Utilizing Open Source Language Models and ChatGPT for Zero-Shot Identification of Drug Discontinuation Events in Online Forums

Supplementary Material

Contents

1 MATERIALS AND METHODS - CONTINUED **1**

1.1 Data Labeling Challenges 1

1.2 Motivation behind ***Classification Strategy (1)***, ***Classification Strategy (2)***, and ***Classification Strategy (3)*** 4

1.2.1 ***Classification Strategy (1)*** 4

1.2.2 ***Classification Strategy (2)*** 5

1.2.3 ***Classification Strategy (3)*** 6

1.3 ChatGPT prompt design 6

1.3.1 ChatGPT prompt used in CS1 and CS2 6

1.3.2 ChatGPT prompt used in CS3 7

2 RESULTS - CONTINUED **9**

2.1 Expanded Discussion Of The False Positive Rates (FPR) And False Negative Rates (FNR) 9

3 REFERENCES **10**

1 MATERIALS AND METHODS - CONTINUED

1.1 Data Labeling Challenges

In this project's initial stages, we attempted to outsource data labeling to Amazon Mechanical Turk (MTurk) before involving three graduate students to label the 1000 comments. MTurk is a "crowdsourcing marketplace" that facilitates outsourcing jobs or processes to a virtually distributed workforce, as described by Amazon [1]. For our DDE labeling tasks, MTurk workers interacted with an interface displayed in Figure 1.

When creating an MTurk task, you need to specify several elements:

1. **$ Pay / Task:** The payment you offer to each worker for task completion. In our case, labeling one comment.

2. **# Workers / Task:** The number of different workers required to complete each task, or the number of unique labelers for each comment in our context.

3. **Worker Requirements (optional):** Any prerequisite qualifications that workers must fulfill to be eligible for your task, such as a high school diploma, or a 95% task acceptance rate on their previous MTurk work.

While MTurk offers quick and cost-effective data gathering, recent studies highlight data quality concerns with this approach [2]. These include labeler honesty and attentiveness issues, and a high prevalence of bots among MTurk "workers" [3–5]. Despite these concerns, we decided to assess the accuracy level at which MTurk workers could distinguish between DDE and non-DDE comments posted on medhelp.org, given MTurk's data procurement speed.

To begin this analysis, we chose 250 medhelp.org comments and had two internal labelers classify each. Of these 250 comments, both labelers agreed on 216. We then defined a task in MTurk using the interface shown in Figure 1 for each of these 216 comments. We conducted three different trials where all 216 comments were labeled by the # Workers / Task indicated in Table

Trial	Worker Req.	$ / Task	# Workers / Task	Overall Acc.	VC Acc.	SC Acc.	NC Acc.	DDE Acc.	non-DDE Acc.
1	None	0.02	3	54.3%	67.4%	49.4%	84.2%	45.6%	82.2%
2	85%	0.03	3	61.1%	66.9%	39.0%	65.8%	41.6%	84.8%
3	95%	0.02	5	58.7%	72.8%	49.4%	57.6%	44.4%	87.3%

Table 1: The parameters selected for the three MTurk trials and the accuracy results from each trial. "Worker Req." refers to the required level of acceptance that workers must have had on previous tasks that they performed on MTurk to be eligible to perform our task; "$ / Task" indicates the amount a labeler is paid for labeling a single comment; "Overall Acc." is the overall accuracy of the MTurk workers' labels; "VC Acc.", "SC Acc.", and "NC Acc." are the accuracy rates of the workers' labels when they selected "Very Confident", "Somewhat Confident", and "Not Confident" respectively; "DDE Acc." and "non-DDE Acc." are the accuracy rates of the labelers when selecting "DDE" and "non-DDE" respectively.

1. The payment, worker requirements, and MTurk workers' accuracy results for each trial are also presented in Table 1.

As shown by the results in Table 1, in all three trials, the labelers agreed with us over 80% of the time when labeling a comment as non-DDE, yet they concurred with us only 40% - 46% of the time when labeling a comment as DDE. These results suggest that the MTurk labelers might have paid less attention or applied a less strict DDE definition despite being given the same instructions as the graduate students who labeled 1000 comments. While the MTurk labelers were more likely to agree with our internal labelers when classifying comments as non-DDEs, it's worth noting that this may be due to the relatively low occurrence rate of DDEs in the labeled comments (approximately 10%).

Following our experience with MTurk, we ensured the three graduate students clearly understood the criteria for a comment to be considered a DDE as outlined in Section 4.3 of the main paper, and we ensured that they agreed with each other more than 90% of the time before labeling the 1000 comments from medhelp.org. We performed multiple practice rounds of labeling and review. Despite these precautions, as illustrated by the example posts in Figure 1, some medhelp.org comments contained spelling and grammatical errors, or acronyms, which could pose interpretation or labeling challenges for those unfamiliar with medhelp.org.

