# Curriculum-guided Abstractive Summarization for Mental Health Online Posts

**Sajad Sotudeh** [1*], **Nazli Goharian**[1], **Hanieh Deilamsalehy**[2], and **Franck Dernoncourt**[2]

[1]IRLab, Georgetown University
{sajad,nazli}@ir.cs.georgetown.edu
[2]Adobe Research
{deilamsa,franck.dernoncourt}@adobe.com

## Abstract

Automatically generating short summaries from users' online mental health posts could save counselors' reading time and reduce their fatigue so that they can provide timely responses to those seeking help for improving their mental state. Recent Transformers-based summarization models have presented a promising approach to abstractive summarization. They go beyond sentence selection and extractive strategies to deal with more complicated tasks such as novel word generation and sentence paraphrasing. Nonetheless, these models have a prominent shortcoming; their training strategy is not quite efficient, which restricts the model's performance. In this paper, we include a curriculum learning approach to reweigh the training samples, bringing about an efficient learning procedure. We apply our model on *extreme* summarization dataset of MENTSUM posts —a dataset of mental health related posts from Reddit social media. Compared to the state-of-the-art model, our proposed method makes substantial gains in terms of ROUGE and BERTSCORE evaluation metrics, yielding 3.5% (ROUGE-1), 10.4% (ROUGE-2), and 4.7% (ROUGE-L), 1.5% (BERTSCORE) relative improvements.

## 1 Introduction

Summarization of mental health online posts is an emerging task that aims to summarize users' posts who are seeking help to enhance their mental state in online networks such as Reddit [1] and Reachout [2]. The post might address several issues of the user's concerns or simply be an elaboration on the user's mental and emotional situation. With user preference, each user-written post can be accompanied by a succinct summary (known as TL;DR [3]), condensing major points of the user post. This TL;DR summary is deemed to urge the counselors for a faster read of the user's posted content before reading the post in its entirety; hence, counsellors can provide responses promptly. Herein, we aim to improve state-of-the-art results reported in (Sotudeh, Goharian, and Young, 2022) for this task.

Large-scale deep neural models are often hard to train, leaning on intricate heuristic set-ups, which can be time-consuming and expensive to tune (Gong et al., 2019; Chen et al., 2021). This is especially the case for the Transformers-based summarizers, which have been shown to consistently outperform the RNN networks when rigorously tuned (Popel and Bojar, 2018), but also require heuristics such as specialized learning rates and large-batch training (Platanios et al., 2019). In this paper, we attempt to overcome the mentioned problem on BART (Lewis et al., 2020) Transformers-based summarizer by introducing a *Curriculum Learning (CL)* strategy (Bengio et al., 2009) for training the summarization model, leading to improved convergence time, and performance.

Inspired by humans' teaching style, *curriculum learning* suggests moving the teaching process from easier samples to more difficult ones and dates back to the nineties (Elman, 1993). The driving idea behind this approach is that networks can accomplish better task learning when the training instances are exposed to the network in a specific and certain order, from easier samples to more difficult ones (Chang et al., 2021). In the context of neural networks, this process can be thought of as a technique that makes the network robust to getting stuck at local optima, which is more likely in the early stages of the training process. Given the mentioned challenge of summarization networks, we utilize the SUPERLOSS (Castells et al., 2020) function that falls into the family of confidence-aware curriculum learning techniques, introducing a new parameter called confidence (i.e., $\sigma$) to the network.

* Work partially done during the internship at Adobe Research.

[1] https://www.reddit.com/

[2] https://au.reachout.com/

[3] TL;DR is the abbreviation of "Too Long, Didn't Read". We use "TL;DR" and "summary" exchangeably in this paper.



*Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 148 - 153

We validate our model on MENTSUM (Sotudeh, Goharian, and Young, 2022) dataset, containing over 24k instances mined from 43 mental health related communities on Reddit social media. Our experimental results show the efficacy of applying curriculum learning objectives on BART summarizer, achieving a new state-of-the-art performance.

