Explainable Depression Detection Using Large Language Models on Social Media Data

Yuxi Wang, Diana Inkpen, Prasadith Buddhitha
School of Electrical Engineering and Computer Science
University of Ottawa
{ywan1225, diana.inkpen, pkiri056}@uottawa.ca

Abstract

Due to the rapid growth of user interaction on different social media platforms, publicly available social media data has increased substantially. The sheer amount of data and level of personal information being shared on such platforms has made analyzing textual information to predict mental disorders such as depression a reliable preliminary step when it comes to psychometrics. In this study, we first proposed a system to search for texts that are related to depression symptoms from the Beck's Depression Inventory (BDI) questionnaire, and provide a ranking for further investigation in a second step. Then, in this second step, we address the even more challenging task of automatic depression level detection, using writings and voluntary answers provided by users on Reddit. Several Large Language Models (LLMs) were applied in experiments. Our proposed system based on LLMs can generate both predictions and explanations for each question. By combining two LLMs for different questions, we achieved better performance on three of four metrics compared to the state-of-the-art and remained competitive on the one remaining metric. In addition, our system is explainable on two levels: first, knowing the answers to the BDI questions provides clues about the possible symptoms that could lead to a clinical diagnosis of depression; second, our system can explain the predicted answer for each question.

1 Introduction

Being one of the leading global public health issues, depression is common, costly, debilitating, and associated with an increased risk of suicide (Marwaha et al., 2023). Since depression has become a prevalent mental health issue, early detection of symptoms could greatly improve the chances of proper treatment. Traditional methods of detection, usually human-led, are expensive to conduct and might be individually biased. In this study, we propose a method to analyze and select social media writings to help identify potential symptoms of depression. Then, we propose an explainable method that uses the selected writings to automatically fill in the Beck's Depression Inventory (BDI) questionnaire (Beck et al., 1961) for the social media user (see Figure A1 for the full questionnaire). The questionnaire then provides the level of depression of the user based on all the answers.

The main contributions of this paper are:

1. Extended the applicability of using Large Language Models (LLMs) to predict mental health status for social media users.

2. Improved the performance on the task of automatically filling in the BDI questionnaire using social media data through manually designed prompts and without further training.

3. Explored the use of LLMs for generating both the predictions and explanations for the predictions.

2 Related Work

To develop computational methods for depression detection using textual information, analyzing word usage became a natural starting point. Through statistical investigation, researchers found that negative emotion, cause, sensory, and the first person singular words were more commonly used when describing activities such as breakup (Boals and Klein, 2005). Linguistic Inquiry and Word Count (LIWC), a computerized text analysis tool, was developed to assess word usage in psychologically meaningful categories (Tausczik and Pennebaker, 2010). The tool was built by creating dictionaries from domain knowledge, with the words categorized into different groups.

In addition to social and semantic features, linguistic n-gram features extracted from social media data were used by Tsugawa et al. (2015) for estimating the degree of depression. Mowery et al. (2016)

Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024), pages 108–126
March 21, 2024 ©2024 Association for Computational Linguistics

further considered using demographic data such as age and gender as features, for classifying depressive symptoms based on social media (Twitter data) on a population level. Term frequency–inverse document frequency (tf–idf), which is a classic method for weighting words, was used to prepare features for predicting mental illness from social media (Thorstad and Wolff, 2019).

Deep learning methods also attracted researchers working on the subject. Yates et al. (2017) proposed a method using a neural network model to identify the risk of self-harm or depression, using data from social media Twitter and Reddit. Researchers participated in CLEF eRisk 2017 (Losada et al., 2017) focused on classifying users into binary targets: at risk or non-risk of depression. Husseini Orabi et al. (2018) explored the effectiveness of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), on detecting signs of depression using unstructured text data extracted from Twitter, released for the shared task on Computational Linguistics and Clinical Psychology (CLPsych) 2015 (Coppersmith et al., 2015).

As part of CLEF eRisk 2019, there was a proposed task about using the BDI questionnaire to automatically predict the depression level of social media users based on their social media writings, and the task continued in 2020 and 2021 (Parapar et al., 2021). Importantly, data was provided for the tasks. This led to somewhat explainable depression predictors, by indicating possible symptoms (such as lack of sleep, loss of appetite, and reduced physical activity). See Figure A1 for the full set of questions from the BDI questionnaire and their possible answers. In general, the performance of the systems that participated in this shared task was poor. Deep learning was used by a few of the participants in the task. For example, while participating in the shared task Task 3 at CLEF eRisk 2021 (Parapar et al., 2021), Inkpen et al. (2021) conducted experiments with Transformer-based models, a Deep Averaging Network (DAN) model, as well as a Hierarchical Attention Network (HAN) for text classification tasks inspired by Yang et al. (2016). On the same task, Maupomé et al. (2021) proposed a system that applied topic modeling using Embedded Topic Model (ETM) (Dieng et al., 2020) which was trained on a depression detection dataset issued from Reddit, and a regression approach with nearest-neighbors on the values of the answers. This system achieved the highest score on one metric and also performed well on the other three metrics from the shared task. In 2022, Skaik and Inkpen (2022) continued working on the task and proposed a method that combined multiple deep learning models to answer different questions. Through all these efforts, a better performance was achieved on some of the metrics. In this paper, we propose new methods to solve the task, with improved performance and with added explainability for the predicted answers.

	Quantity
Number of TREC files	3,107
Number of subjects	3,107
Number of sentences	4,264,693

Table 1: Statistics of the dataset for depression symptom search

3 Datasets

3.1 Dataset for Depression Symptoms Search

This dataset was shared for Task 1 of the eRisk 2023 (Search for symptoms of depression) (Parapar et al., 2023). The participants in the shared task were given files in the TREC format containing documents (sentences) of each user. Each document has a document ID as well as the text of the document. The corpus provided to the participants was a sentence-tagged dataset based on eRisk's past data.

The dataset contains only the derived sentences from social media, with no labels included. Languages other than English were not filtered out. The aim of the task was to extract the top-1000 relevant documents for each of the 21 symptoms in the BDI questionnaire and provide rankings for the extracted documents. Some statistics of this dataset are shown in Table 1.