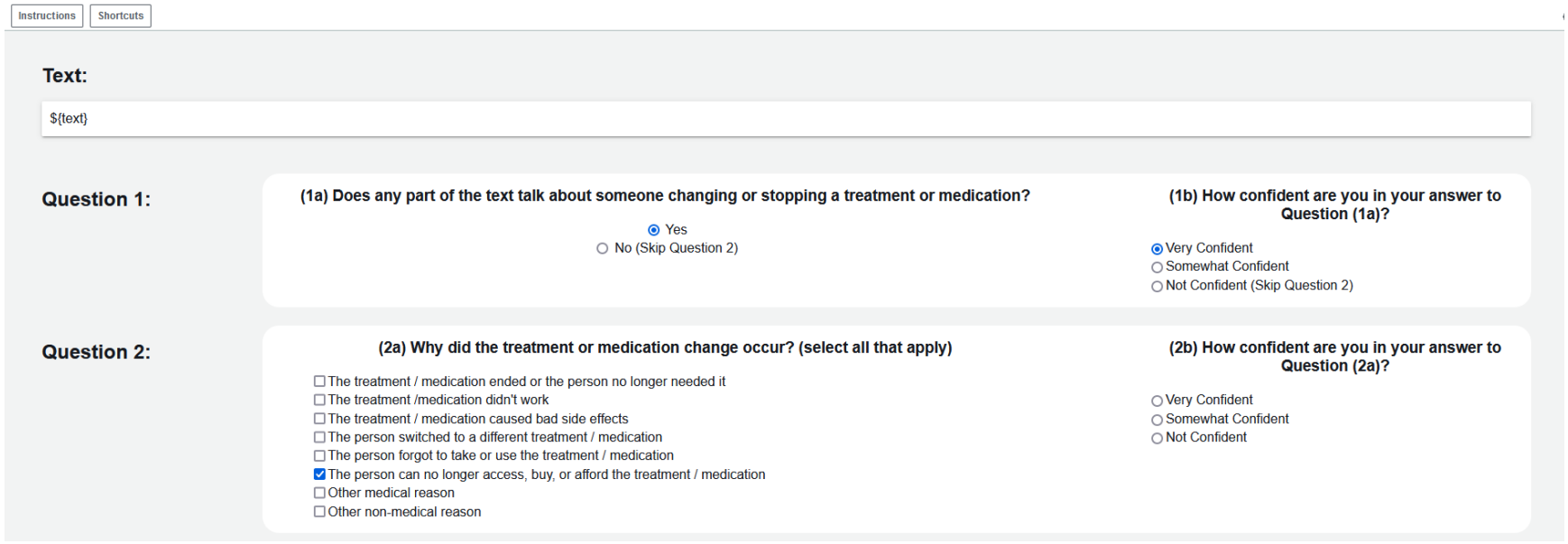

Figure 1: The data labeling interface utilized by labelers hired from MTurk

1.2 Motivation behind *Classification Strategy (1)*, *Classification Strategy (2)*, and *Classification Strategy (3)*

The maximum token input length for NLI-DeBERTa-Base and DistilBERT-Base-Uncased-MNLI is 512 tokens, for RoBERTa-Large-MNLI and NLI-DistilRoBERTa-Base it is 514 tokens, for BART-Large-MNLI it is 1024 tokens, for GPT-3.5 Turbo it is 4096 tokens, and for GPT-4 it is 8192. It is worth noting that an even larger GPT-4 model exists, GPT-4-32k, which can support a context up to 32768 tokens [6].

1.2.1 *Classification Strategy (1)*

In CS1, by classifying each sentence individually, we hypothesized that the models may be able to more effectively classify questions or answers which are long but which only briefly mention a DDE. In such a case, we hypothesized that if a long question or answer largely talked about things not related to a DDE but briefly mentioned a DDE somewhere in the text, the models may have a hard time classifying such questions or answers as DDEs due to the "noise" throughout the text. By running the model on each individual sentence in a comment, then if any sentence discusses a DDE, other irrelevant sentences would not hinder the ability of the models to detect the DDE as the maximum model output for any sentence is taken to be the "model prediction" under CS1. However, by passing each sentence into the model individually, each sentence is taken out of context of the larger comment. Therefore, if a question or answer describes a DDE in an indirect way throughout many sentences, using CS1, the models may not be able to detect the DDE since each sentence is taken out of context of the larger comment. An example of how this strategy is implemented is shown in Figure 4. In the computation results in Section 2, Columns with a (1) such as "DeBERTa (1)" indicate that CS1 was utilized.

