Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2682–2692
Marseille, 20-25 June 2022

MentSum: A Resource for Exploring Summarization of Mental Health Online Posts

Sajad Sotudeh[1,2]*, Nazli Goharian[1,2]*, Zachary Young[2]
[1]Information Retrieval Lab, Georgetown University
[2]Department of Computer Science, Georgetown University
{sajad, nazli}@ir.cs.georgetown.edu, zjy2@georgetown.edu

Abstract

Mental health remains a significant challenge of public health worldwide. With increasing popularity of online platforms, many use the platforms to share their mental health conditions, express their feelings, and seek help from the community and counselors. Some of these platforms, such as Reachout, are dedicated forums where the users register to seek help. Others such as Reddit provide subreddits where the users publicly but anonymously post their mental health distress. Although posts are of varying length, it is beneficial to provide a short, but informative summary for fast processing by the counselors. To facilitate research in summarization of mental health online posts, we introduce **Ment**al Health **Sum**marization dataset, MENTSUM, containing over 24k carefully selected user posts from Reddit, along with their short user-written summary (called TLDR) in English from 43 mental health subreddits. This domain-specific dataset could be of interest not only for generating short summaries on Reddit, but also for generating summaries of posts on the dedicated mental health forums such as Reachout. We further evaluate both extractive and abstractive state-of-the-art summarization baselines in terms of ROUGE scores, and finally conduct an in-depth human evaluation study of both user-written and system-generated summaries, highlighting challenges in this research.

Keywords: Text Summarization, Summarization Dataset, Mental Health Summarization

1. Introduction

Mental health has been a global public health challenge for many years and even more so since the COVID-19 pandemic (Holmes et al., 2020; Pfefferbaum and North, 2020; Otu et al., 2020). Social media has served as a viable platform for many to share their frustrations, emotions, depressions, and also their already diagnosed mental disorders. Figure 1 depicts the growing popularity (measured by the number of subscribers) of discussion forums dedicated to three mental disorders in Reddit social discussion website over the years. [1].

Online social platforms such as *Reddit* [2] and *Reachout* [3] have become increasingly popular over the recent years due to the vital networking facets that they offer to the community users. These platforms provide users with an opportunity to share different types of user-curated and user-generated content, ranging from daily updates/statuses to sharing personal anecdotes and mental conditions. Users can also interact with other users, carry on conversations through which they can express their feelings and views regarding a specific topic. Platforms such as *Reachout* are not public, requiring users to register; users' content are not visible to anyone but to the permitted users and counselors. On the other hand, in the public platforms such as *Reddit*, users can openly exchange information with each other through community-based *subreddits*, each of which specified with a certain theme or condition, such as suicide watch, mental health, alcoholism, attention-deficit/hyperactivity disorder (ADHD), depression, anxiety, etc. Each post in any of these subreddits, however, may report more than one past or present condition and what the user is distressed about.

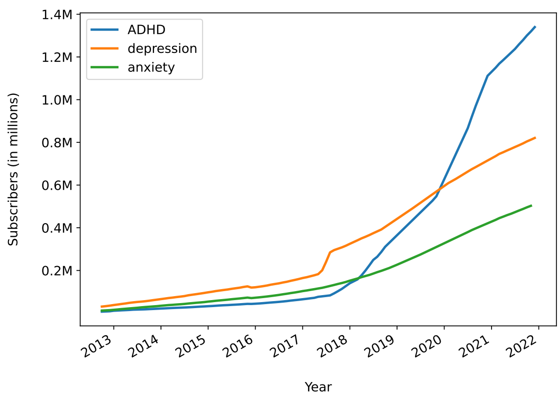

Figure 1: Growing popularity of mental health related forums in Reddit.

The user-generated content on many of such platforms might be of varying length. Longer posts may address multitude of issues of concern or simply be a lengthy elaboration of the user on the situation. The longer a post is, the more time it requires a counselor for reading the post which leads to fatigue and/or delay in a timely response. Our hypothesis is that a short yet informative summary of each user's post provides the counselors with the important information of the post in a glimpse before reading the details. Hence, in this research we

*Equal contribution
[1]Statistics from `https://subredditstats.com/`
[2]`https://www.reddit.com/`
[3]`https://au.reachout.com/`

create a dataset resource for the research community to be utilized in the short text (known as TLDR) summarization of mental health related social media posts.

A great deal of research studies in social media mental health domain have focused on developing classification models and their needed datasets to either triage the severity of the potential harm or to identify the type of mental disorders; among these efforts are (Choudhury et al., 2013; Coppersmith et al., 2014a; Yates et al., 2017; Coppersmith et al., 2018; Cohan et al., 2018; Garg et al., 2021). Our goal is not to undermine classification of disorders but potentially serve as yet another additional form of guidance to the readers/counselors of the posts through short summaries of the posts. To this end, we had to create a domain-specific dataset to contain social media mental health related posts, along with their short user-written summary as the gold standard. This dataset, MENTSUM, contains over 24k posts with pairwise user-written summaries. We hope that MENTSUM would expedite future work in the social media mental health text summarization task. In short, our contributions are:

- Creating Mental Health Summarization dataset, called MENTSUM that includes over 24k user posts from the online mental health discussion forums along with their user-written short summaries.

- Providing the results of existing strong baselines in summarization, covering both extractive and abstractive approaches.

- Carrying out a human evaluation and error analysis in terms of fluency, informativeness, and conciseness of the system generated summaries, on its own, and in comparison with the user-written summaries.

  We believe that insights from our human evaluation study can be further used by future work to investigate more sophisticated models for moving the field forward.

