Mohamed Abdelwahab 1009200339
Jiading Zhu 1002445874

Word Count: 1998
Penalty: 0%

Final Report: MICon

Introduction

Smoking is a very common habit on a worldwide scale and it leads to the death of 8 million people every year. [1] Motivational Interviewing (MI) is a psychotherapy method that is used to resolve ambivalence and encourage change, and it is used with smokers, as well as alcoholics and drug addicts, to help them to decide to quit their unhealthy habits. The sentences within a MI session are coded using the Motivational Interviewing Skills Code (MISC) to determine their MI-consistency. [2] An example of MI-inconsistent sentences is "you should try to quit" and is given an *Advise Without Permission (ADW)* code. An MI-consistent alternative would be: "If you don't mind my asking, have you considered quitting this habit?", which is coded with *Advise With Permission (ADP)*.

The task of determining the MI-consistency of a sentence presents a suitable challenge for large language models. In this project, we are developing two classifiers, the first of which is an end-to-end classifier that determines the MI-consistency of the last response of a therapist given a partial conversation, and the second classifies the MISC code of the last response, as illustrated in Figures 1 & 2. The ultimate end goal of this project is to use these classifiers in generative chatbots that deliver MI-styled conversations, which serves as a cheaper alternative to talking to a human therapist.

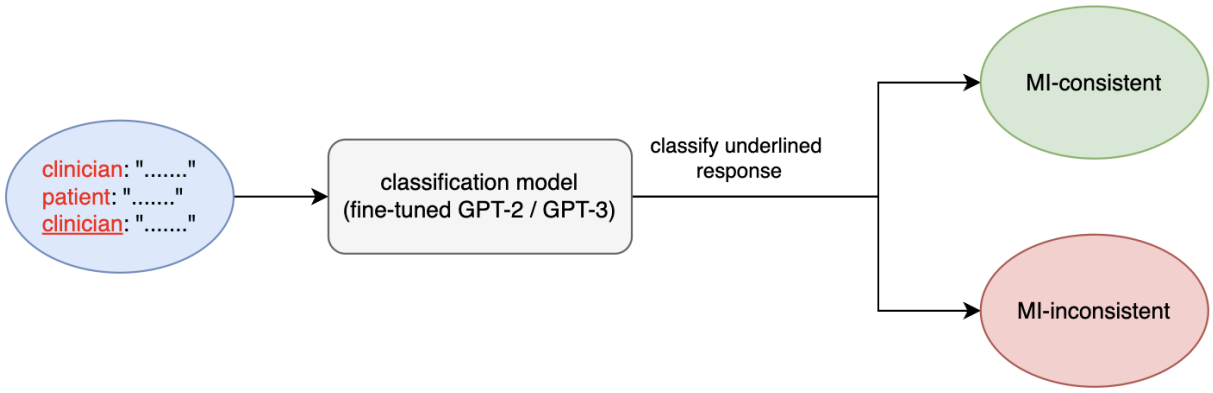

Figure 1: End-to-End MI-consistency Classifier

Background & Related Work

Research is being done into automating MI coding. The authors of [3] present a preliminary study in using computers to replace humans for coding therapy sessions for inspecting MI fidelity. They are concerned that manual coding cannot be scaled up with the huge amount of time spent in therapy

sessions. They use labeled topic models to cluster words and phrases and relate them to each MI code given by MISC. Results show that their proposed method is not reliable when coding individual sentences (called utterances) but is comparable to human coding in case of session-level coding.