3.2 Dataset for Depression Estimation

This dataset was shared at eRisk 2021 Task 3 (Measuring the severity of the signs of depression) and was built upon data shared at eRisk 2020 and eRisk 2019 for the same task. The dataset contains a training dataset and a test dataset.

The training dataset contains 90 examples, which consist of 43,514 writings written by 90 users from the 2020 and 2019 tasks, as well as their answers to the 21 questions of the BDI Questionnaire. The test dataset consists of 19,803 posts and comments written by 80 users, and their answers to the questionnaire. The labels for questions 16 and

minimal depression	depression levels 0-9
mild depression	depression levels 10-18
moderate depression	depression levels 19-29
severe depression	depression levels 30-63

Figure 1: Depression categories associated with depression levels

Category	# of Subjects[1]
minimal depression	14 (15%)
mild depression	27 (30%)
moderate depression	22 (24%)
severe depression	27 (30%)

Table 2: Statistics of depression category in the training data

18, which have different answer sets, are revised so that the answers with letters are merged into a single answer (for example, 1a and 1b are merged into 1). Each of the remaining questions has four answers: 0, 1, 2, and 3.

Statistics for the user answers (labels) are shown in Table A2. Through investigations, it can be seen that most of the symptoms users have are minor, as about 68% of users answered 0 or 1 (with about 35% answered 0 and about 32% answered 1); a few users reported severe symptoms. Specific questions may have a different distribution, such as question 16. As label frequency distributes among multiple labels, and varies by question, we can see that simply choosing a label as the default value is not practical.

To calculate an overall level of depression for a user, depression categories introduced in Figure 1 (Losada et al., 2019) were considered. The calculated levels of depression are shown in Table 2. We could conclude that users are distributed in all four categories, with most users having mild (30%) or severe (30%) depression. It is worth noting that this finding does not fully comply with the findings we had while investigating the label distribution.

4 Methodology

4.1 Search for Depression-related Writings

A writing (sentence in this context) is considered relevant to a symptom if it provides information, ideally explicit, about the user's status of that particular symptom.

The task was considered as an information retrieval task, where user-written sentences are stored as documents. We first transformed the 21 questions from the BDI questionnaire into 21 queries. We then used contextual text embedding methods for transfer learning. To accelerate the calculation of contextual representations, many keywords were selected from the questions in the BDI questionnaire, in order to filter out unrelated documents. The queries and keywords are shown in Table A1. To calculate the relevance of a document to a BDI question, we used cosine similarity between two vector representations. Our developed system can extract the most relevant sentences and provide a ranking of them for each of the 21 symptoms in the BDI questionnaire.

4.1.1 Data Normalization and Text Processing

When obtaining the vector representations, we filtered out the documents (sentences) that did not contain symptom-related keywords. Then we used a transfer learning strategy, by employing the knowledge from the language models directly to build the semantic representations. Traditional preprocessing methods were not applied, but the specific tokenization used by each contextual embedding model was used. The processing steps were applied to both the documents and the queries.

4.1.2 Universal Sentence Encoder with Cosine Similarity

The Universal Sentence Encoder (USE) is a text encoder that directly encodes sentences into vectors. It is specifically designed for transfer learning of various types of tasks. The encoder based on the transformer architecture was trained in the following way: the word representations acquired through the transformer were converted to a fixed-length encoding vector by summing the element-wise representations at each word position, and then the vector was divided by the square root of the length of the sentence to reduce sentence length effects. The inputs to the encoder are lowercased strings that are tokenized using the Penn Treebank Tokenizer (PTB), and the outputs are 512-dimensional vector representations. Since the model was designed to be of general purpose, multi-task learning was conducted (Cer et al., 2018).

We used the USE to obtain embeddings of queries and sentences, and calculated the distance between them to obtain rankings using the cosine similarity. As a result, this system named "USESim" achieved a precision of 0.60 for the top-10

[1] Percentage numbers were rounded.

documents retrieved and an average precision of 0.16 for the top-1000. We experimented with other sentence representations, such as RoBERTa (Liu et al., 2019), but the results for the retrieval task were lower.

4.1.3 Adapting USESim for Writing Selection

The dataset for depression estimation contains a large number of user writings. It is good to have a lot of information, but too many texts for a user could introduce noise, and it is difficult and costly for models to process all of them. This is why we need to filter out the less relevant writings.

We use our above-mentioned document retrieval system "USESim" for pre-processing the dataset for depression estimation to generate a smaller and more relevant dataset, by keeping only relevant writings.

4.2 Estimate Level of Depression

As mentioned, the final goal is to automatically fill in a standard depression questionnaire, the BDI questionnaire, by using LLMs to do multi-class predictions of answers to questions in the questionnaire. The questionnaire has 21 questions in total, which can be used to evaluate conditions of feelings about sadness, sleeping, etc. Each question has 4 answers, except 2 questions (question 16 about sleep patterns and question 18 about appetite) have more than 4 answers of which answers were transformed into 4 classes.

4.2.1 Data Preparation

As discussed in Section 3.2, the dataset contains a training dataset and a test dataset. The training dataset 90 users' 43,514 Reddit writings, and their answers to the BDI questionnaire. The test dataset consists of 80 users's 19,803 posts and comments as well as their responses to the questionnaire. The writing-selection system USESim is adopted for cleaning the dataset, by selecting only symptom-relevant user writings and forming a more useful dataset. Based on our settings, two types of datasets were generated using USESim:

1. **The Top-5 Dataset**
   Collected by applying the USESim to collect the top-5 relevant writings for each symptom in the BDI questionnaire. The statistics for the text length of this dataset are shown in Table A3.
2. **The Top-1 Dataset**
   Formed by applying the USESim to collect the top-1 relevant writings (the most relevant post or comment) for each of the 21 symptoms. The statistics for the text length of this dataset are shown in Table A4.

Top-5 and top-1 relevant writings were selected with the consideration of the maximum length: as many LLMs have a short context length which refers to the maximum number of tokens that the model can process, shorter input texts are desired. For example, the Llama 2 models have a context length of 4096 tokens.

