the numerical data in order to observe if we have better classification results. The multimodal transformer toolkit provides multiple ways to work around the combination of the two types of data. An example of the combining architecture is as follows:

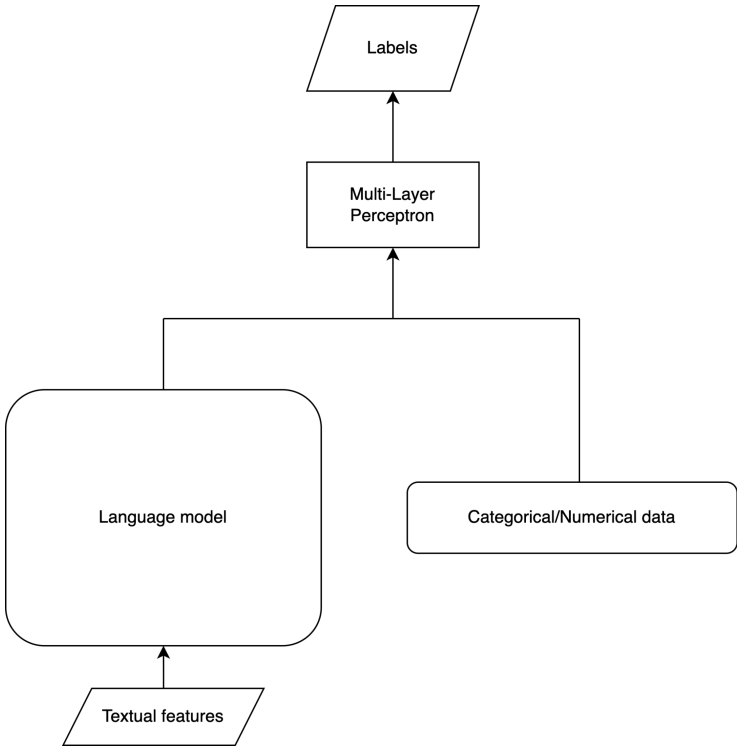


Figure 3 Multimodal Architecture

The BERT architecture uses 12 hidden layers with each layer consisting of 768 units which takes in only the textual sentences. The non-textual features are processed (zero mean, unit variance) and in the next step both of these outputs are concatenated and passed into a MLP with 2 layers and ReLU activation function. The final layer does the classification after applying the softmax function. Two different combining methods were tried on our dataset as per the toolkit [4].

Combining Method	Description
individual_mlps_on_cat_and_numerical_feats_then_concat	Inputting categorical/numerical features through an MLP then concatenation with the transformer output.
weighted_feature_sum_on_transformer_cat_and_numerical_feats	Learnable weighted feature-wise sum of transformer outputs, numerical feats and categorical feats for each feature dimension before final classifier layer(s)

Table 2 Multimodal combining methods

One can replicate these results by using the default parameters in the library without any hyperparameter tuning.

Topic modeling

Topic modeling aims to identify different themes, entities and topics from a piece of text document/corpus to extract patterns depending on how they are touched on in a particular model.

Topic modeling can be challenging at times, especially with short responses like our survey responses. This can be taxing because short sentences:

- Lack context which leads to sparse data and makes a model unfit for semantic analysis
- Might require configuration which requires some domain knowledge
- May contain abbreviations and slangs
- Can cause overfitting

To resolve this issue, we made sure to diversify our topic modeling techniques and extract information using more than one method. Hence, we used Latent Dirichlet Allocation(LDA), BERTopic and Top2Vec for our analysis.

Baseline model

The baseline aims to address the sentiment of the response into positive and negative examples by training on the subjective data only. For the baseline model our approach is to fine tune the large language models to classify the responses. We chose both BERT and GPT-2 to fine tune our data on both the datasets to see if there are any differences in performance.

Following Hyper parameters were used in training:

Hyperparameter	Value
Batch size	16
Epochs	5
Learning rate	2e-5
Weight decay	0.01

Table 3 Hyperparameters for baseline training

All other parameters are left to default as per hugging face trainer function.

Quantitative Results

Baseline results:

Model	Class Balance	Validation loss	Validation accuracy
BERT	Unbalanced	0.246	0.89
	Balanced	0.327	0.91
GPT-2	Unbalanced	0.41	0.87
	Balanced	0.514	0.88

Table 4 Baseline results

Higher accuracy was observed in both datasets when using BERT as compared to GPT-2. However we observe that there is no major impact on performance by balancing the dataset.

Multimodal results

Combining Method	Accuracy Score	F1-score
MLP on all numerical features and then concat with transformer output	0.91	0.96
MLP on numerical feats then take weighted avg summation with transformer output	0.86	0.90

Table 5 Multimodal results

We observe no significant improvement over the baseline models.