## 2 Related Work

While majority of works in mental health research have focused on studying users' behavioral patterns through classification and prediction tasks (Choudhury et al., 2013; Resnik et al., 2013; Coppersmith et al., 2014; Yates et al., 2017; Cohan et al., 2017, 2018; MacAvaney et al., 2018), summarization of online mental health posts has been recently made viable. Sotudeh, Goharian, and Young (2022) were the first to step on this direction via introducing MENTSUM dataset of online mental health posts, pinpointing the baseline results. Curriculum Learning (CL) has gained growing interest from the research communities during the last decade (Tay et al., 2019; MacAvaney et al., 2020; Xu et al., 2020). Bengio et al. (2015) were the first to apply this strategy in the context of sequence prediction through *scheduled sampling* approach, which gently changes the training process from ground truth tokens to model-generated ones during decoding. Sample's *difficulty* is a key concept in this scheme as it is used to distinguish easy examples from difficult ones. Researchers have used many textual features as the "difficulty measure" including n-gram frequency (Haffari et al., 2009), word rarity and sentence length (Platanios et al., 2019). Recent works (Saxena et al., 2019; Cachola et al., 2020) have made use of confidence-aware approaches that learn the difficulty of training samples and dynamically reweight samples in the training process.

## 3 Our Approach

In this section, we describe the details of our proposed model, where a curriculum learning architecture is added to the BART's Transformers-based framework, upweighting easier training samples; hence, increasing their contribution in learning stage.

**Curricular Learner for BART.** Considering the applicability of curriculum learning in training large-scale networks, we aim to use it in our summarization framework. Before incorporating the curriculum learning strategy into our model's training stage, we first need to define the *difficulty* metric to distinguish the hardness of samples. In practice, estimating a prior difficulty for each sample is considered a complex task, so we propose to discriminate the samples with progressive signals —such as the respective sample loss at each training iteration— in the training process. In this context, CL is achieved by predicting the difficulty of each sample at the training iterations in the form of a weight, such that difficult samples receive lower weights during the early stages of training and vice versa. To model the curriculum, we propose to use SUPERLOSS (Castells et al., 2020) which is a generic loss criterion built upon the task loss function as shown in Figure 1.

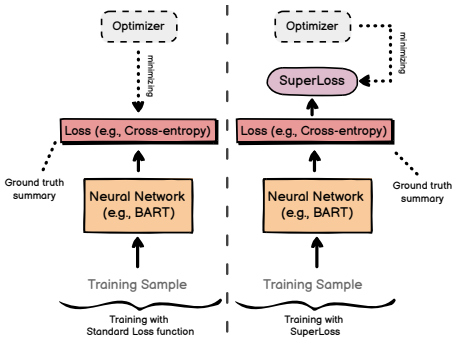


Figure 1: Training with standard loss function (left-hand side) and SuperLoss criteria (right-hand side)

.

More specifically, SUPERLOSS is a task-agnostic confidence-aware loss function that takes in two parameters: (1) the task loss $\mathcal{L}_i = \ell(y_i, \widehat{y_i})$, where $y_i$ is neural network's (i.e., BART's generated summary) output and $\widehat{y_i}$ is the gold label (i.e., ground-truth summary); and (2) $\sigma_i$ as the confidence parameter of the $i$th sample. SUPERLOSS is framed as $\mathrm{L}_\lambda(\mathcal{L}_i, \sigma_i)$ and computed as follows,

$$\mathrm{L}_\lambda(\mathcal{L}_i, \sigma_i) = (\mathcal{L}_i - \tau)\sigma_i + \lambda(\log \sigma_i)^2 \quad (1)$$

in which $\lambda$ is the regularization parameter, and $\tau$ is the running or static average of task loss (i.e., $\mathcal{L}$) during the training. While SUPERLOSS provides a well-defined approach to curriculum learning strategy, learning $\sigma$ parameter is not tractable for tasks with abundant training instances such as text summarization. To circumvent this issue and hinder imposing new learnable parameters, SUPERLOSS suggests using the converged value of $\sigma_i$ at the limit,

$$\sigma^*_\lambda(\ell_i) = \arg\min_{\sigma_i} \mathrm{L}_\lambda(\ell_i, \sigma_i)$$
$$\mathrm{SL}_\lambda(\ell_i) = \mathrm{L}_\lambda(\ell_i, \sigma^*_\lambda(\ell_i, \sigma_i)) = \min_{\sigma_i} \mathrm{L}_\lambda(\ell_i, \sigma_i), \quad (2)$$

Using this technique, the confidence parameters are not required to be learned during the training. Castells et al. (2020) found out that $\sigma^*_\lambda(\ell_i)$ has a closed-form solution, computed as follows,

$$\sigma^*_\lambda(\ell_i) = e^{-W(\frac{1}{2}\max(-\frac{2}{e}, \beta))}, \beta = \frac{\ell_i - \tau}{\lambda} \quad (3)$$

in which $W$ is the Lambert W function. With this in mind, SUPERLOSS upweights easier samples dynamically during the training, providing a curriculum learning approach to summarization. We call this model CURRSUM in our experiments.