4.2.2 Classification Using Large Language Models

Prompt learning is a new paradigm that is showing promising results. Large language models are essentially language models that are trained to estimate the probability $P(x;\theta)$ for text x. Prompt learning techniques utilize the probability P, to predict the output y. As an example, the output y can be the label in a classification task, and it can be extracted or transformed from the text generated by the LLM.

Selection of the LLM

A wide variety of pre-trained language models are available. In this study, the following open LLMs were applied in experiments:

- **Llama-2-13b-chat**
  Meta's Llama 2 models (Touvron et al., 2023) are LLMs that are well-supported and powerful. As auto-regressive language models, they are particularly useful for Natural Language Generation (NLG) tasks, which means that not only the output label for our classification task could be generated, but also the LLMs' explanations for predictions.
  Llama-2-13b-chat[2], which is optimized for dialogue use cases was applied in this study.
- **SUS-Chat-34B**
  Released by the Southern University of Science and Technology (SUSTech) and IDEA-CCNL, SUS-Chat-34B[3] is a bilingual (Chinese-English) dialogue model. It has significant improvements on many benchmarks of evaluation; it achieved high scores among

[2] https://huggingface.co/meta-llama/Llama-2-13b-chat-hf

[3] https://huggingface.co/SUSTech/SUS-Chat-34B

open source models of similar size (34 billion parameters), and is one of the best models with a size below 70B.

The model which was based on Yi-34B[4] was trained with 1.4 billion tokens of complex instruction data, including multi-turn dialogues, mathematics, reasoning, and others, thus the model is capable of focusing on long-text dialogue and of imitating human thought processes.

- **Neural-chat-7b-v3**
  Based on Mistral-7B-v0.1[5], the Intel neural-chat-7b-v3-1[6] is a LLM that was fine-tuned on the SlimOrca[7] dataset and with the Direct Preference Optimization (DPO) algorithm to align with human preferences. The DPO derives the probability of human preference data for an optimal policy to replace the reward model used by the Reinforcement Learning from Human Feedback (RLHF) (Lv et al., 2023).
  The model performed exceptionally well among 7-billion-parameter models.

Design of Prompts

In this step, as discussed in (Liu et al., 2021), a prompting function f_{prompt} is designed, which could then be applied to transform the input text x into the prompt $x' = f_{prompt}(x)$.

Prompt engineering is the process of designing the most effective prompting function f_{prompt}. With the knowledge we learned that most of the symptoms users have are minor, we added "suppose you are a psychiatrist who prefer to give minor diagnoses rather than serious ones" in the instruction to alleviate the models' tendency of exaggeration. Instead of giving exact meanings of each question's answers, such as "0. I do not feel sad." and "3. I am so sad or unhappy that I can't stand it." for the symptom sadness, we give only the labels to the model. In this way, the labels 0, 1, 2, and 3 are considered levels of intensity or the magnitude of the change, rather than solely markers of classes. The first consideration is the length of the prompt: it is inefficient to reserve a diverse space for this information, and models may not have a large enough context window. Another consideration is that some label meanings might confuse the model. Examples of options that could be confusing are shown in Figure 2.

9. Suicidal Thoughts or Wishes
0. I don't have any thoughts of killing myself.
1. I have thoughts of killing myself, but I would not carry them out.
2. I would like to kill myself.
3. I would kill myself if I had the chance.

10. Crying
0. I don't cry anymore than I used to.
1. I cry more than I used to.
2. I cry over every little thing.
3. I feel like crying, but I can't.

Figure 2: Some options that might be confusing to models

To clarify the symptom that the model should work on, we rephrased the 21 questions in the BDI questionnaire into 21 sentences. All the rephrased sentences talk about the levels of the symptom, to help the model's understanding. The rephrased symptoms are summarized in Table A5.

Our designed prompt template is shown in Figure 3. During experiments, "[$rephrased\ symptom$]" will be filled in using rephrased symptoms, and "[$input\ text$]" will be substituted with the concatenated user writings from the cleaned dataset: texts for each symptom when using top-5 or all texts when using top-1.

Instruction: The following paragraph was concatenated from a user's posts on social media. Suppose you are a psychiatrist who prefer to give minor diagnoses rather than serious ones, read the posts as a whole, determine the level of "[$rephrased\ symptom$]" and give a number in 0, 1, 2 or 3, then explain why.

User's posts: [$input\ text$]

Level (0, 1, 2 or 3):

Figure 3: Designed prompt template for symptom estimation

Using LLMs

The same configurations are applied to all the experimented LLMs.

The tokenizers for each model were used to encode prompts and decode outputs. The maximum

[4] https://huggingface.co/01-ai/Yi-34B
[5] https://huggingface.co/mistralai/Mistral-7B-v0.1
[6] https://huggingface.co/Intel/neural-chat-7b-v3-1
[7] https://huggingface.co/datasets/Open-Orca/SlimOrca

length for the models was set to 4,096, and only the first sentence of a pair would be truncated if longer.

To reduce the costs of utilizing LLMs, QLoRA (Dettmers et al., 2023) was used: models are run in 4-bit precision, using the NF4 (Normal Float 4) data type, double quantization, and the computational type of torch.bfloat16.

While generating texts, the models are set to use multinomial sampling, keeping the top-10 highest probability vocabulary tokens and the smallest set of most probable tokens with probabilities that add up to 0.7 or higher. The maximum length is 4,096, and the temperature (the value used to modulate probabilities of tokens) was set to be 0.1.

We experimented with a single LLM, as well as combining two LLMs, working on separate questions based on the results from experiments using training data. The experiments were conducted using the top-1 and top-5 datasets. The outputs generated by the LLMs contain labels for the predicted answers, and explanations for the predictions. We extracted the labels using regular expression (regex) and recorded them as the automated responses to the BDI questionnaire for each user in the test set.

5 Evaluation

The same evaluation metrics were applied for the task in eRisk 2019, 2020,and 2021 (Losada et al., 2019). The four metrics used for evaluation are:

- **Average Hit Rate (AHR)**
  The AHR is the hit rate averaged over all the users. The hit rate measures the number of answers systems automatically fill in that are exactly the same as the actual answers provided by the users.