2. Related work

Online social platforms provide a considerable wealth of textual data, attracting attention of those who study the users' mental conditions in social environments. Early works in mental health research have put their focus on understanding and identifying mental health conditions in social media platforms such as Reddit and Twitter (Choudhury et al., 2013; Resnik et al., 2013; Coppersmith et al., 2014b; Mowery et al., 2017b); particularly, predicting mental state of the users in online media (Hao et al., 2013; Wang et al., 2013; Mowery et al., 2017a), exploring the types of mental health condition (Wilson et al., 2014), detecting the severity level of mental disorders (O'Dea et al., 2015; Chancellor et al., 2016), studying mental health discourse (Choudhury and De, 2014), studying language of users and identifying those with a high risk of mental illness (Milne et al., 2016; Cohan et al., 2017), and analyzing the impact of conversation between a target user and participants (Soldaini et al., 2018). Some research focused on creating large-scale mental health datasets such as RSDD (Yates et al., 2017), SMHD (Cohan et al., 2018) to detect potential of mental health conditions through the general language of users, and RSDD-TIME (MacAvaney et al., 2018) to study the temporal information of diagnoses. Unlike existing work whose main focus has been on classification tasks, we define a text summarization task over users' mental health content in online social media platforms.

While summarization of clinical reports has already attracted the attention of researchers (Mishra et al., 2014; Goldstein and Shahar, 2016; MacAvaney et al., 2019; Zhang et al., 2020; Sotudeh et al., 2020), the summarization of social media mental health posts has not been explored previously, which could be due to lack of large-scale mental health summarization datasets. The closest work but yet different than ours is done by Manas et al. (2021) that aim to summarize the mental health diagnostic interviews on a small-scale conversational dataset (189 patient interviews) without human-written gold summaries. Hence, to the best of our knowledge, we are the first to propose a relatively large scale mental health text summarization dataset based off social media users' content with user-written short summaries.

3. MENTSUM dataset

In this section, we elaborate on construction of our dataset, provide dataset statistics, and analyze the characteristics of the data. Subsequently, we provide the ethics and privacy of the dataset.

3.1. Dataset construction

Our constructed dataset is based off Reddit mental health related posts of the users along with their user-written short summaries (called TLDR) as the ground-truth. Note that the author of the post and TLDR is the same; hence, the goldness of this ground-truth TLDR, in respect to its fluency, and completeness might be impacted by the emotional state of the post's author. The choice of Reddit for building our dataset is motivated by being a public and popular platform, namely its public content and also the availability of the short summary ground-truth for each post. Reddit is a social media platform that supports communities called *subreddits*, each dedicated to a specific topic.

We used Pushshift (Baumgartner et al., 2020) which is a social media data repository containing recent and historical dumps of posted content on Reddit, which are made publicly available to the Natural Language Processing (NLP) community for research studies. We downloaded the Reddit data dumps covering the period of 2005-2021, and filtered the posts based on a set of pre-defined 43 mental health subreddits. [4] As not all of these users' posts have short summaries (i.e., TLDR), using regular expression, we harvested posts that contain a TLDR summary as done in (Völske et al., 2017; Sotudeh et al., 2021). The regular expression matches keywords that begin with uncased "TL" and end with uncased "DR", allowing up to three characters in between.

Social media texts are generally unstructured and noisy in terms of having chunky sentences, and grammatical errors (Baldwin et al., 2013). This is due to the fact that users can freely express themselves (Liu et al., 2016). To further preserve high-quality instances, we applied the following filtering using a set of hand-crafted filtering rules listed below.

1. **Token filtering:** We remove a set of markup characters such as "&lt", "&gt", "amp" , etc. that frequently occur within the harvested instances. URLs are also removed and replaced with "@http" tokens. We further remove all non-ASCII characters that may happen within the social media text; hence, preventing their negative effect in the summarization process. We further replace the user IDs or users' names with "@user" tokens to hinder the possibility of users' identities being disclosed.

2. **Instance filtering** We define two instance sampling criteria which should be met by each instance to be included in the final MENTSUM dataset. First, we identify the most frequent word bigram of the post's TLDR; if it occurs more than 3 times (empirically determined), we exclude the instance from the final dataset, otherwise, we keep the instance. This is based on our observation that TLDR summaries with more than 3 identical bigrams contain redundant information, not conveying enough information about the posted content in a short summary. Second, we apply a filtering rule based on the *compression ratio* [5] of instances. Specifically, we only retain instances whose compression ratio falls in the range of [2-13] [6] (i.e., user's post should be between 2x and 13x longer than the associated TLDR summary). This filtering decision is based on the notion that to have short summaries we do not want too small of compression ratio; to have informative enough summaries we do not want too large of compression ratio.

The pipeline that we mentioned above reduced the initial set of 42k instances to the final dataset with 24,119 English post-TLDR user-written summary pairs.

Dataset size	24,119 posts
Training set size	21,695 posts
Subreddit coverage	43 subreddits
Average post length (word/sent.)	327.5 / 16.9
Min post length (word/sent.)	50 / 1
Max post length (word/sent.)	2,979 / 192
Median post length (word/sent.)	267.0 / 14.0
Average TLDR length (word/sent.)	43.5 / 2.6
Min TLDR length (word/sent.)	15 / 1
Max TLDR length (word/sent.)	596 / 36
Median TLDR length (word/sent.)	35.0 / 2.0
Average sentence length (post/TLDR)	19.4 / 18.3 words
Median sentence length (post/TLDR)	17.0 / 16.0 words
Compression ratio (average)	7.5

(a)

Total vocabulary size	76,411 words
Occurring 10+ times	17,486 words
Training vocabulary size	49,037 words
Validation vocabulary size	13,609 words
Test vocabulary size	13,765 words
Training/Test vocabulary overlap	91.4%

(b)

Table 1: (a) Statistics of MENTSUM dataset. (b) Vocabulary statistics over distinct and uncased vocabulary terms.