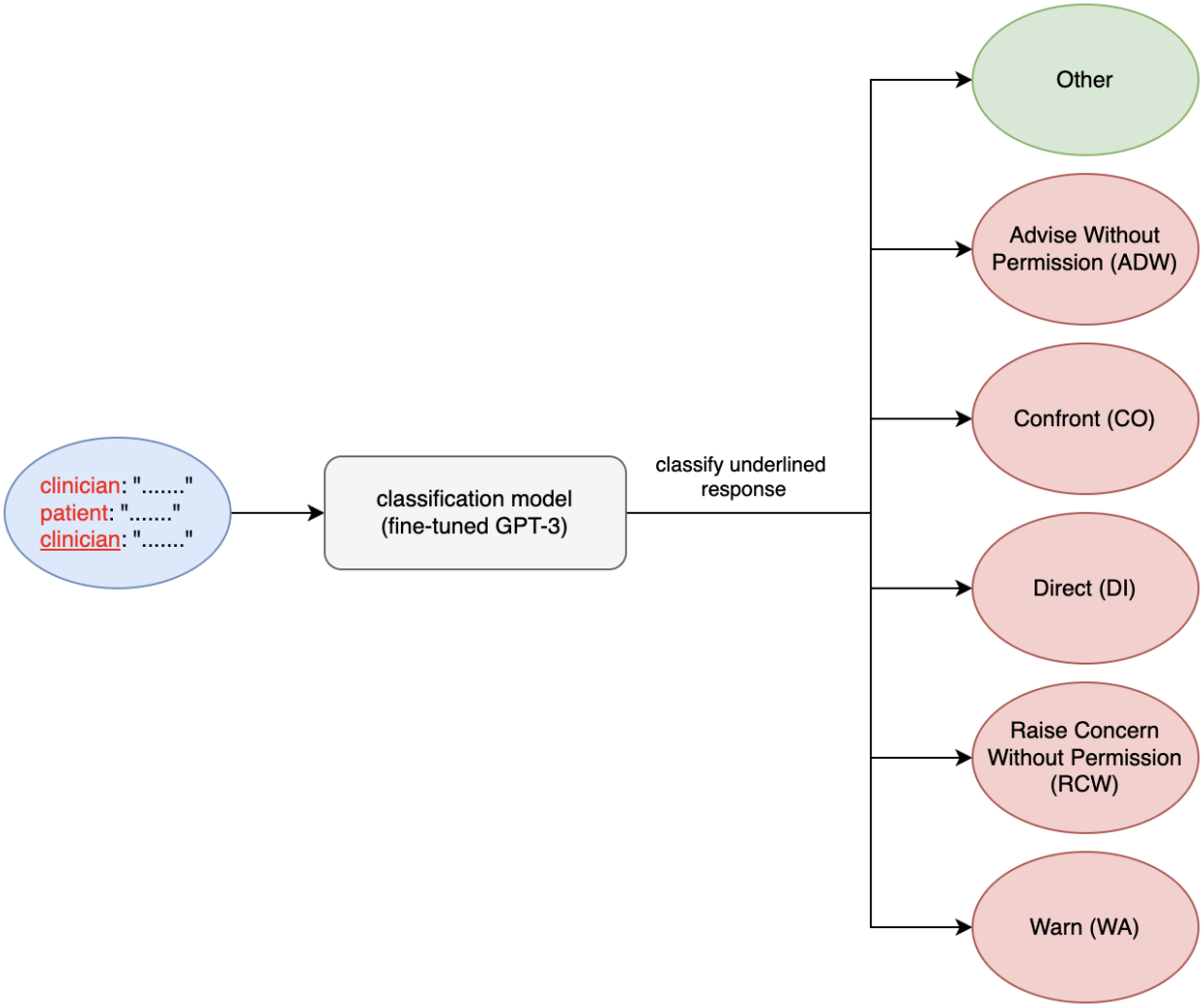

Figure 2: MISC-code Classifier

Natural Language Processing methods are used in [4] to automate MI coding. The authors are also concerned with the number of resources required for manually coding therapy sessions. They propose using 2 methods, the first of which is a Discrete Sentence Feature (DSF) model, which uses dependency trees and N-grams, along with the actual utterance, as inputs to a Multiclass classifier for determining the code that this sentence abides by. The second method uses a Recurrent Neural Network (RNN) in combination with the GloVe embeddings. The embeddings of a sentence are fed into the RNN, the output of which is then fed to a multiclass classifier to determine its code. Their results for utterance-level testing show that both models exhibit comparable performance to human coding for most codes. Their session-level results are much closer to human coding than utterance-level results. The results also show that for all codes, DSF's performance is either similar to or better than the RNN's.

Data and Data Processing

Reflection Triplets

Our initial choice for MI-consistent data was a dataset from our own research group. This dataset is composed of triplets, in which the therapist (in the case of our research group, a chatbot) initiates with a sentence, then the patient replies, and then finally the therapist provides a response to that reply. Since the purpose of the dataset was to train a classifier for reflections, all of the responses provided by therapists were of the MISC coding of Reflect (RE), including both Simple Reflections (RES) and Complex Reflections (REC).

Due to lack of variety within the MI-consistent MISC codings, along with suboptimal results from preliminary experiments, we eventually decided to pivot to another dataset instead. However, we persisted with the triplet format for subsequent datasets, since it offers the model context from both the therapist and the patient which is an essential deciding factor for the MISC codings.