- **Average Closeness Rate (ACR)**
  The ACR is the Closeness Rate averaged over all the users. It takes into account that the answers represent an ordinal scale, rather than merely separate options. To get the closeness rate, first compute the absolute difference between the automatically filled answer and the actual answer, then transform the calculated absolute difference into an effectiveness score as follows:

  $$CR = \frac{maxad - ad}{maxad} \quad (1)$$

  where $maxad$ stands for the maximum absolute difference, and ad is the absolute difference.

- **Average DODL (ADODL)**
  The ADODL is the difference between the system's and actual overall depression levels averaged over all users. The Difference between Overall Depression Levels (DODL) is obtained by first calculating the overall depression levels for the system-filled and actual questionnaire, then computing the absolute difference between the two overall scores. The DODL is normalized as follows:

  $$DODL = \frac{63 - ad}{63} \quad (2)$$

  where ad is the absolute difference between the automated and actual overall score.

- **Depression Category Hit Rate (DCHR)**
  The DCHR measures the closeness of the depression estimation achieved over all users according to the established depression categories introduced in Figure 1. It calculates the fraction of cases where the automated questionnaire led to a category that is identical to the user's actual depression category.

6 Results and Discussion

The experimental results of our systems using LLMs are shown in Table 3.

We can learn that the usage of USESim for user writing selection is helpful, and it is generally better to have more writings kept so that the model could have more information about the user, and the writings would be more focused on the specific question. In our experiments, the usage of top-5 dataset leads to a better performance than using the top-1 dataset.

When using the top-5 dataset, the model neural-chat-7b-v3-1 performed better than SUS-Chat-34B on the metrics AHR, ADODL, and DCHR. This is surprising since the neural-chat-7b-v3-1 is much smaller than SUS-Chat-34B in terms of size/number of parameters. The reason could be the language focus and the application of the DPO algorithm. The Llama-2-13b-chat model did not perform well on any of the metrics.

Through experimenting on the training data, the neural-chat-7b-v3-1's answers on questions 4, 8, 9, 11, 12, 16, 18, 19, 20 and 21 are combined with SUS-Chat-34B's answers on questions 1, 2, 3, 5, 6,

Run	AHR	ACR	ADODL	DCHR
Llama-2-13b-chat_top1	21.90	63.29	72.22	42.5
Llama-2-13b-chat_top5	22.32	63.51	72.16	42.5
neural-chat-7b-v3-1_top1	31.96	71.82	84.12	48.75
neural-chat-7b-v3-1_top5	33.63	70.83	**85.87**	**52.5**
SUS-Chat_top1	32.61	72.02	84.64	50.0
SUS-Chat_top5	33.51	72.57	83.53	**52.5**
neural-chat+SUS-Chat_top1	34.70	72.91	85.41	48.75
neural-chat+SUS-Chat_top5	**37.32**	**73.25**	85.63	50.0

Table 3: Results of LLM-based systems

7, 10, 13, 14, 15 and 17 as the combined system. The combined runs performed well on the AHR, ACR and ADODL metrics.

Although the performance on some metrics is still not outstanding, our systems scored over 85% on ADODL, which is an improvement considering that ADODL is the most critical metric for measuring depression at the population level (Skaik and Inkpen, 2022). Many runs scored over 50 on DCHR, meaning that they predicted correctly for more than half of the test subjects on predicting their depression category.

Our experiments proved that LLMs have learned knowledge about various depression-related symptoms, and they can make better inferences than supervised deep learning techniques, with zero-shot learning (no training) and properly designed prompts.

6.1 Local Explanations of LLMs

Through prompts, the LLMs were asked to provide explanations for their predictions. Even though these explanations are not necessarily factual, they provide insights about the important information in the given user writings. In Figure 4, an example of user writings is given, which is answered by LLMs for Q18 as a change in appetite.

Figure 5 shows the prediction and explanation from Llama-2-13b-chat for text from Figure 4. In the explanation, the model mentioned several physical and mental issues described in the user's writings, such as inflammation and mental health issues. The model predicted 2 as the answer.

In Figure 6, the classification and explanation for the given example generated by neural-chat-7b-v3-1 are shown. The model mentioned that the user needed to set alarms to eat, having stomach flu and fluctuating weights, which could affect the user's appetite. An answer label of 2 is given by the model.

SUS-Chat-34B's prediction and explanation are shown in Figure 7. The model presumed that the user had a higher level of change in appetite since the user had to set alarms to eat at some points, but the model also mentioned that no significant weight changes were presented. The model generated 2 as the answer to the question.

6.1.1 Evaluation of Explanations

In (Rajagopal et al., 2021), several criteria were introduced for evaluating the explanations, including sufficiency (via BERT-score), plausibility, and trustability. Due to limitations on the amount of time available for conducting evaluations, we only evaluated explanations for the best-performing system on one metric: sufficiency – to automatically evaluate how well the explanations reflect the system's predictions. Manual evaluation with experts to calculate the other measures is left as future work.

Rajagopal et al. (2021) used the "Faithfulness-by-construction" (FRESH) pipeline (Jain et al., 2020) to evaluate sufficiency: a BERT (Devlin et al., 2019) based classifier is trained to perform a task using only the extracted explanations. A high accuracy would indicate a high sufficiency of the explanations, which means that the explanations are indicative of the predicted labels. Based on this method, we implemented a BERT classifier, trained it with the generated explanations for the training dataset as the training data, and evaluated the classifier with the explanations for the test data.

The sufficiency score for the explanations generated by our best system was 80.77%. This indicates the effectiveness of generated explanations of the system. The configurations are as follows: a fast Rust-based pretrained tokenizer of "bert-base-cased", with default settings on the rest; the pretrained "bert-base-cased" model for the classifica-

User's posts: I would say coriander. Grab a package and eat it. **Great snacks.** I mean the psychosis is still untreated I'm thinking about what to do next. **My weight fluctuates** because, in addition to severe inflammation, I also have some mental health issues that cause loss of memory. I actually **have to set alerts to eat** for certain times. I had a huge success that day and was approved to work from home. Now I **got stomach flu**. I don't feel accomplished enough. I'm **vomiting for days** and my mental state is getting worse and I hear constant noise from the church. Oh my goodness, my back pain, my **constant digestive issues**, plus my mental illness, have all gotten worse. I think all I really need is a happy message. Usually good news upsets me, but I really need it today.