3.2. Dataset statistics

The overall dataset statistics are shown in Tables 1 (a). As shown, the compression ratio of this dataset amounts to 7.5 which shows that TLDR summaries are the extremely short version of their associated users' post. Table 1 (b) presents the statistics on distinct and uncased vocabularies included in the dataset. As shown, about 23% of the vocabularies occur more than 10 times within the dataset. Figure 2 depicts the count and percentage of the posts for top 10 subreddits that have the highest count of posts. As indicated, ADHD has the highest count of posts in MENTSUM, followed by depression, anxiety, SuicideWatch, and the rest[7]. In order to make different sets for training and evaluating summarization models, we randomly divide the data into 21,695 training (90%), 1,209 validation (5%) and 1,215 test (5%) instances.

3.3. Dataset analysis

In this section, we provide an analysis of MENTSUM dataset in terms of lead bias, abstractive, and extractive characteristics of data.

Lead bias. Lead bias (Hong and Nenkova, 2014) is a common phenomenon in News summarization

[4] Subreddits are available at https://ir.cs.georgetown.edu/resources/data/mentsum/

[5] Compression ratio = $\frac{\text{count of words in user's post}}{\text{count of words in TLDR}}$

[6] [2-13] was decided empirically in our experiments.

[7] While our dataset covers a range of 43 mental health subreddits, we only show top 10 of them in terms of post frequency due to space constraints.

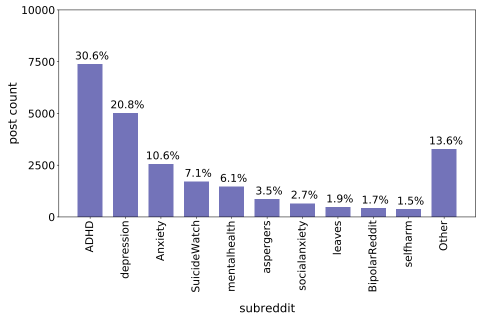

Figure 2: Count (y-axis) and proportion (above each bar) of posts of top 10 frequent subreddits in MENTSUM dataset. "Other" includes the posts in the remaining 33 mental health subreddits.

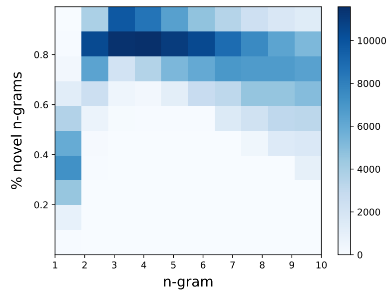

Figure 4: Percentage of novel n-grams (y-axis) across increasing word n-grams (x-axis) in user-written TLDRs. The heat extent shows the frequency of word n-grams for a certain percentage bin.

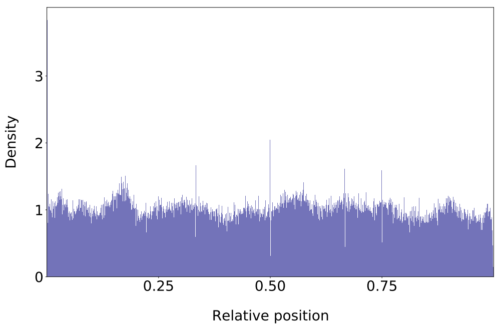

Figure 3: The relative position of word bigrams of the user-written TLDR summary across users' post in MENTSUM dataset.

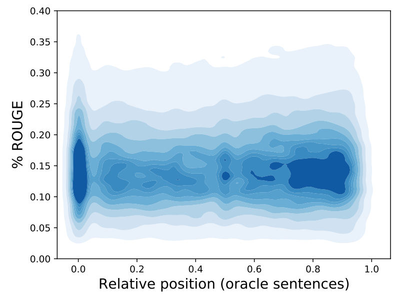

Figure 5: Kernel density estimation diagram of oracle sentences' relative position and the ROUGE score of oracle sentences in respect to user-written TLDRs.

datasets, where the early parts of the source document contain the most salient information. However, this characteristic does not hold in the social media posts as the salient information are scattered in the entire user's post (Kim et al., 2019); hence, imposing a challenge for the summarization task. Figure 3 demonstrates the relative position of word bigrams of the TLDR summary within users' posts in MENTSUM dataset. As observed, the TLDR summary's bigrams are uniformly distributed along the users' post text, exhibiting a weak lead bias.

Abstractiveness. We further measure the abstractiveness of the purposed MENTSUM dataset to verify its applicability for abstractive text summarization models. Figure 4 shows the percentage of novel words for different n-grams. As observed, the heat is mainly populated in the upper bins (i.e., higher probabilities), particularly for $\{n|n = 2, 3, 4, 5, 6\}$, which shows strong abstractive characteristics of the MENTSUM dataset, making it suitable for abstractive summarization task.

Extractiveness. Figure 5 shows the density estimation diagram of oracle sentences' relative position in the users' post in regard to the ROUGE score [8] of the oracle sentences. Oracle sentences are up to 3 summary-worthy sentences which are labelled using a greedy labeling approach proposed in Liu and Lapata (2019). As indicated, the oracle sentences appear across various positions of the users' post. Considering the diagram and the range of oracle sentences' scores, MENTSUM dataset still expresses extractive characteristics in addition to its high abstractiveness.