1.2.2 *Classification Strategy (2)*

Next, to address situations where a DDE may be described in an indirect way throughout many sentences, we developed CS2 to explore if some models might perform better if they had access to the entire context of the comment at once, or at least multiple consecutive sentences concatenated together. In this way, we developed CS2 to evaluate how well the models would perform if we passed in multiple sentences (potentially the entire comment) as single blocks of text. For each model we broke each comment into groups of consecutive sentences which were as large as possible while staying under the maximum token input length of the model. For example, consider a question / answer comprised of five sentences, each of which is 250 tokens in length when tokenized by DistilBERT-Base-Uncased-MNLI's tokenizer. Since DistilBERT-Base-Uncased-MNLI's max input token length is 512 tokens, we would group the five sentences into three groups. Group one would consist of the first and second sentences (500 tokens in length), group two would consist of the third and fourth sentences (500 tokens in length), and group five would consist of the fifth sentence (250 tokens in length). Here, we have three groups, each of which is less than the max number of tokens for the model, and the first two are as large as they could be since adding the next consecutive sentence to the group would cause the total length to exceed the 512 token limit. In our method, the groups are defined one at a time (group one is formed, then group two, then group three, etc.), and each group will be comprised of as many consecutive sentences as possible while maintaining a total token length less than the maximum length for the model. More examples showing how this strategy is implemented are shown in Figures 5 and 6. In the computation results in Section 2, Columns with a (2) such as "DeBERTa (2)" indicate that CS2 was utilized.

It is worth noting that the maximum token input length varies significantly across the models we tested, ranging from 512 tokens to 8192 tokens. Specifically, the maximum token input length for NLI-DeBERTa-Base and DistilBERT-Base-Uncased-MNLI is 512 tokens, for RoBERTa-Large-MNLI and NLI-DistilRoBERTa-Base it is 514 tokens, for BART-Large-MNLI it is 1024 tokens, for GPT-3.5 Turbo it is 4096 tokens, and for GPT-4 it is 8192. Although the base GPT-4 model offered by OpenAI was sufficient for our research, an even larger GPT-4 model is also offered by

OpenAI, GPT-4-32k, which can support a context up to 32768 tokens [6].

1.2.3 *Classification Strategy (3)*

While the HuggingFace models naturally output a probability of entailment for each Premise, carefully curated prompts were designed for the ChatGPT models to reliably obtain a structured response which could be parsed to obtain a probability of entailment to allow for a direct comparison with the HuggingFace models using CS1 and CS2. Furthermore, since ChatGPT does not naturally output a probability of entailment like the HuggingFace models do, we hypothesized that asking the model to give a probability of entailment might not align well with its inherent text generation functionality. Instead, we hypothesized that a binary classification task where the model is explicitly asked to predict whether a text does or does not describe a DDE could be a better fit for the capabilities of ChatGPT, and potentially yield better performance. As a result, we developed CS3 to test this hypothesis on the GPT-3.5 Turbo and GPT-4 models specifically.

1.3 ChatGPT prompt design

As breifly discussed in the main document of this manuscript, GPT models including GPT-3.5-Turbo and GPT-4 are sequence-to-sequence models meaning that the output is narrative, and as a result, prompts for GPT models must be carefully curated in order to obtain results in consistent formats.

1.3.1 ChatGPT prompt used in CS1 and CS2

The following prompt was provided to both GPT-3.5-Turbo and GPT-4 when using CS1 and CS2:

Your task as an AI model is to determine the probability that a given comment from medhelp.org entails a "Drug Discontinuation Event".

The concept of a "Drug Discontinuation Event" involves a specific individual stopping a recurring medication or treatment. It includes instances where the individual has switched from one

medication to another, but not one-time treatments.

To formulate your task, consider each comment as a premise and the statement "A person stopped taking a medication" as the hypothesis. Your goal is to estimate the probability that the hypothesis is true given the premise. Express this probability as a percentage.

For example, if a comment strongly implies a Drug Discontinuation Event, you might respond with "Probability of entailment: 95%". Conversely, if a comment does not suggest a Drug Discontinuation Event, you might respond with "Probability of entailment: 5%".

Examples:

1. Comment: "I stopped taking my birth control medication." Response: "Probability of entailment: 100%"

2. Comment: "I took the plan B pill yesterday." Response: "Probability of entailment: 5%"

3. Comment: "I changed my birth control medication because of side effects." Response: "Probability of entailment: 95%"

4. Comment: "I stopped taking my birth control medication because I was feeling worse, but then I started taking it again." Response: "Probability of entailment: 100%"

Next, I will provide a comment, and you will estimate the probability of entailment as instructed. Remember, the output should be in the format of "Probability of entailment: X%" where X is the estimated probability.