Figure 4: An example of concatenated writings. Some words were highlighted for readability. In fact, we rephrased the sentences in the example, because the data sharing agreement that we signed when we obtained the dataset does not allow us to use the real example directly.

Metric	Our system	Our best	Best system	State-of-the-art
AHR	neural-chat+SUS-Chat_top5	37.32	BDI_Multi_Model	**41.25**
ACR	neural-chat+SUS-Chat_top5	**73.25**	UPV-Symanto 4_symanto_upv_lingfeat_cors	73.17
ADODL	neural-chat-7b-v3-1_top5	**85.87**	BDI_Multi_Model	83.79
DCHR	neural-chat-7b-v3-1_top5	**52.5**	BDI_Multi_Model	48.75

Table 4: Our results compared to the state-of-the-art

Level (0, 1, 2 or 3): 2

Explanation:

The user's posts suggest that they have experienced significant changes in appetite, as they mention "weight fluctuations" and setting alerts to eat. They also mention "severe inflammation" and "mental health issues" which could be contributing to their appetite changes. Additionally, they mention "stomach flu" and "constant digestive issues" which could also be affecting their appetite. Based on this information, I would rate the user's experience of changes in appetite as a level 2.

Figure 5: Classification and explanation by Llama-2-13b-chat

Level (0, 1, 2 or 3): 2

The user experienced changes in appetite at a level of 2. This is because they mentioned having mental health issues that cause loss of memory, which can affect appetite. They also mentioned setting alerts to eat for certain times, indicating that they have to be more mindful of their eating habits. Additionally, they mentioned having stomach flu, which can also affect appetite. The user's fluctuating weight and other health issues, such as back pain, digestive issues, and mental illness, further contribute to the complexity of their situation.

Figure 6: Classification and explanation by neural-chat-7b-v3-1

tion into the 4 classes/labels; and with default values for the hyperparameters except for the total number of training epochs performed that was set to 8.

6.2 Comparison to Related Work

Skaik and Inkpen (2022) proposed a method for this task that first applies topic-based filtering to clean the irrelevant posts, then conducts multiple iterations to select and assign different models to different question sets. The final model named "BDI_Multi_Model" was formed using 5 HANs, 3 Long Short-Term Memory (LSTM) models, and 2 transformers. Their proposed system sets new records for several metrics. The comparison between our best systems and the state-of-the-art is concluded in Table 4.

Through comparisons, we can see that our systems achieved the best scores on three out of four metrics: ACR, ADODL, and DCHR. It is worth noting that the run neural-chat+SUS-Chat_top5 achieved 85.63 on ADODL and 50.0 on DCHR, which indicates that the system neural-chat+SUS-

```
### Level (0, 1, 2 or 3): 2

Explanation: The user mentions having to set alerts to eat for certain times, which indicates that they have experienced changes in appetite. However, they do not mention any significant weight loss or gain, which would indicate a higher level of change in appetite. Therefore, a level of 2 is appropriate.
```

Figure 7: Classification and explanation by SUS-Chat-34B

Chat_top5 performed better on 3 metrics (that were achieved by different systems in previous work).

In addition to answering questions without further training, our LLM-based systems could generate local explanations for all their answers, and these explanations are more readable and straightforward compared with the features-based model explanations. More importantly, no examples with labels are needed while using these LLMs, which means that the methods could fit the tasks that lack labeled data, and could be adapted into other tasks and domains. We believe that with the development of computing power, if more user writings (not just selected ones) are included in the prompts and a larger model could be used, higher scores could be achieved using this method.

7 Conclusion and Future Work

In the study, we designed a system that searches for relevant sentences in numerous user writings, and applied it to provide cleaner data for a depression estimation system based on LLMs that automatically answers the questions from the BDI questionnaire. The resulting depression detection system has good performance on several metrics, and could sufficiently explain its answers to every question on the questionnaire for every user, without training on labeled data.

However, due to the randomness of cyberspace, users' speeches on social media platforms cannot fully, objectively, accurately, and consistently describe their status of various depression-related symptoms. In the future, it would be good to collect larger high-quality datasets, so that we can run more experiments to calibrate our system and verify its effectiveness.

Also, with more computing resources and more powerful LLMs, much more user writings could be given to the model rather than filtered out, and it is expected that this would improve the performance. Since our system does not need a large amount of training data, only a small set of labeled examples to design prompts, this is a promising avenue for automatically answering other types of mental health questionnaires, such as PHQ-9, anxiety questionnaires, etc.

Ethics Statement

This study complies with the ACL Ethics Policy[8]. Since the datasets are collected from Reddit and are anonymized, privacy is respected, and no bias is introduced. The filled questionnaires are meant to be for initial information and used as references by professionals, not for self-diagnosis. Dictionaries and Grammarly were used when writing this paper, but no AI assistance was involved in the writing or in the programming.

Limitations

The proposed system on user writing selection would result in datasets mostly in English; thus, the system is limited to English-written texts. The texts in foreign languages were filtered out; therefore, more investigation will be needed in multilingual settings.

We set many restrictions on context length, sampling and model size due to the high requirements of computing resources. These restrictions could affect the performance but can be removed if more resources are available.

In addition, all evaluations are conducted without human health practitioners. It is better to have mental health practitioners review system predictions and explanations and test the system in clinical settings.

Acknowledgements

We thank the Natural Sciences and Engineering Research Council of Canada (NSERC) for supporting our research. We thank the CLEF eRisk organizers, for providing great opportunities for us to explore the task of automatic depression detection based on the BDI questionnaire.