3.4. Ethics and privacy

Although we use publicly available Reddit data in our research to construct the MENTSUM dataset, mental health is a sensitive topic and special care should be taken when such data is used in social media research (Thomas et al., 2002; Moreno et al., 2013; Suster et al., 2017; Benton et al., 2017; Nicholas et al., 2020). Hence, we made no attempt to identify and contact the users, or discover user relations with other social media accounts. In preprocessing step, we have de-identified the usernames with `@user` tokens to prevent the user identities from being known. The MENTSUM dataset can be accessed through a Data Usage Agreement (DUA) [9]. The DUA particularly ensures that no attempts should be made to distribute portions of dataset (which could result in revealing users' identity), identify users, and contact users.

[8] We have taken average of ROUGE-1, ROUGE-2, and ROUGE-L scores w.r.t the TLDR summary.

[9] https://ir.cs.georgetown.edu/resources/

4. Experimental setup

To evaluate the quality of MENTSUM dataset for the summarization task of mental health related posts, and to provide strong baselines for further research, we explored several baselines. In this section, we present the baselines and the implementation details.

4.1. Baselines

We explored various extractive and abstractive baselines which are listed below.

- **LEAD-2**: A simple extractive baseline that selects the first two leading sentences as the summary.
- **LSA** (Steinberger and Ježek, 2004): A non-neural extractive vector-based model that adopts the mathematical concept of Singular Value Decomposition (SVD) to find hidden semantic structures of words and sentences.
- **LEXRANK** (Erkan and Radev, 2004): An unsupervised extractive model that makes use of graph centrality network to find important sentences and concatenate them to form the final summary.
- **BERTSUMEXT** (Liu and Lapata, 2019): A neural extractive model that fine-tunes BERT (Devlin et al., 2019) language model on extractive summarization text. This model runs by appending `[CLS]` tokens to the start of each input sentence, and use the representations associated with `[CLS]` tokens to predict sentence importance. `[CLS]` is the classification head in BERT model that aggregates the contextualized embeddings of preceding tokens.
- **MATCHSUM** (Zhong et al., 2020): A state-of-the-art extractive summarization model that first composes candidate summaries given the salient set of source sentences scored by BERTSUMEXT model, and then ranks them using the Siamese neural networks. Top-ranked candidate summary is retrieved as the final extractive summary of the post.
- **BERTSUMABS** (Liu and Lapata, 2019): The abstractive variant of BERTSUM framework, where the encoder is simply a BERT model, which is trained alongside a Transformers-based (Vaswani et al., 2017) decoder from scratch.
- **BERTSUMEXTABS** (Liu and Lapata, 2019): A two-stage fine-tuned abstractive model that exploits a pre-trained BERTSUMEXT summarizer (i.e., first stage) which is further fine-tuned along with a decoder on abstractive summarization task (i.e., second stage).
- **BART** (Lewis et al., 2020): An abstractive model that is currently amongst the most powerful state-of-the-art summarization models. BART extends the BERT's intuition by adding up a couple of pre-training objectives including token deletion, text in-filling, sentence permutation, and document rotation. Unlike BERT, BART utilizes a pre-trained encoder-decoder framework for language generation task, summarization being one of them.

4.2. Implementation details

We used Sumy package [10] for running non-neural extractive models. To find extractive oracle labels (i.e., important sentences) of users' post, we ran a greedy labeling approach (Liu and Lapata, 2019) over the entire set of source sentences and retrieved up to 3 sentences as the extractive summary. For BERTSUM models, we used the official codebase [11] with `BERT-base-uncased` and ran all of the models with default hyper-parameters as suggested by Liu and Lapata (2019) besides the learning rate of $1e-3$ and warmup steps of 5k. For MATCHSUM, we used `RoBERTa-base` as the encoder with the same default hyper-parameters as initialized in the original paper (Zhong et al., 2020) and candidate summaries of lengths 2 and 3. We utilized Huggingface Transformers' (Wolf et al., 2020) implementation for training BART model. All the experimented neural models were trained for 5 epochs and then the best checkpoint that attains the highest ROUGE-L score during validation was picked for inference time. As the optimizer, we used AdamW (Loshchilov and Hutter, 2019) initialized with learning ratio of $3e-5$, $(\beta_1, \beta_2) = (0.9, 0.98)$, and a weight decay of 0.01. Cross-entropy loss was employed for all of the experimented models. We also utilized Weights & Biases toolkit (Biewald, 2020) to keep track of training and validation progress.

5. Results and discussion

Table 2 presents the performance of summarization models on the test set of MENTSUM dataset in terms of widely adopted ROUGE (F1) metric (Lin, 2004) that measures the text overlap between the system-generated TLDR summaries and user-written TLDR summaries. As expected and was shown in the data analysis, LEAD-2 is unable to perform well on MENTSUM dataset. As addressed earlier, this is due to scattered salient information throughout the post.

Comparing extractive models with the abstractive ones, we notice that extractive models lag behind the best performing abstractive model (i.e., BART) by a huge margin; this is not surprising as summaries are paraphrased by users, as also confirmed by Figure 4 in Section 3.3. Even when comparing abstractive variants of BERTSUM model (i.e., BERTSUMABS and BERTSUMEXTABS) with the extractive models, we see a similar trend of gaining more significant improvements via abstractive models.

