SigBART: Enhanced Pre-training via Salient Content Representation Learning for Social Media Summarization

Sajad Sotudeh
IRLab, Georgetown University
Washington, D.C., USA
sajad@ir.cs.georgetown.edu

Nazli Goharian
IRLab, Georgetown University
Washington, D.C., USA
nazli@ir.cs.georgetown.edu

ABSTRACT
Our approach to automatically summarizing online mental health posts could help counselors by reducing their reading time, enabling quicker and more effective support for individuals seeking mental health assistance. Neural text summarization methods demonstrate promising performance owing to their strong pre-training procedure. Random token/span masking technique is often relied upon by existing pre-trained language models; an approach that overlooks the importance of content when learning word representations. In an attempt to rectify this, we propose using source and summary alignments as a saliency signal to enhance the pre-training strategy of language model for better representation learning of important content, paving the way for a positive impact on the model fine-tuning phase. Our experiments on a mental health-related dataset for user post summarization (MENTSUM) reveal improved performance, as evidenced by human evaluation metrics, surpassing the current state-of-the-art system.

CCS CONCEPTS
• Computing methodologies → Natural language generation.

KEYWORDS
Text Summarization, Social Media Text Summarization, Pretrained Text Summarization

ACM Reference Format:

1 INTRODUCTION AND RELATED WORK
Social media platforms such as Reddit have become popular spaces for people to discuss a range of topics, including mental health. These forums are crucial for individuals seeking mental health support or advice. However, for mental health professionals navigating and understanding the vast amount of user-generated content is challenging [4, 9, 23, 26]. Creating summaries of these posts is not only a technical task but also a means to grasp the nuance aspects of mental health more effectively [13, 20, 24, 33]. In this context, summarization efforts aim to create concise, abstractive summaries, known as TL;DR (Too Long; Didn’t Read), from user posts.

The majority of research in the mental health domain has primarily concentrated on predictive tasks [6–8, 18, 22, 32]. However, the development of the MentSum dataset, as introduced in [25], has made the task viable. Building on this, [24] achieved the state-of-the-art (SOTA) performance by implementing a curriculum-guided abstractive summarizer on this dataset. In a broader context, pre-trained language models [10, 16, 17, 30] have revolutionized NLP by two-step process of pre-training and fine-tuning. Existing models [1, 2, 14] in the social media domain have primarily focused on fine-tuning. However, none have focused specifically on pre-training the language model for summarization-specific purposes in the social media domain. Our work fills this gap by enhancing the pre-training stage, identifying salient content, and masking it for representation learning. Despite some work showing that further pre-training on in-domain data can enhance model performance [12, 15, 29], none have considered content importance 1. Our work is the first to incorporate content saliency in the pre-training stage for the social media domain. We identify important content by aligning it with the TL;DR summary, mask it, and have the pretrained language model (i.e., BART) predict these masked spans. This replaces BART’s random text infilling pre-training objective with masked important content. We then use this pre-trained model to fine-tune the state-of-the-art summarizer, BART + R3F [2], for mental-health social media summarization. The work that comes closest to ours is PEGASUS [30], which masks the top-m important sentences. However, we diverge by using token alignments with the summary as a measure of content saliency. Our evaluations reveal our method outperforms the SOTA system in fluency, informativeness, and faithfulness, while also improving automatic evaluation metrics. In summary, our contributions include proposing and evaluating a novel model that pre-trains the language model on masked salient content over social media data to improve mental health post summarization.

2 APPROACH
Aiming to improve important span representations (as defined earlier), we replace BART’s typical random masking approach with salient span masking. We explain the steps in the following in more detail.

Identification of salient spans. To identify the salient spans, we employ two orthogonal methods to make alignments between

1In our framework, content importance in pre-training is defined as those spans of text that are likely to be present in summary.
users’ posts and their associated TLDR summary. Specifically, we frame the alignment problem as a sequence-tagging task, where the post terms are tagged if they are copied into the TLDR. Formally written, let $b = \{b_0, b_1, \ldots, b_n\}$ be the binary tags over the post words (excluding stop words) $x = \{x_0, x_1, \ldots, x_n\}$ and TLDR terms $y = \{y_0, y_1, \ldots, y_m\}$, with $n$ and $m$ being the length of the user post and TLDR, respectively. To align the user post to the TLDR, we denote the word $x_i$ as copied if (1) it is included in the longest possible sub-sequence of words $s = \{x_{i-j-i,j+i+k}\}$ $(j \leq i, k \leq (n-i))$ $(s \in x, s \in y)$, and (2) there exists no earlier sequence $u$ with $s = u$. While the proposed method generates supervised labels for token masking, it does not guarantee that enough labels will be generated due to the abstractive nature of the pretraining dataset. To address this issue, we identify frequently occurring spans across the data collection and annotate their occurrence within each sample. Our investigations show that around 32% of tokens (on average) are identified as being part of the salient spans.  