References

A. T. Beck, C. H. Ward, M. Mendelson, J. Mock, and J. Erbaugh. 1961. An Inventory for Measuring De-

[8] https://www.aclweb.org/portal/content/acl-code-ethics

pression. *Archives of General Psychiatry*, 4(6):561–571.

Adriel Boals and Kitty Klein. 2005. Word Use in Emotional Narratives about Failed Romantic Relationships and Subsequent Mental Health. *Journal of Language and Social Psychology*, 24(3):252–268. Publisher: SAGE Publications Inc.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal Sentence Encoder. ArXiv:1803.11175 [cs].

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. CLPsych 2015 Shared Task: Depression and PTSD on Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39, Denver, Colorado. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. ArXiv:2305.14314 [cs].

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv:1810.04805 [cs].

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. Topic Modeling in Embedding Spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. 2018. Deep Learning for Depression Detection of Twitter Users. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 88–97, New Orleans, LA. Association for Computational Linguistics.

Diana Inkpen, Ruba Skaik, Prasadith Buddhitha, Dimo Angelov, and Maxwell Thomas Fredenburgh. 2021. uOttawa at eRisk 2021: Automatic Filling of the Beck's Depression Inventory Questionnaire using Deep Learning.

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C. Wallace. 2020. Learning to Faithfully Rationalize by Construction. ArXiv:2005.00115 [cs].

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pretrain, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. ArXiv:2107.13586 [cs] version: 1.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv:1907.11692 [cs].

David E. Losada, Fabio Crestani, and Javier Parapar. 2017. eRISK 2017: CLEF Lab on Early Risk Prediction on the Internet: Experimental Foundations. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Lecture Notes in Computer Science, pages 346–360, Cham. Springer International Publishing.

David E. Losada, Fabio Crestani, and Javier Parapar. 2019. Overview of eRisk 2019 Early Risk Prediction on the Internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Lecture Notes in Computer Science, pages 340–357, Cham. Springer International Publishing.

Kaokao Lv, Wenxin Zhang, and Haihao Shen. 2023. Supervised Fine-Tuning and Direct Preference Optimization on Intel Gaudi2.

Steven Marwaha, Edward Palmer, Trisha Suppes, Emily Cons, Allan H. Young, and Rachel Upthegrove. 2023. Novel and emerging treatments for major depression. *The Lancet*, 401(10371):141–153. Publisher: Elsevier.

Diego Maupomé, Maxime D Armstrong, Fanny Rancourt, and Thomas Soulas. 2021. Early Detection of Signs of Pathological Gambling, Self-Harm and Depression through Topic Extraction and Neural Networks.

Danielle L. Mowery, Albert Park, Craig Bryan, and Mike Conway. 2016. Towards Automatically Classifying Depressive Symptoms from Twitter Data for Population Health. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 182–191, Osaka, Japan. The COLING 2016 Organizing Committee.

Javier Parapar, Patricia Martín-Rodilla, David E. Losada, and Fabio Crestani. 2021. Overview of eRisk 2021: Early Risk Prediction on the Internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings*, pages 324–344, Berlin, Heidelberg. Springer-Verlag.

Javier Parapar, Patricia Martín-Rodilla, David E. Losada, and Fabio Crestani. 2023. Overview of eRisk 2023: Early Risk Prediction on the Internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Lecture Notes in Computer Science, pages 294–315, Cham. Springer Nature Switzerland.

Dheeraj Rajagopal, Vidhisha Balachandran, Eduard Hovy, and Yulia Tsvetkov. 2021. SelfExplain: A Self-Explaining Architecture for Neural Text Classifiers. ArXiv:2103.12279 [cs].

Ruba S. Skaik and Diana Inkpen. 2022. Predicting Depression in Canada by Automatic Filling of Beck's

Depression Inventory Questionnaire. *IEEE Access*, 10:102033–102047. Conference Name: IEEE Access.

Yla R. Tausczik and James W. Pennebaker. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29(1):24–54. Publisher: SAGE Publications Inc.

Robert Thorstad and Phillip Wolff. 2019. Predicting future mental illness from social media: A big-data approach. *Behavior Research Methods*, 51(4):1586–1600.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. ArXiv:2307.09288 [cs].

Sho Tsugawa, Yusuke Kikuchi, Fumio Kishino, Kosuke Nakajima, Yuichi Itoh, and Hiroyuki Ohsaki. 2015. Recognizing Depression from Twitter Activity. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 3187–3196, New York, NY, USA. Association for Computing Machinery.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical Attention Networks for Document Classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California. Association for Computational Linguistics.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and Self-Harm Risk Assessment in Online Forums. ArXiv:1709.01848 [cs].

A Appendix

Question	Keywords	Query
Q1	sadness, sad, unhappy	Sadness. I feel sad unhappy cannot stand it.
Q2	pessimism, discouraged, hopeless	Pessimism. I feel discouraged about my future is hopeless and will get worse.
Q3	failure, fail	Past Failure. I have failed.
Q4	pleasure, enjoy	Loss of Pleasure. I don't enjoy things.
Q5	guilty	Guilty Feelings. I feel guilty.
Q6	punishment, punish	Punishment Feelings. I am being punished.
Q7	confidence, disappointed	Self-Dislike. I have lost confidence. I am disappointed in myself.
Q8	criticalness, critical, criticize, blame, fault	Self-Criticalness. I criticize myself blame myself for my faults.
Q9	suicidal, suicide, kill	Suicidal Thoughts or Wishes. I kill myself.
Q10	crying, cry	Crying. I cry.
Q11	agitation, agitate, restless	Agitation. I am restless or agitated keep moving.
Q12	interest, interested	Loss of Interest. It's hard to get interested.
Q13	indecisiveness, decision, decide	Indecisiveness. I find it difficult to make decisions.
Q14	worthlessness, worthless, worthwhile, useful	Worthlessness. I feel worthless not useful.
Q15	energy, energetic	Loss of Energy. I don't have enough energy.
Q16	sleep, sleeping	Changes in Sleeping Pattern. I sleep more or less than usual.
Q17	irritability, irritable, angry	Irritability. I am irritable.
Q18	appetite, food, eat	Changes in Appetite. My appetite is greater or less.
Q19	concentration, concentrate	Concentration Difficulty. It's hard to keep my mind. I can't concentrate.
Q20	tiredness, fatigue, tired	Tiredness or Fatigue. I am tired or fatigued.
Q21	sex	Loss of Interest in Sex. I am less interested in sex.