While comparing the extractive models with each other, we observe that BERTSUMEXT model considerably outperforms the other unsupervised models (i.e., LSA amd LEXRANK), though, it lags behind the

[10] `https://github.com/miso-belica/sumy`

[11] `https://github.com/nlpyang/PreSumm`

Method	RG-1	RG-2	RG-L
LEAD-2	20.09	3.57	13.70
ORACLEEXT	35.98	11.59	23.21
LSA (Steinberger and Ježek, 2004)	22.96	3.98	14.50
LEXRANK (Erkan and Radev, 2004)	23.21	4.42	15.19
BERTSUMEXT (Liu and Lapata, 2019)	24.64	5.83	16.66
MATCHSUM (Zhong et al., 2020)	26.29	6.32	17.12
BERTSUMABS (Liu and Lapata, 2019)	25.35	6.49	17.11
BERTSUMEXTABS (Liu and Lapata, 2019)	25.75	6.80	17.48
BART (Lewis et al., 2020)	**29.13**	**7.98**	**20.27**

Table 2: ROUGE (F1) results on MENTSUM's test set.

MATCHSUM model as the best-perfomring extractive model. This shows the effect of learning salient sentences in supervised manner, accounting for the improvements of both BERTSUMEXT and MATCHSUM models. Noting that ORACLEEXT shows the upper bound performance of extractive summarization models, it is noticeable that we see a large gap between the extractive systems and ORACLEEXT's performances. This observation is a clear justification that there is still a large room for improvement in extractive summarization setting, which can be further approached by future work.

While looking at the performances of abstractive summarization models, we see a remarkable improvement of BART over other systems; particularly, with relative improvements of 13.12% (RG-1), 17.35% (RG-2), and 15.96% (RG-L) compared to BERTSUMEXTABS model [12]. This is expected as BART uses pre-trained encoder-decoder framework, unlike abstractive models of BERTSUM that only employ the pre-trained encoder (i.e., BERT) while training the decoder from scratch. It is also observed that BERTSUMEXTABS model surpasses BERTSUMABS model, showing the effectiveness of extractive objectives in abstractive summary generation process as also suggested by Gehrmann et al. (2018).

Figure 6 demonstrates the percentage of novel n-grams in BART-generated TLDRs over the test set.[13] As indicated, with increasing n-grams, the rate of novel words generation grows, showing that BART does the abstraction by rephrasing/paraphrasing the source sentences. Also, compared to the user-written TLDRs, BART does more extraction (on word level) than users do; however, it is still hard to attribute this observation to the quality of generated TLDRs as the goldness of ground-truth TLDRs may be influenced by the emotional state of the author at the time of writing as noted in Section 3.1.

[12]The relative improvement of system A over system B is calculated as $\frac{\text{Rouge}_A - \text{Rouge}_B}{\text{Rouge}_B}$

[13]The novel n-grams diagram on user-written TLDRs in test set is quite similar to Figure 4 and is not shown due to space limitations.

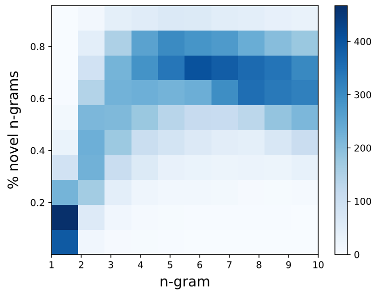

Figure 6: Percentage of novel n-grams across increasing word n-grams in BART's system-generated TLDRs on test set. The heat extent shows the frequency of word n-grams for a certain percentage bin.

6. Human analysis

A few prior studies have recognized the limitations of widely used ROUGE metric on qualitative evaluation as it is rather biased towards surface lexical similarities (Ng and Abrecht, 2015; Cohan and Goharian, 2016). To shed light on the qualities and limitations of the state-of-the-art summarizer (i.e., BART), and provide insightful directions for future work, we carried out a human evaluation study. To this end, we randomly selected 100 posts along with their associated user-written and system-generated TLDR summaries from the test set. Following prior work (Grusky et al., 2018; Zhang et al., 2020; Cho et al., 2021; Sotudeh et al., 2021), we defined three qualitative metrics: **(1) Fluency:** is the summary well-written and easy to understand?; **(2) Informativeness:** does the TLDR summary provide useful information (i.e., the most important information) about user's post?; **(3) Conciseness:** does the summary briefly provide comprehensive information (i.e., majority of important information) about user's post? Two human annotators evaluated the provided cases using 5-point Likert scale (1 = worst, 5 = best). For disagreement cases, where the assigned scores are 2 levels or more different from each other, a third annotator broke the disagreement and made the final decision. We then provided selected cases through

System	Fluency	Info.	Conc.
BART	**4.60**	**3.65**	**3.51**
User	4.28	3.46	3.15

Table 3: Results of the human evaluation comparing the systems in terms of Fluency, Informativeness, and Conciseness. Winning scores are shown in bold. Scores are in 5-point Likert scale (1=worst, 5=best).

System	Fluency	Info.	Conc.
BART	17.4%	28.2%	28.0%
User	19.5%	25.2%	24.6%

Table 4: System-wise Cohen's kappa inter-rater agreement.

the following multi-stage pipeline:

1. In the first stage, we provided 100 posts with one randomly selected summary (from either user-written or system-generated), and asked the annotators to independently score the given summary in terms of the qualitative criteria.

2. In the second stage, we provided the same 100 posts but with the other summary that was not shown to the annotators in the first stage, and asked the annotators to independently evaluate the given summary on the qualitative criteria.

3. In the final stage, we provided the same 100 posts with both summaries (i.e., user-written and system-generated) side-by-side, shuffled, non known to annotators and asked the annotators to specify which summary they prefer the most. The annotators were further asked to stipulate which major pieces of information are captured by or missing from each given summary. We will discuss the details of this stage in Section 7 (i.e., Error Analysis).