AnnoMI

AnnoMI is a dataset from Kaggle which contains transcripts of whole therapy sessions, with each session being labeled as being of high-quality or low-quality MI. [5] There are about 10,000 rows of text dialogues in the dataset, leading to having diversity in the MISC codes of the therapist's responses.

For MI-consistent examples, we extracted triplets from the high-quality session transcripts. We manually filtered triplets to only keep the ones where the therapist's response makes sense in the MI context. For each triplet, we labeled the therapist's response with its corresponding MISC code. When the therapist's response contains more than one MISC codings, we split the response into two parts, and we fill the missing parts with a natural sentence that does not conflict with the current context. We collected MI-inconsistent triplets from the low-quality session transcripts in the same way.

In total, we collected 118 MI-consistent triplets and 84 MI-inconsistent triplets from AnnoMI, each MISC code having 10-30 triplets. The limiting factor is the MI-inconsistent ones, due to the fact that only 9% of the AnnoMI dataset are low-quality sessions, and the 84 triplets are all we got from going through every single low-quality session transcript. To make our dataset more balanced, we resorted to using an additional source of data for more MI-inconsistent examples.

Reddit

To gather additional MI-inconsistent examples, we parsed subreddits using Python's PRAW library for speech that are judgemental, confrontational, or even abusive in nature. We focused on 4 subreddits in particular: *r/gaslighting*, *r/Manipulation*, *r/abusiverelationships* and *r/abusiveparents*, and iteratively parsed through every post and comment in each subreddit. We manually picked speech text that could appear in a therapy session, and formed it into a triplet by using the chosen sentence as the therapist's final response and making up the missing parts appropriately. We collected an additional 20 triplets from this approach, reaching a total of 104 MI-inconsistent triplets.

Baseline Model

For our baseline models, we are using OpenAI's "Davinci-003" GPT-3 model in a zero-shot setting. For the end-to-end model, the prompt is formatted in the following way:

> *Decide if the final sentence said by the clinician in the following dialogue is adherent to Motivational Interviewing practices, in TRUE or FALSE:*
> *Clinician: ……*
> *Patient: ……*
> *Clinician: ……*

The expected output is a singular English word in all-lower case, either "true" or "false".

For the MISC classifier model, the prompt is instead the following:

> *Classify the final sentence said by the clinician into the following classes: ADW (advice without permission), CO (Confront), DI (direct), RCW (raise concern without permission), WA (warn), or other:*
> *Clinician: ……*
> *Patient: ……*
> *Clinician: ……*

The expected output are the five MI-inconsistent MISC codings in their abbreviations, or a singular English word "other", in all-lower case, as shown in Figure 2.

The model is evaluated by querying OpenAI's Python API while feeding in the same test data prepared for our fine-tuned GPT-3 end-to-end and MISC classifier models.

Architecture and Software

Our best performing models including 2 end-to-end models and 1 MISC classifier model:

1. A fine-tuned end-to-end GPT-2 model (1.5 billion parameters, retrieved from Hugging Face)
2. OpenAI's "Davinci-003" GPT-3 model, fine-tuned as an end-to-end model
3. OpenAI's "Davinci-003" GPT-3 model, fine-tuned as a MISC classifier

For the MISC classifier model, we made the design decision of only classifying the MI-inconsistent codes, since there exist a lot more MI-consistent codes and not enough data per code.

Similar to the baseline models, the fine-tuned GPT-3 models are evaluated by querying OpenAI's Python API. Following OpenAI's guidelines, the GPT-3 models used in a classification context performs best when setting *max_tokens=1* and *temperature=0*, which is how we performed our evaluations.

Data Preparations

For the GPT-2 model, the input only includes the Clinician-Patient-Clinician triplet, tokenized by GPT-2's tokenizer. For the GPT-3 models, the input data is prepared using OpenAI's CLI data preparation tool into JSONL format. [6] In particular, the following operations were performed:

1. The data is put in a dictionary with two keys: "prompt" and "completion", where the "prompt" values are the corresponding prompts for each model, and "completion" are the corresponding expected outputs;
2. The expected completions are converted to all-lower cases;
3. A suffix separator "\n\n###\n\n" is appended to all prompts;
4. A whitespace character is added to the beginning of all completions;
5. The dataset is split into a train-test split of 4-1 ratio stratified.

End-to-End Model Results

The baseline model was only able to reach an accuracy of 51.1%. We observed that it classified nearly all examples as being MI-inconsistent, even for the examples that were clearly MI-consistent by MISC standards. For the fine-tuned GPT-2 model, it achieved a classification accuracy of 62.2%. In contrast, we noticed that this model classified most examples as being MI-consistent, even for the examples that were clearly MI-inconsistent.