Table A1: Queries and keywords for each question

	Answer 0	Answer 1	Answer 2	Answer 3
Q1	27 (30%)	47 (52%)	11 (12%)	5 (5%)
Q2	22 (24%)	34 (37%)	20 (22%)	14 (15%)
Q3	22 (24%)	35 (38%)	18 (20%)	15 (16%)
Q4	28 (31%)	33 (36%)	23 (25%)	6 (6%)
Q5	34 (37%)	32 (35%)	12 (13%)	12 (13%)
Q6	60 (66%)	13 (14%)	11 (12%)	6 (6%)
Q7	28 (31%)	17 (18%)	23 (25%)	22 (24%)
Q8	28 (31%)	27 (30%)	23 (25%)	12 (13%)
Q9	41 (45%)	37 (41%)	7 (7%)	5 (5%)
Q10	42 (46%)	23 (25%)	8 (8%)	17 (18%)
Q11	37 (41%)	31 (34%)	14 (15%)	8 (8%)
Q12	28 (31%)	32 (35%)	8 (8%)	22 (24%)
Q13	38 (42%)	21 (23%)	16 (17%)	15 (16%)
Q14	38 (42%)	21 (23%)	20 (22%)	11 (12%)
Q15	17 (18%)	32 (35%)	28 (31%)	13 (14%)
Q16	17 (18%)	36 (40%)	24 (26%)	13 (14%)
Q17	38 (42%)	31 (34%)	16 (17%)	5 (5%)
Q18	32 (35%)	30 (33%)	15 (16%)	13 (14%)
Q19	29 (32%)	25 (27%)	25 (27%)	11 (12%)
Q20	21 (23%)	34 (37%)	21 (23%)	14 (15%)
Q21	51 (56%)	18 (20%)	11 (12%)	10 (11%)
Total	678 (35%)	609 (32%)	354 (19%)	249 (13%)

Table A2: Statistics of labels in the training data for depression estimating

Instructions:

This questionnaire consists of 21 groups of statements. Please read each group of statements carefully, and then pick out the one statement in each group that best describes the way you feel.
If several statements in the group seem to apply equally well, choose the highest number for that group.

1. Sadness
0. I do not feel sad.
1. I feel sad much of the time.
2. I am sad all the time.
3. I am so sad or unhappy that I can't stand it.

2. Pessimism
0. I am not discouraged about my future.
1. I feel more discouraged about my future than I used to be.
2. I do not expect things to work out for me.
3. I feel my future is hopeless and will only get worse.

3. Past Failure
0. I do not feel like a failure.
1. I have failed more than I should have.
2. As I look back, I see a lot of failures.
3. I feel I am a total failure as a person.

4. Loss of Pleasure
0. I get as much pleasure as I ever did from the things I enjoy.
1. I don't enjoy things as much as I used to.
2. I get very little pleasure from the things I used to enjoy.
3. I can't get any pleasure from the things I used to enjoy.

5. Guilty Feelings
0. I don't feel particularly guilty.
1. I feel guilty over many things I have done or should have done.
2. I feel quite guilty most of the time.
3. I feel guilty all of the time.

6. Punishment Feelings
0. I don't feel I am being punished.
1. I feel I may be punished.
2. I expect to be punished.
3. I feel I am being punished.

7. Self-Dislike
0. I feel the same about myself as ever.
1. I have lost confidence in myself.
2. I am disappointed in myself.
3. I dislike myself.

Figure A1: Beck's Depression Inventory (part 1)

8. Self-Criticalness
0. I don’t criticize or blame myself more than usual.
1. I am more critical of myself than I used to be.
2. I criticize myself for all of my faults.
3. I blame myself for everything bad that happens.

9. Suicidal Thoughts or Wishes
0. I don’t have any thoughts of killing myself.
1. I have thoughts of killing myself, but I would not carry them out.
2. I would like to kill myself.
3. I would kill myself if I had the chance.

10. Crying
0. I don’t cry anymore than I used to.
1. I cry more than I used to.
2. I cry over every little thing.
3. I feel like crying, but I can’t.

11. Agitation
0. I am no more restless or wound up than usual.
1. I feel more restless or wound up than usual.
2. I am so restless or agitated that it’s hard to stay still.
3. I am so restless or agitated that I have to keep moving or doing something.

12. Loss of Interest
0. I have not lost interest in other people or activities.
1. I am less interested in other people or things than before.
2. I have lost most of my interest in other people or things.
3. It’s hard to get interested in anything.

13. Indecisiveness
0. I make decisions about as well as ever.
1. I find it more difficult to make decisions than usual.
2. I have much greater difficulty in making decisions than I used to.
3. I have trouble making any decisions.

14. Worthlessness
0. I do not feel I am worthless.
1. I don’t consider myself as worthwhile and useful as I used to.
2. I feel more worthless as compared to other people.
3. I feel utterly worthless.

15. Loss of Energy
0. I have as much energy as ever.
1. I have less energy than I used to have.
2. I don’t have enough energy to do very much.
3. I don’t have enough energy to do anything.

Figure A1: Beck’s Depression Inventory (part 2)

16. Changes in Sleeping Pattern
0. I have not experienced any change in my sleeping pattern.
la. I sleep somewhat more than usual.
lb. I sleep somewhat less than usual.
2a. I sleep a lot more than usual.
2b. I sleep a lot less than usual.
3a. I sleep most of the day.
3b. I wake up 1-2 hours early and can't get back to sleep.

17. Irritability
0. I am no more irritable than usual.
1. I am more irritable than usual.
2. I am much more irritable than usual.
3. I am irritable all the time.

18. Changes in Appetite
0. I have not experienced any change in my appetite.
la. My appetite is somewhat less than usual.
lb. My appetite is somewhat greater than usual.
2a. My appetite is much less than before.
2b. My appetite is much greater than usual.
3a. I have no appetite at all.
3b. I crave food all the time.