While stages (1) and (2) could have been combined into one stage, to avoid any bias in scoring these two summaries, we decided it was better that the annotators would score each of these two summaries independently without any comparison. It has to be mentioned that the order of summaries was shuffled; that is, the annotators were not aware of which of the two summaries they were evaluating.
The evaluation scores are averaged and shown in Table 3. As shown, the BART summarizer outperforms the user-written TLDR summaries across all metrics with relative improvements of 7.47% (Fluency), 5.49% (Informativeness), and 11.42% (Conciseness). Observing the significant improvement gain by the state-of-the-art summarizer, it does seem interesting that contextualized language modelling based approaches such as BART are becoming a firm touchstone to be compared with the human system. This finding is consistent with the observations that have been made recently by a prior work (Fabbri et al., 2021).
Table 4 reports the Cohen's kappa (Cohen, 1960) inter-rater agreement for both system-generated and user-written TLDR summaries. With regard to the Cohen's kappa range interpretation (McHugh, 2012), the agreements obtained on informativeness and Conciseness metrics fall into the "fair" agreement, while there is a "slight" agreement observed on fluency. The slight agreement on fluency could be attributed to the subjective nature of this metric in evaluation process (van der Lee et al., 2021), leading to a high variability in assigned scores by the annotators. In addition to the overall promising performance of BART, we observed that the nature of data is also impactful in the process of human study. In other words, users in social media are free to publish their content of discussion in whatever style they like to as there is usually not any supervision over the posted content in terms of quality, which leads to having gibberish TLDRs in some cases.
As the results of stage (3) of the evaluation, the annotators reported that they prefer system-generated summaries in 59% of the cases, and user-written summaries in 41% of the cases with an agreement rate of 59.6% (moderate).

System	Win rate	Fluency	Info.	Conc.
BART	59%	17%	64%	19%
User	41%	8%	81%	11%

Table 5: Distribution of leading criteria for summary selection in human evaluation process.

7. Error analysis

In this section, we showcase where the system summarizer lacks in or improves compared to the user-written summaries in more detail. As reported by the annotators, the distribution of qualitative criteria for determining which summary is more preferable is shown in Table 5. For instance, annotators pick the BART-generated summaries in 59% of cases. In 64% of BART's *win* cases, the system-generated summary was preferred due to its informativeness as compared to the user-written summary. As indicated, out of 3 evaluation metrics, informativeness is the most dominant one in the selection of BART-generated and user-written summaries.
We report the findings for each of the three metrics of fluency, informativeness, and conciseness [14]:

- **Fluency.** Few reasons contributed to annotators' lower score for user-written TLDRs in terms of fluency. One of them is grammatical issues in the users' language in user-written TLDR summaries. Examples of such grammatical error is "`...good`

[14]Examples are truncated in the paper for extra privacy of the users.

`friend and ex girlfriend has...`". Another observation showed the complex structure of the user-written sentence which makes the summary less understandable such as in "`...she can't be arsed trying to communicate with myself...`". Interestingly, the system-generated TLDRs did not suffer from grammatical errors. This might be due to the fact that the system summarizer is exposed to the structured, grammatical, and fluent textual data from books and Wikipedia during pre-training, and hence with further fine-tuning on the MENTSUM dataset, it is yet able to produce understandable TLDRs. On the other hand, for the cases where system-generated summaries underperform the user-written ones, we observed examples of repetition such as in "`i'm having massive mood swings and mood swings...`", and complexity without specifying sentence boundaries such as in "`...i was @ageX i am @ageY now and i haven't been able to find a therapist until my @ageZ birthday and i want to offer myself...`".

- **Informativeness.** We observed that the users provide some history of their mental health and disorders within their posts. It is important for these vital information to be captured in the system-generated summaries. Interestingly, the system summarizer attends more to such mental disorders and includes them into the generated summaries. For instance, in the system-generated summary "`...diagnosed with anxiety, GAD, and agoraphobia...`" all mental illnesses of the user are captured. However, the user-written TLDR "`...was diagnosed with anxiety...`" does not include "`GAD, and agoraphobia`". Furthermore, we observed that parts of user-written summaries contain information that is not directly mentioned in the user's post, but was inferred from it. Such cases are still a challenge for the system summarizers. As an example, consider user-written TLDR of "`...I'm holding it all in which is making me and everyone around me worse...`", which conveys information that is not directly mentioned in the post, but could be inferred from the post. While user-written TLDR contains such salient information, the system-generated TLDR was not able to infer and generate that. In underperformed cases, we observed that the model focuses on a few (one or two) important points of the user's post, while the user-written TLDR includes more important points from diverse parts of the user's post.

- **Conciseness.** The aforementioned limitations of informativeness affect the conciseness negatively. That is, some of important information that are required to be in a concise summary are missed. In particular, for the cases in which user-written TLDRs are scored lower than system-generated ones in terms of conciseness, the user-written TLDRs are verbose and very detailed. However, the system is able to skip unnecessary information, focus on discerning important information, and verbalize them within a comparably shorter text. For instance, consider system-generated TLDR "`...been on meds and seeing a psych for a while, now finally going to a dr for help...`" in comparison with its user-written TLDR "`...saw a psych about it but then stopped. felt like garbage last year. seeing a doctor about it now. happy hopeful noises...`". As seen, the system verbalizes the important points using a comparably briefer usage of language.