1.3.2 ChatGPT prompt used in CS3

The following prompt was provided to both GPT-3.5-Turbo and GPT-4 when using CS3:

As an AI model, your task is to classify each comment from medhelp.org into one of two categories based on the content of the comment. The categories are: "Drug Discontinuation Event" (1) and "Non-Drug Discontinuation Event" (0).

A “Drug Discontinuation Event” (1) is any instance where it can be deduced from the comment that a specific individual has stopped a recurring medication or treatment. This includes cases where the individual has switched from one medication to another. It does not include one-time treatments.

A “Non-Drug Discontinuation Event” (0) is any instance where it cannot be inferred from the comment that a specific individual has stopped a recurring medication or treatment.

Examples:

1. Comment: “I stopped taking my birth control medication” Response: “1”
2. Comment: “I took the plan B pill yesterday” Response: “0”
3. Comment: “I changed my birth control medication because of side effects” Response: “1”
4. Comment: “I stopped taking my birth control medication because I was feeling worse, but then I started taking it again” Response: “1”

Next, I will give you a comment, and you will classify it according to these instructions. Remember to only respond with “1” or “0”.

2 RESULTS - CONTINUED

2.1 Expanded Discussion Of The False Positive Rates (FPR) And False Negative Rates (FNR)

Upon further analysis of the FPR and FNR results, it is apparent that for all the models except GPT-4 at cutoffs 0.9 and 0.95, the FNR is *higher* under CS2 than in CS1. Conversely, the FPR is *lower* under CS2 than in CS1. This phenomenon could be due to the possibility that feeding multiple sentences simultaneously may cause comments that merely mention cessation of medication within a larger text to be classified as non-DDE ("Negative"). This might suggest that the additional contextual information provided to the models in CS2 may result in some loss of sensitivity towards the detection of DDEs. However, the trade-off is that the models may also be less prone to mistakenly classifying non-DDE comments as DDEs.

However, the result that GPT-4 achieved a lower false negative rate under CS2 than CS1 at cutoffs of 0.9 and 0.95 is highly intriguing as it stands in contrast to the phenomonenon exhibited by all other models, including GPT-3.5 Turbo. Further investigation would be needed to better understand this phenomenon, but our hypothesis is that GPT-4 is more highly capable of handling complex and detailed contexts than the other models, and by using CS2, GPT-4 can leverage the entire content of the comment to make a more "informed" classification. It is also interesting that the false negative rate was only higher for GPT-4 under CS2 than CS1 at the two highest cutoff values we tested, 0.9 and 0.95. Further investigation is required to better understand this phenomenon, which may be an outlier or a consequence of the sample size of this study.

Analyzing the results obtained using CS3, we can see that the performance of GPT-4 under CS3 is slightly underperformed compared to GPT-4 under CS2 with a cutoff of 0.75. Specifically, GPT-4 under CS2 with a cutoff of 0.75 obtained an overall accuracy of 86.9%, a FPR of 12.654%, and a FNR of 16.822%, all of which are slightly better than those obtained using GPT-4 with CS3. This result challenges our initial hypothesis that the GPT-3.5 Turbo and GPT-4 models would perform

better under the binary classification task considered in CS3. Consequently, further investigation into how to best design a prompt for the ChatGPT models is needed to explore how the performance of these models could be further improved.

3 REFERENCES

[1] "Amazon Mechanical Turk — mturk.com." `https://www.mturk.com/`. [Accessed 26-Apr-2023].

[2] Z. Shakeri Hossein Abad, G. P. Butler, W. Thompson, and J. Lee, "Crowdsourcing for machine learning in public health surveillance: Lessons learned from amazon mechanical turk," *J Med Internet Res*, vol. 24, p. e28749, Jan 2022.

[3] A. Saravanos, S. Zervoudakis, D. Zheng, N. Stott, B. Hawryluk, and D. Delfino, "The hidden cost of using amazon mechanical turk for research," in *HCI International 2021 - Late Breaking Papers: Design and User Experience* (C. Stephanidis, M. M. Soares, E. Rosenzweig, A. Marcus, S. Yamamoto, H. Mori, P.-L. P. Rau, G. Meiselwitz, X. Fang, and A. Moallem, eds.), (Cham), pp. 147–164, Springer International Publishing, 2021.

[4] M. A. Webb and J. P. Tangney, "Too good to be true: Bots and bad data from mechanical turk," 2021.

[5] H. Aguinis, I. Villamor, and R. S. Ramani, "Mturk research: Review and recommendations," *Journal of Management*, vol. 47, no. 4, pp. 823–837, 2021.

[6] "OpenAI Platform — platform.openai.com." `https://platform.openai.com/docs/models`. [Accessed 23-Jul-2023].