19. Concentration Difficulty
0. I can concentrate as well as ever.
1. I can't concentrate as well as usual.
2. It's hard to keep my mind on anything for very long.
3. I find I can't concentrate on anything.

20. Tiredness or Fatigue
0. I am no more tired or fatigued than usual.
1. I get more tired or fatigued more easily than usual.
2. I am too tired or fatigued to do a lot of the things I used to do.
3. I am too tired or fatigued to do most of the things I used to do.

21. Loss of Interest in Sex
0. I have not noticed any recent change in my interest in sex.
1. I am less interested in sex than I used to be.
2. I am much less interested in sex now.
3. I have lost interest in sex completely.

Figure A1: Beck's Depression Inventory (part 3)

	Mean	Min	Max
Training-Q1	183.78	31.00	890.00
Training-Q2	276.74	37.00	958.00
Training-Q3	177.38	32.00	743.00
Training-Q4	206.89	39.00	1505.00
Training-Q5	171.72	32.00	994.00
Training-Q6	177.67	27.00	1171.00
Training-Q7	197.49	41.00	923.00
Training-Q8	161.58	44.00	783.00
Training-Q9	245.66	27.00	962.00
Training-Q10	112.19	27.00	445.00
Training-Q11	268.56	33.00	1502.00
Training-Q12	159.20	43.00	567.00
Training-Q13	211.44	25.00	851.00
Training-Q14	201.67	39.00	759.00
Training-Q15	248.41	35.00	716.00
Training-Q16	207.50	50.00	744.00
Training-Q17	217.79	31.00	1406.00
Training-Q18	177.93	32.00	630.00
Training-Q19	223.30	45.00	866.00
Training-Q20	228.73	49.00	960.00
Training-Q21	276.11	61.00	811.00
Training-All	2122.86	287.00	11039.00
Test-Q1	304.71	36.00	1020.00
Test-Q2	413.99	43.00	2312.00
Test-Q3	237.18	29.00	1261.00
Test-Q4	274.61	33.00	945.00
Test-Q5	246.57	28.00	955.00
Test-Q6	250.69	26.00	1154.00
Test-Q7	314.34	42.00	1091.00
Test-Q8	232.59	33.00	693.00
Test-Q9	329.64	26.00	1703.00
Test-Q10	206.88	21.00	978.00
Test-Q11	382.46	37.00	1579.00
Test-Q12	225.05	28.00	960.00
Test-Q13	295.59	22.00	1571.00
Test-Q14	283.14	30.00	1083.00
Test-Q15	357.49	34.00	1115.00
Test-Q16	253.56	42.00	768.00
Test-Q17	322.65	25.00	1080.00
Test-Q18	230.90	40.00	857.00
Test-Q19	321.32	38.00	1022.00
Test-Q20	345.68	41.00	1665.00
Test-Q21	367.70	52.00	1467.00
Test-All	2561.28	202.00	11424.00

Table A3: Statistics of text length of the cleaned data (top-5)

	Mean	Min	Max
Training-Q1	29.69	5.00	181.00
Training-Q2	51.88	5.00	234.00
Training-Q3	26.08	4.00	245.00
Training-Q4	36.30	4.00	255.00
Training-Q5	28.13	4.00	266.00
Training-Q6	31.40	4.00	693.00
Training-Q7	31.10	5.00	314.00
Training-Q8	31.19	4.00	331.00
Training-Q9	43.63	5.00	305.00
Training-Q10	25.19	4.00	213.00
Training-Q11	50.84	6.00	260.00
Training-Q12	27.90	5.00	146.00
Training-Q13	34.90	4.00	159.00
Training-Q14	35.27	5.00	255.00
Training-Q15	47.36	5.00	260.00
Training-Q16	44.61	5.00	260.00
Training-Q17	37.88	4.00	304.00
Training-Q18	35.23	4.00	260.00
Training-Q19	40.43	5.00	245.00
Training-Q20	59.67	4.00	382.00
Training-Q21	47.08	4.00	204.00
Training-All	524.31	77.00	2261.00
Test-Q1	56.11	4.00	438.00
Test-Q2	84.61	6.00	438.00
Test-Q3	37.94	4.00	321.00
Test-Q4	44.86	5.00	438.00
Test-Q5	40.58	4.00	405.00
Test-Q6	42.67	5.00	532.00
Test-Q7	60.31	4.00	438.00
Test-Q8	46.88	4.00	509.00
Test-Q9	76.28	4.00	438.00
Test-Q10	38.52	4.00	326.00
Test-Q11	77.05	6.00	887.00
Test-Q12	43.34	4.00	429.00
Test-Q13	59.45	5.00	752.00
Test-Q14	57.23	6.00	438.00
Test-Q15	78.53	6.00	438.00
Test-Q16	59.17	6.00	398.00
Test-Q17	76.19	4.00	446.00
Test-Q18	38.81	4.00	242.00
Test-Q19	65.39	7.00	367.00
Test-Q20	63.86	5.00	398.00
Test-Q21	90.78	6.00	445.00
Test-All	742.86	61.00	2777.00

Table A4: Statistics of text length of the cleaned data (top-1)

Question	Rephrased symptom
Q1	how sad the user feels
Q2	how discouraged the user is about future
Q3	how much the user feels like a failure
Q4	how much the user loses pleasure from things
Q5	how often the user feels guilty
Q6	how much the user feels punished
Q7	how much the user feels disappointed about him/herself
Q8	how often the user criticizes or blames him/herself
Q9	how much the user thinks about killing him/herself
Q10	how often the user cries
Q11	how much the user feels restless or agitated
Q12	how much the user loses interest in things
Q13	how difficult the user to make decisions
Q14	how much the user feels worthless
Q15	how much the user loses energy
Q16	how much the user experienced changes in sleeping
Q17	how much the user feels irritable
Q18	how much the user experienced changes in appetite
Q19	how difficult the user to concentrate
Q20	how much the user feels tired or fatigued
Q21	how much the user loses interest in sex

Table A5: Rephrased symptoms on the BDI questionnaire