## 4 Experimental Setup

### 4.1 Dataset

We use the MENTSUM dataset in our experiments. This dataset contains over 24k post-TL;DR pairs, making up 21,695 (train), 1209 (validation), and 1215 (test) instances, and is gathered by mining 43 mental health subreddit communities on Reddit with rigorous filtering rules. We refer the readers to the main paper for more details on this dataset (Sotudeh, Goharian, and Young, 2022).

### 4.2 Comparison

We compare our model against the BART (Lewis et al., 2020) baseline, which does not utilize the curriculum learning objective. BART is an abstractive model that uses a pre-trained encoder-decoder architecture, unlike BERT which only utilizes a pre-trained encoder. As shown in (Sotudeh, Goharian, and Young, 2022), BART is the strongest baseline; thus, we apply CL on it to evaluate its impact on summarization. We refer the reader to the original paper for more extractive and abstractive baselines.

### 4.3 Implementation details

We use the Huggingface's Transformers library (Wolf et al., 2020) [4] to implement our models. We train all of our models for 8 epochs, performing evaluation step in intervals of 0.5 epochs, and use the checkpoint that achieves the best ROUGE-L score in the validation for the inference. AdamW optimizer (Loshchilov and Hutter, 2019) initialized with learning rate of $3e-5$, $(\beta_1, \beta_2) = (0.9, 0.98)$, and weight decay of 0.01 is used for all of our summarization models, as well as for BART. Cross-entropy loss is used for all models. To keep track of the learning process, we use Weights & Biases (Biewald, 2020) toolkit [5].

| Model | R-1 | R-2 | R-L | BS |
|---|---|---|---|---|
| ORACLEEXT | 35.98 | 11.59 | 23.21 | 82.72 |
| BART (2020) | 29.13 | 7.98 | 20.27 | 85.01 |
| CURRSUM (Ours) | **30.16** | **8.82** | **21.24** | **86.32** |

Table 1: ROUGE and BERTSCORE results on test set of MENTSUM dataset. As BART was the most performant baseline provided in (Sotudeh, Goharian, and Young, 2022), we evaluate the effectiveness of Curriculum on BART in this work. For other baselines, we refer the reader to the main paper.

## 5 Ethics and privacy

While we use publicly available Reddit data to train and validate our model, we recognize that special care should be taken when dealing with this type of data due to its sensitivity and users' privacy (Benton et al., 2017; Nicholas et al., 2020). Thus, we have made no attempts to identify, contact the users, or explore user relations to trace their social media account. It has to be mentioned that the MENTSUM dataset is distributed through Data Usage Agreement (DUA) [6] document to further ensure that users' identity would never be disclosed.

## 6 Results

**Automatic evaluation.** Table 1 reports the performance of the baseline model along with our model's in terms of ROUGE score variants (Lin, 2004) and BERTSCORE (Zhang et al., 2020) over MENTSUM dataset. As indicated, the best model is our CURRSUM that uses SUPERLOSS curriculum directly on top of the BART model and is a clear improvement over it across all metrics, achieving the current state-of-the-art performance. Specifically, our model outperforms its ground baseline that has no curriculum (i.e., BART) by improvement gains of 3.5%, 10.4%, 4.7%, 1.5% for ROUGE-1, ROUGE-2, ROUGE-L, BERTSCORE, respectively. Having looked at the ORACLEEXT scores which shows