8. Summary and conclusion

Mental health and illnesses have become a major challenge in public health. The prevalence of user-generated and user-curated content in online social media platforms has provided a large amount of accessible information for those who seek to study user behavioural patterns in social settings. While most previous research studies in mental health domain have focused on developing datasets for the classification tasks to triage and detect the type and severity of mental health concerns, there has not been any dataset to support the research in summarizing mental health related social media posts. In this paper, we introduced our large-scale novel dataset resource, MENTSUM for the task of summarization of users' mental health posts on Reddit social media. Although the dataset is gathered from 43 publicly available Reddit subreddits, it will be provided to researchers through a Data Usage Agreement (DUA) to protect the privacy of the users. Our detailed dataset analyses revealed characteristics such as weak lead bias, strong abstractiveness, and extractiveness. We further evaluated various state-of-the-art extractive and abstractive summarization models on MENTSUM dataset. This study showed that while the abstractive method (i.e., BART) outperformed the extractive models, the gap between the extractive models' performance and ORACLEEXT (i.e., the upper bound performance of extractive methods) indicates a large room for improvement in extractive settings, calling for future research. Our human evaluation over 100 randomly selected post summary pairs in terms of fluency, informativeness, and conciseness of system-generated and user-written summaries revealed interesting information; the system-generated summaries were preferred on average over user-written ones, showing the promising performance of contextualized language modelling based summarization approaches. Finally, we provided an error analysis to show the current challenges. We hope our provided MENTSUM dataset paves the path for further exploration in summarization research of social media mental health posts.

9. Bibliographical References

Baldwin, T., Cook, P., Lui, M., MacKinlay, A. D., and Wang, L. (2013). How noisy social media text, how diffrnt social media sources? In *IJCNLP*.

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., and Blackburn, J. (2020). The pushshift reddit dataset. In *ICWSM*.

Benton, A., Coppersmith, G. A., and Dredze, M. (2017). Ethical research protocols for social media health research. In *EthNLP@EACL*.

Biewald, L. (2020). Experiment tracking with weights and biases. Software available from wandb.com.

Chancellor, S., Lin, Z. J., Goodman, E. L., Zerwas, S., and Choudhury, M. D. (2016). Quantifying and predicting mental illness severity in online proeating disorder communities. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*.

Cho, S., Dernoncourt, F., Ganter, T., Bui, T., Lipka, N., Chang, W., Jin, H., Brandt, J., Foroosh, H., and Liu, F. (2021). Streamhover: Livestream transcript summarization and annotation. *ArXiv preprint*, abs/2109.05160.

Choudhury, M. D. and De, S. (2014). Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *ICWSM*.

Choudhury, M. D., Gamon, M., Counts, S., and Horvitz, E. (2013). Predicting depression via social media. In *ICWSM*.

Cohan, A. and Goharian, N. (2016). Revisiting summarization evaluation for scientific articles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 806–813, Portorož, Slovenia. European Language Resources Association (ELRA).

Cohan, A., Young, S., Yates, A., and Goharian, N. (2017). Triaging content severity in online mental health forums. *Journal of the Association for Information Science and Technology*, 68.

Cohan, A., Desmet, B., Yates, A., Soldaini, L., MacAvaney, S., and Goharian, N. (2018). Smhd: a large-scale resource for exploring online language usage for multiple mental health conditions. In *COLING*.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37 – 46.

Coppersmith, G. A., Dredze, M., and Harman, C. (2014a). Quantifying mental health signals in twitter. In *CLPsych@ACL*.

Coppersmith, G. A., Harman, C., and Dredze, M. (2014b). Measuring post traumatic stress disorder in twitter. In *ICWSM*.

Coppersmith, G. A., Leary, R., Crutchley, P., and Fine, A. B. (2018). Natural language processing of social media as screening for suicide risk. *Biomedical Informatics Insights*, 10.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Erkan, G. and Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.*, 22:457–479.

Fabbri, A. R., Kryscinski, W., McCann, B., Socher, R., and Radev, D. (2021). Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Garg, S., Taylor, J., Sherief, M. E., Kasson, E., Aledavood, T., Riordan, R., Kaiser, N. M., Cavazos-Rehg, P. A., and Choudhury, M. D. (2021). Detecting risk level in individuals misusing fentanyl utilizing posts from an online community on reddit. *Internet Interventions*, 26.

Gehrmann, S., Deng, Y., and Rush, A. (2018). Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, Brussels, Belgium. Association for Computational Linguistics.

Goldstein, A. and Shahar, Y. (2016). An automated knowledge-based textual summarization system for longitudinal, multivariate clinical data. *Journal of biomedical informatics*, 61:159–75.

Grusky, M., Naaman, M., and Artzi, Y. (2018). Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Hao, B., Li, L., Li, A., and Zhu, T. (2013). Predicting mental health status on social media - a preliminary study on microblog. In *HCI*.

Holmes, E. A., O'Connor, R. C., Perry, V. H., Tracey, I., Wessely, S., Arseneault, L., Ballard, C. G., Christensen, H., Silver, R. C., Everall, I., Ford, T. J., John, A., Kabir, T., King, K., Madan, I., Michie, S., Przybylski, A. K., Shafran, R., Sweeney, A., Worthman, C. M., Yardley, L., Cowan, K. C., Cope, C., Hotopf, M., and Bullmore, E. T. (2020). Multidisciplinary research priorities for the covid-19 pandemic: a call for action for mental health science. *The Lancet. Psychiatry*, 7:547 – 560.

Hong, K. and Nenkova, A. (2014). Improving the estimation of word importance for news multi-document summarization. In *EACL*.

Kim, B., Kim, H., and Kim, G. (2019). Abstractive summarization of Reddit posts with multi-level memory networks. In *Proceedings of the 2019 Con-*

ference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 2519–2531, Minneapolis, Minnesota. Association for Computational Linguistics.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Liu, Y. and Lapata, M. (2019). Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Liu, H., Morstatter, F., Tang, J., and Zafarani, R. (2016). The good, the bad, and the ugly: uncovering novel research opportunities in social media mining. *International Journal of Data Science and Analytics*, 1:137–143.