**Pre-training Bart on salient spans.** Our goal is to learn effective representations of the salient content (i.e., odds-on-copied spans as discussed in Section 2) for better summarization in the fine-tuning stage. In this sense, after identifying and masking salient spans, we pre-train the Bart model over the masked spans. A sample of pre-training iteration is shown in Figure 1. The Bart model is trained to generate the output sequence by predicting the masked tokens within the input. We choose to pre-train the model using text infilling pre-training objective, considering its effectiveness as reported by Lewis et al. [16]. Our pre-training strategy, as well as the associated pre-trained checkpoint, is referred to as SigBART, which is short for Significant Bart.

### 3 EXPERIMENTAL SETUP

The experimental setup is set to address the following research questions:

- **RQ1:** Does pre-training on salient content/spans lead to improved summaries in terms of automatic and qualitative metrics as compared to the SOTA Bart + R3F model?  
- **RQ2:** Does the proposed pre-training method lead to improvements as compared to the existing baselines that differ in pre-training and fine-tuning stages, in terms of ROUGE evaluation metrics?

**Datasets.** We use WEBS-TLDR-17[27] for pre-training, which comprises 3.8M user posts and their paired TLDR summaries. For fine-tuning, we use MENTSUM [25] with over 24k mental-health post, summary pairs. For more information on the datasets, please refer to the respective papers.

**Comparison.** We compare our model with several extractive and abstractive baselines as listed below. The Bart + R3F baseline is for evaluating RQ1, and the rest for RQ2.

- **Curriculum** [24]: a SOTA curriculum-guided abstractive model that exploits SUPERLOSS (a confidence-aware curriculum loss) as a progressive training signal to enhance the learning efficiency and performance of the summarization model.
- **BART + R3F** [2]: a SOTA abstractive system that discourages representation change during fine-tuning when possible without hurting the system performance. In the experiment, we include this model to assess the impact of SigBart’s pre-training strategy in comparison to a scenario when it is not employed (i.e., RQ1).
- **MATCHSUM** [31]: an extractive summarizer that matches text spans of source and summary in semantic space for selecting the most important sentences in fine-tuning time.

---

1To increase the match between user post and summary, we use the lemmatization.

2Bart randomly masks 30% of tokens (default setting).

3We used bart-large as the initial checkpoint.

4Note that the fine-tuning stage of both models is the same, and the only difference is the pre-training stage to validate our hypothesis.

5The pre-training dataset comes from social media domain.
Table 1: Results for Rouge (F1) and Bertscore (F1) metrics comparing our SigBart model with (a) Bart + R3F (RQ1); (b) existing baseline systems (RQ2), on the MentSum dataset. † shows statistical significance (paired t-test, $p < 0.05$).

<table>
<thead>
<tr>
<th>Model</th>
<th>Rg-1 (%)</th>
<th>Rg-2 (%)</th>
<th>Rg-L (%)</th>
<th>BS (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SigBart (ours)</td>
<td>30.79*</td>
<td>9.41*</td>
<td>21.72*</td>
<td>86.69*</td>
</tr>
<tr>
<td>Bart + R3F [2021b]</td>
<td>29.66</td>
<td>8.78</td>
<td>21.25</td>
<td>85.79</td>
</tr>
<tr>
<td>CurrSum† [2022b]</td>
<td>30.16</td>
<td>8.82</td>
<td>21.24</td>
<td>86.32</td>
</tr>
<tr>
<td>MatchSum† [2020]</td>
<td>26.29</td>
<td>6.32</td>
<td>17.12</td>
<td>–</td>
</tr>
<tr>
<td>SocialPegasus</td>
<td>29.71</td>
<td>8.49</td>
<td>20.66</td>
<td>–</td>
</tr>
<tr>
<td>Bart† [2020]</td>
<td>29.13</td>
<td>7.98</td>
<td>20.27</td>
<td>–</td>
</tr>
</tbody>
</table>

Table 2: Human evaluation scores, agreement rates, and relative improvement of SigBart over Bart + R3F, on MentSum.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Bart + R3F</td>
<td>4.43</td>
<td>3.92</td>
<td>4.26</td>
</tr>
<tr>
<td>agr. rate (%)</td>
<td>47</td>
<td>52</td>
<td>53</td>
</tr>
<tr>
<td>SigBart</td>
<td>4.51</td>
<td>4.18</td>
<td>4.41</td>
</tr>
<tr>
<td>agr. rate (%)</td>
<td>51</td>
<td>54</td>
<td>54</td>
</tr>
<tr>
<td>Relative imp. (%)</td>
<td>+1.8</td>
<td>+6.4</td>
<td>+3.4</td>
</tr>
</tbody>
</table>

4 RESULTS AND DISCUSSION

4.1 Automatic evaluation.

Table 1 presents experimental results with respect to automatic evaluation metrics over MentSum dataset. In what follows, we present the results and discussions around each research question.