As for the fine-tuned GPT-3 model, it outperformed the two other models, with an accuracy of 91.1%. Nearly all of the examples that it misclassified were ambiguous in nature, as the one shown below:

> ***Clinician***: *"So you're doing more than that right now. And when you drink ..."*
> ***Patient***: *"Okay."*
> ***Clinician***: *"And, uh-- so I'm going to recommend that you cut down- cut down that amount to the recommended limit. Again, that's no more than seven drinks a week and no more than three on a given day. Do you think you could do that? Would that be a good goal for you?"*

We coded this example as being MI-inconsistent because at the beginning of the response the clinician was advising without permission. However, the questions at the end make it seem like the clinician is emphasizing the patient's control over their decisions, so one could make an argument for this sentence being MI-consistent instead.

MISC Classifier Results

The MISC classifier baseline model achieved an accuracy of 22.2%. Using the results for this classification as a binary classifier for MI-consistency led to a 70.4% binary classification accuracy. The fine-tuned GPT-3 model achieved a 63.0% accuracy, and when used as a binary classifier, the model achieved a 92.6% accuracy. Below is an example of a misclassification using this model:

> ***Clinician**: "Well, to be honest with you, you mentioned the patch. I can't recommend a patch …"*
> ***Patient**: "Well, I mean, I s-- I-I-I smoke barely a pack a day. I mean, I've probably …"*
> ***Clinician**: "Okay. Well, we actually consider 20 cigarettes a day to be a significant amount, and in fact, I would then recommend to you that you would start on the highest dose of nicotine replacement."*

The model classified this example as *Confront (CO)*, possibly based on the first part of the response disagreeing with the patient. We labeled this example as *Advise Without Permission (ADW)* because of the second part. Such misclassifications could potentially be alleviated with a larger and more accurately labeled training set.

Discussion and Learnings

Overall, the fine-tuned models meet our expectations by reaching satisfactory performance. Results for the end-to-end model experiments show that a zero-shot GPT-3 does not possess a good understanding of MI. Comparing the performances of fine-tuned GPT-2 vs GPT-3 suggests that the bigger the model, the more powerful it could be. Using the MISC classifier as a binary classifier achieves higher accuracy than the end-to-end models, but no definitive conclusions could be made due to the small size of the test size.

It is interesting to learn that GPT-3 is not omnipotent out-of-the-box, but can be quite powerful after fine-tuning. In future projects, we would spend more time on data collection and labeling, especially for a multi-class classification task.

Individual Contributions

Mohamed retrieved the AnnoMI dataset and extracted and hand-labeled the 118 MI-consistent examples with their MISC codes. He was also responsible for fine-tuning the GPT-2 model for the end-to-end classifier, and wrote an evaluation script for end-to-end GPT-3 models.

Jiading hand-labeled 84 MI-inconsistent examples with MISC codes from AnnoMI, and wrote scripts to parse data from Reddit and hand-picked 20 more MI-inconsistent examples. He also fine-tuned GPT-3 into end-to-end and MISC classifier models, prepared datasets into allowed formats, and wrote an evaluation script for the MISC classifier models.

References

[1] “Tobacco,” World Health Organization. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/tobacco [Accessed: 21-Nov-2022].
[2] Miller WR, Moyers TB, Ernst DB, Amrhein PC: Manual for the Motivational Interviewing Skill Code (MISC), Version 2.1. New Mexico: Center on Alcoholism, Substance Abuse, and Addictions, The University of New Mexico, 2008.
[3] Atkins et al.: Scaling up the evaluation of psychotherapy: evaluating motivational interviewing fidelity via statistical text classification. Implementation Science 2014 9:49.
[4] Tanana et al. “A comparison of natural language processing methods for automated coding of motivational interviewing.” Journal of substance abuse treatment 65 (2016): 43-50.
[5] “AnnoMI,” RAHULBABURAJ. [Online]. Available: https://www.kaggle.com/datasets/rahulbaburaj/annomi [Accessed: 12-Dec-2022].
[6] “Fine-tuning,” OpenAI [Online]. Available: https://beta.openai.com/docs/guides/fine-tuning/cli-data-preparation-tool [Accessed: 12-Dec-2022].

Permissions

	Post video	Post final report	Post source to code
Mohamed Abdelwahab	yes	yes	no
Jiading Zhu	yes	yes	no