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

MacAvaney, S., Desmet, B., Cohan, A., Soldaini, L., Yates, A., Zirikly, A., and Goharian, N. (2018). Rsdd-time: Temporal annotation of self-reported mental health diagnoses. In *CLPsych@NAACL-HTL*.

MacAvaney, S., Sotudeh, S., Cohan, A., Goharian, N., Talati, I. A., and Filice, R. W. (2019). Ontology-aware clinical abstractive summarization. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Manas, G., Aribandi, V., Kursuncu, U., Alambo, A., Shalin, V. L., Thirunarayan, K., Beich, J., Narasimhan, M., and Sheth, A. (2021). Knowledge-infused abstractive summarization of clinical diagnostic interviews: Framework development study. *JMIR Mental Health*, 8.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, 22:276 – 282.

Milne, D. N., Pink, G., Hachey, B., and Calvo, R. A. (2016). Clpsych 2016 shared task: Triaging content in online peer-support forums. In *CLPsych@HLT-NAACL*.

Mishra, R., Bian, J., Fiszman, M., Weir, C. R., Jonnalagadda, S. R., Mostafa, J., and Fiol, G. D. (2014). Text summarization in the biomedical domain: A systematic review of recent research. *Journal of biomedical informatics*, 52:457–67.

Moreno, M. A., Goniu, N., Moreno, P. S., and Diekema, D. (2013). Ethics of social media research: Common concerns and practical considerations. *Cyberpsychology, behavior and social networking*, 16 9:708–13.

Mowery, D. L., Bryan, C. J., and Conway, M. (2017a). Feature studies to inform the classification of depressive symptoms from twitter data for population health. *ArXiv*, abs/1701.08229.

Mowery, D. L., Smith, H., Cheney, T., Stoddard, G., Coppersmith, G. A., Bryan, C. J., and Conway, M. (2017b). Understanding depressive symptoms and psychosocial stressors on twitter: A corpus-based study. *Journal of Medical Internet Research*, 19.

Ng, J.-P. and Abrecht, V. (2015). Better summarization evaluation with word embeddings for ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1925–1930, Lisbon, Portugal. Association for Computational Linguistics.

Nicholas, J., Onie, S., and Larsen, M. E. (2020). Ethics and privacy in social media research for mental health. *Current Psychiatry Reports*, 22:1–7.

O'Dea, B., Wan, S., Batterham, P. J., Calear, A. L., Paris, C., and Christensen, H. (2015). Detecting suicidality on twitter. *Internet Interventions*, 2:183–188.

Otu, A. A., Charles, C. H., and Yaya, S. (2020). Mental health and psychosocial well-being during the covid-19 pandemic: the invisible elephant in the room. *International Journal of Mental Health Systems*, 14.

Pfefferbaum, B. and North, C. S. (2020). Mental health and the covid-19 pandemic. *The New England journal of medicine*.

Resnik, P., Garron, A., and Resnik, R. (2013). Using topic modeling to improve prediction of neuroticism and depression in college students. In *EMNLP*.

Soldaini, L., Walsh, T., Cohan, A., Han, J., and Goharian, N. (2018). Helping or hurting? predicting changes in users' risk of self-harm through online community interactions. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 194–203, New Orleans, LA, June. Association for Computational Linguistics.

Sotudeh, S., Goharian, N., and Filice, R. (2020). Attend to medical ontologies: Content selection for clinical abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1899–1905, Online. Association for Computational Linguistics.

Sotudeh, S., Deilamsalehy, H., Dernoncourt, F., and Goharian, N. (2021). TLDR9+: A large scale resource for extreme summarization of social media posts. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 142–151, Online and in Dominican Republic, November. Association for Computational Linguistics.

Steinberger, J. and Ježek, K. (2004). Using latent semantic analysis in text summarization and summary evaluation. In *ISIM*.

Suster, S., Tulkens, S., and Daelemans, W. (2017). A short review of ethical challenges in clinical natural language processing. In *EthNLP@EACL*.

Thomas, J. C., Sage, M., Dillenberg, J., and Guillory, V. J. (2002). A code of ethics for public health. *American journal of public health*, 92 7:1057–9.

van der Lee, C., Gatt, A., van Miltenburg, E., and Krahmer, E. J. (2021). Human evaluation of automatically generated text: Current trends and best practice guidelines. *Comput. Speech Lang.*, 67:101151.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In Isabelle Guyon, et al., editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Völske, M., Potthast, M., Syed, S., and Stein, B. (2017). TL;DR: Mining Reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63, Copenhagen, Denmark. Association for Computational Linguistics.

Wang, X., Zhang, C., and Sun, L. (2013). An improved model for depression detection in micro-blog social network. *2013 IEEE 13th International Conference on Data Mining Workshops*, pages 80–87.

Wilson, M. L., Ali, S., and Valstar, M. F. (2014). Finding information about mental health in microblogging platforms: a case study of depression. *Proceedings of the 5th Information Interaction in Context Symposium*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yates, A., Cohan, A., and Goharian, N. (2017). Depression and self-harm risk assessment in online forums. In *EMNLP*.

Zhang, Y., Merck, D., Tsai, E., Manning, C. D., and Langlotz, C. (2020). Optimizing the factual correctness of a summary: A study of summarizing radiology reports. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5108–5120, Online. Association for Computational Linguistics.

Zhong, M., Liu, P., Chen, Y., Wang, D., Qiu, X., and Huang, X. (2020). Extractive summarization as text matching. In *ACL*.