For Topic Modeling

Below are the results from topic modeling using three approaches specified before:

Model	Topics generated	Topic themes
Top2Vec	2	the room provided for the training is very uncomfortable, Unnecessary time spend on mindfulness exercise - everyone has their own way/strategies for mindfulness - not just breathing exercises
BERTopic	5	Restraints_restraint_client_and, Videos_practice_controls_handout, debrief_weekly_debriefing_form, training_the_and_is, to_be_and_would
LDA	5	unit refreshments time strive minimize provided restraints team consuming water, need place waist concise better debrief could wrist practice consider, session debriefing shortern responsible class people training time hour condense, need training hands anchor low bed back strained staff floor

Table 6 Topic modeling results

Qualitative Results

Baseline

We deployed our models on HuggingFace portal and using the hosted interface API we tested out some of our own examples :

Examples of our model working :

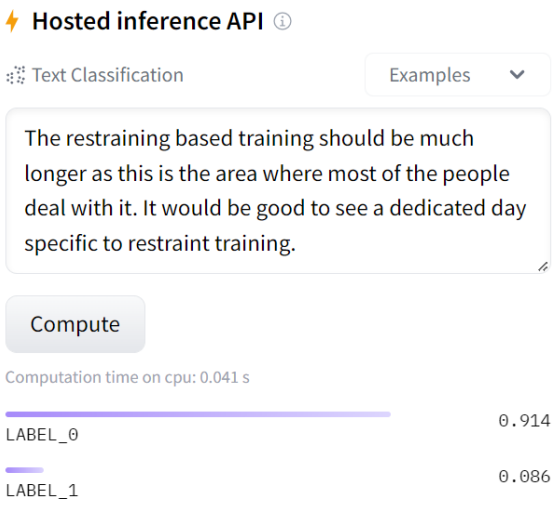


Figure 4 An input sentence with correct prediction

Examples where the model fails:

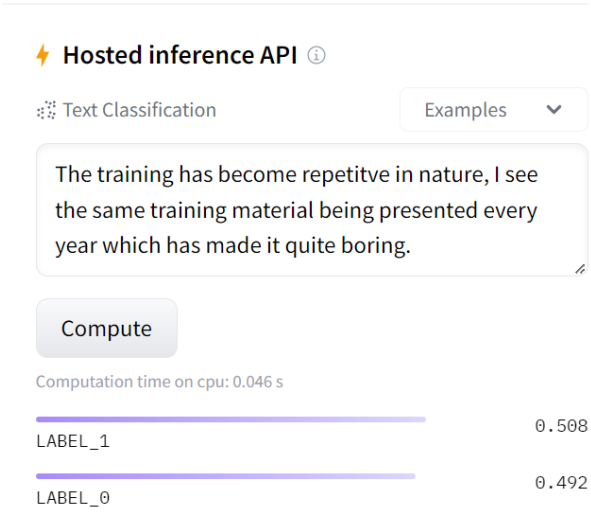


Figure 5 An input sentence where model fails borderline

One of the main reasons why we feel the model says as label 1 is lack of depth (sentiment). Had the user added some sort of suggestion the model prediction would have changed as shown below with a slightly modified example. This also relates to the few training examples where the model failed.

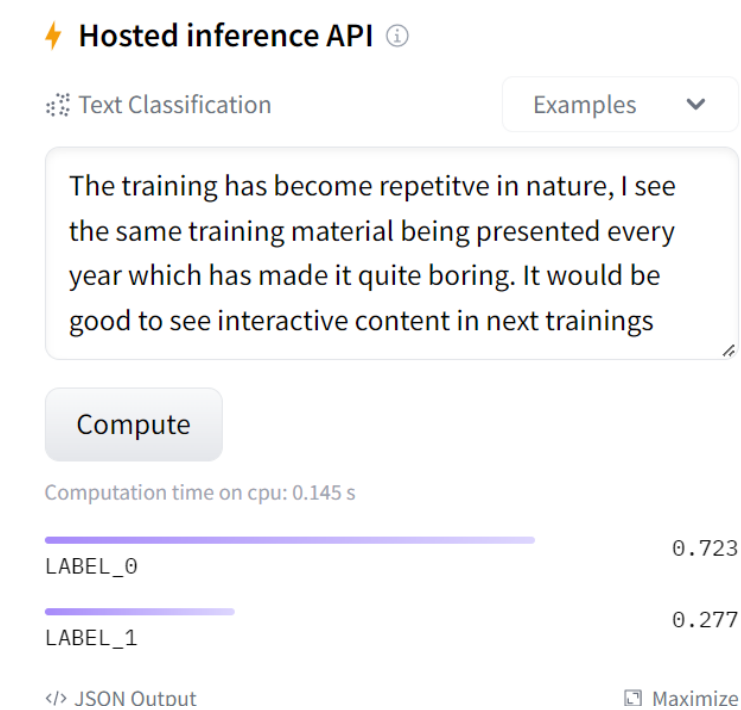


Figure 6 A modified input giving correct results

For Topic Modeling

Application of multiple Topic Modelling packages resulted in some common noticeable themes across the useful(negative) responses from the Survey Data. Observed were the following few themes across all three methods of topic modeling:

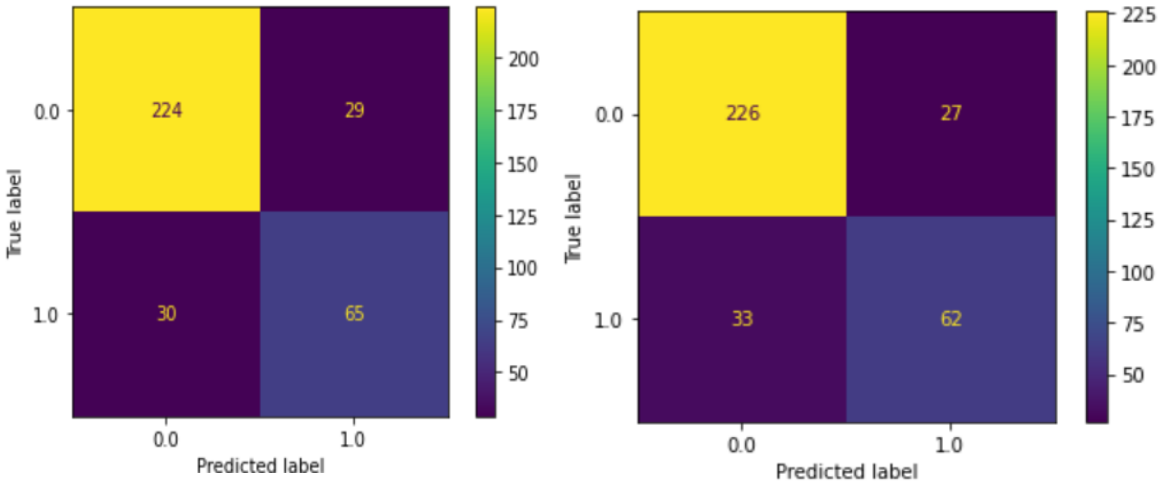
- Room provided was too small, hot and lots of external noise
- Training was too long and the pace was too slow
- There was no water provided during training
- Time provided to read training material was less
- Lack of hands-on training
- No hand-sanitizer was provided

Topic modeling was effective in providing some themes repeated across the whole textual data. Multiple topic modeling algorithms helped in elucidating the recurrent suggestion and responses from the data. However, because of responses being short and most of them lacking context makes topic modeling a challenging task.

For future projects topic modeling can be applied on just long sentences to observe any difference in results. Exploring other non-machine learning approaches could be effective in extracting recurring themes from the textual data.

Testing Baseline Model

As highlighted before we tested the baseline model using the nursing data and the results are as follows :



(a) **(b)**

Figure 7 Confusion matrix of fine tuned BERT on (a) Balanced and (b) Unbalanced dataset

Metric	Accuracy	Recall	F1	Precision
BERT on Balanced Dataset	0.83	0.68	0.68	0.69
BERT on Unbalanced Dataset	0.83	0.68	0.68	0.69

Table 7 Results of Fine tune BERT on balanced and unbalanced dataset

A decent performance on the nursing dataset though we have to keep in mind that it was fine tuned on TIDES.

Discussion and Learnings

Our discussions are summarized below:

- Baseline model performs satisfactorily as evident from the metrics used when training/testing. Testing the model out on additional data gave us reinforcement on the same.
- There was no significant performance improvement even after the numerical data was added. A reason could be the lack of variance between the pre and post scores In addition, the lack of data could also be a factor.

Our learnings are summarized below :

- Working with short sentences with Topic Modeling is challenging. We would try some simple non-ML based approaches as a workaround as well.
- We were unable to debug much of the multimodal model results (final layers) as the library was built on top of HuggingFace version 3.1 (Current Version of HF-4.24) and is no longer maintained.
- Focus more on quality and quantity of data

Individual Contributions

Karan's contribution:

- Manually labeling of the primary i.e. TIDES dataset
- Responsible for coding and training the baseline model on textual data
- Augmented the dataset for topic modeling
- Performed Topic modeling and carried out the analysis

Sai's contribution:

- Manually labeling of Nursing Dataset
- Tested out baseline models on Nursing Dataset
- Responsible for the multimodal architecture by utilizing the multimodal toolkit
- Summarized results for the baseline and final model

References

[1] Ostendorff, Malte, Peter Bourgonje, Maria Berger, Julian Moreno-Schneider, Georg Rehm, and Bela Gipp. "Enriching bert with knowledge graph embeddings for document classification." *arXiv preprint arXiv:1909.08402* (2019).
[2] Georgian. (2021, February 26). *How to incorporate tabular data with Huggingface Transformers*. Medium. Retrieved October 31, 2022, from https://medium.com/georgian-impact-blog/how-to-incorporate-tabular-data-with-huggingface-transformers-b70ac45fcfb4
[3] Huang, Xin, Ashish Khetan, Milan Cvitkovic, and Zohar Karnin. "Tabtransformer: Tabular data modeling using contextual embeddings." *arXiv preprint arXiv:2012.06678* (2020).
[4] *Multimodal Transformers Documentation*¶ (no date) *Multimodal Transformers Documentation - Multimodal Transformers documentation*. Available at: https://multimodal-toolkit.readthedocs.io/en/latest/ (Accessed: October 31, 2022).