RQ1: As demonstrated in Table 1 (a), SigBart improves Rouge-1 by 3.8%, Rouge-2 by 6.6%, Rouge-L by 2.2%, and Bertscore by 1.1% in comparison to the Bart + R3F model. As the fine-tuning stage of these two models are the same, with difference being on pre-trained stage, this improvement backs our hypothesis on effectiveness of learning salient representations during the pre-training stage.

RQ2: Table 1 (b) provides the automatic evaluation of other existing baselines that have different pre-training and fine-tuning. The results indicate that SigBart outperforms these existing systems, including the best performing baseline (i.e., CurrSum) by relative improvements of 2.1% (Rouge-1), 6.3% (Rouge-2), 2.3% (Rouge-L), and 0.5% (Bertscore). Moreover, a comparative analysis of Pegasus and SocialPegasus shows only a marginal improvement, indicating that pre-training Pegasus on short documents, akin to those found in social media, may not lead to considerable benefits. SocialPegasus, despite its approach, still lags behind SigBart model. This discrepancy could potentially be attributed to the divergence between the pre-training strategies of SigBart and SocialPegasus, the latter of which adopts a similar but deviating pre-training.

4.2 Human evaluation.

We undertook a human evaluation to contrast SigBart with the Bart + R3F baseline (i.e., the baseline with the same fine-tuning, yet different pre-training), using three qualitative metrics: Fluency (is the summary easy to read and understand?), Informativeness (does the summary provide key information about the user post?), and Faithfulness (is the summary’s information supported by the user post?). The evaluation task was double-blinded, involving 50 randomly picked cases from the MentSum’s test set. To minimize bias, we randomized the order of the summaries for evaluation. For the scoring process, each case was independently reviewed by two assessors, scoring each sample from 1 (worst) to 5 (best) based on these metrics. A third assessor resolved any scoring differences exceeding 2 points.

As presented in Table 2, SigBart model consistently surpasses the baseline in all qualitative metrics. The improvements are specifically prominent in the informativeness and faithfulness metrics, presenting relative improvements exceeding 6% and 3%, respectively. This highlights the effectiveness of the SigBart-specific pre-training method in improving the qualitative metrics. We also attribute gains in fluency to the augmented transferability achieved
### 5 ERROR ANALYSIS

Evaluating underperformed cases, we found that SigBART generates text spans by muddling different source materials, which is a pressing issue of SOTA abstractive summarizers [5, 19], requiring them to deal with complex reasoning over user post. For instance, SigBART unfaithfully produces phrase “...i was born in saudi...” by attending to “born” and “saudi” from two different regions of the user post. Furthermore, SigBART misses producing segments that are not directly mentioned within the user post, but inferred from it. This is a common phenomenon in social media summarization as also denoted by Sotudeh et al. [25], and it remains yet an open question for pre-trained summarization models. For example, we have a phrase “…saved 317 gb worth of data...” in the gold summary, where “317” is inferred from the user post; neither SigBART nor Bart + R3F infers “317” to produce in the summary. Lastly, the current aligner employs exact matching between the user post and the summary, leading to certain limitations. For instance, a phrase like “…saw food being wasted...” in the gold summary does not align with “…seeing perfectly good food thrown away...” in the user post. Similarly, “…meandering around...” in the gold summary does not align with “…start walking away...” in the user post. Despite these phrases conveying similar meanings, they are not detected as alignments by the current system. Future research could address this by employing semantic textual alignment methods, as suggested by Ernst et al. [11]. Examples of generated summaries on two samples are shown in Table 3.

### 6 ETHICS AND PRIVACY

In compliance with similar research practices on health domain [3, 21], Our research strictly adhered to ethical standards by using only anonymized data, ensuring no identification and exploration of users. We maintained stringent data security measures, with access restricted to the research team, to uphold confidentiality and integrity in our mental health study. While our automated summarization system advances state-of-the-art capabilities, it is crucial to acknowledge the potential for misrepresentation or oversimplification of complex mental health narratives within the posts. Given the sensitivity of such topics, special care must be exercised when implementing these technologies on mental health-related datasets at a production scale, ensuring that their application does not compromise the nuanced understanding of mental health issues.