[4] https://github.com/huggingface/transformers

[5] https://wandb.ai/

[6] https://ir.cs.georgetown.edu/resources/mentsum.html

| Human-written | BART-generated | CURRSUM-generated |
|---|---|---|
| *... having kids ... my ADHD ... opinions on parenthood...* | *I think it's better to not have a child now than never having a child in the future.* | *I'm **thinking of having a child** and I **don't want to be a shi**y dad** for my **ADHD**. Any **parents with ADHD experience?*** |
| *... symptoms of depression ... go to get help ...* | *I think I **am clinically depressed**, and I have pretty much **all the symptoms of depression**, so I am asking for help* | *I think I am **clinically depressed** xxx. I have pretty much **every symptom on this list**.* |
| *... free of anxiety for a decade.. symptoms come ...* | *I've been **free of anxiety** for less than a week and I'm afraid that it's only a temporary solution. I don't know what to do.* | *I've been **free of anxiety for a decade** and I'm afraid that if the **symptoms start to come back**, I'll be able to **slip back into it** again.* |
| *... ask her out ... barely talked to her... mess up my chance...* | *I like a girl at work and I want to **ask her out**, but I don't want to **ruin that chance** in the process.* | *I like a girl at work, but I don't know how to **ask her out.*** |

Table 2: Four samples of the the human-written, BART-generated, and CURRSUM-generated TL;DR summaries. The human-written samples are partially shown to preserve users' privacy. That is, we have only shown the important human-written phrases to trace them within the generated summaries. The text that is unfaithful to the post (i.e., not supported by the user post) is in Gray and the salient information that is picked up by the summarization systems is shown in **Bold**.

the upper bound performance of an ideal extractive summarizer, it seems that there is room for improvement on the extractive setting to achieve state-of-the-art performance. More sophisticated models can invest in extractive or hybrid summarization models such as those done in (Gehrmann et al., 2018; MacAvaney et al., 2019; Sotudeh et al., 2020).

**Case study and analysis.** While our proposed model significantly improves upon the BART baseline, we recognize the limitations of ROUGE metric in evaluation of summarization systems (Cohan and Goharian, 2016). In order to explore the qualities and limitations of our work, we analyze the human-written TL;DRs along with the generated results by BART and our model, comparing them against each other. Table 2 shows four samples of the human and systems generated TL;DRs. As seen, our model can improve the faithfulness of the summary [7] in the first, second, and third samples. Having looked at other cases in our study, it appears that curriculum learning positively mitigates faithfulness errors. This might be attributed to the fact that the summarizer can achieve an improved *understanding* of the source document when the contribution of each sample is controlled in each iteration of the learning process. Looking at the second sample, it turns out that our model can improve the conciseness of the summaries; that is, briefly conveying the main points within the summary. Comparing system-generated summaries in the fourth sample, it is observed that our system generated a phrase (shown in Gray) by mixing up different regions of the user post. Surprisingly, it appears that "*I don't know how/what to*" is a common phrase used in most human-written summaries that are seeking advice from the community. The experimented summarization systems (i.e., BART and ours) adhere to overgenerating this phrase.

[7] Faithfulness is defined as generating output text that is supported by the user post.

## 7 Conclusion

Generating short summaries given the users' online posts can save counselors' reading time, and reduce their fatigue. On this basis, they can provide faster responses to community users. While neural Transformers-based summarization models have shown to be promising, they suffer from *inefficient training process* that hinders their potentials for showing a promising performance. To compensate for this issue, in this paper, we incorporated a confidence-aware curriculum learning approach, which uses task-agnostic SUPERLOSS, to the summarization framework in the hope of increasing model's generalization, and ultimately improving model performance. Our automatic evaluations over MENTSUM dataset of mental health posts show the effectiveness of our model, yielding 3.5%, 10.4%, 4.7%, 1.5% relative improvements over BART summarizer on ROUGE-1, ROUGE-2, ROUGE-L, and BERTSCORE, respectively. Our model tailors the new state-of-the-art performance on MENTSUM dataset. We further showed various system-generated summaries to showcase the qualities and limitations of our proposed model.

## References


Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1171–1179.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, volume 382 of *ACM International Conference Proceeding Series*, pages 41–48. ACM.

Adrian Benton, Glen A. Coppersmith, and Mark Dredze. 2017. Ethical research protocols for social media health research. In *EthNLP@EACL*.

Lukas Biewald. 2020. Experiment tracking with weights and biases. Software available from wandb.com.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel Weld. 2020. TLDR: Extreme summarization of scientific documents. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4766–4777, Online. Association for Computational Linguistics.

Thibault Castells, Philippe Weinzaepfel, and Jérôme Revaud. 2020. Superloss: A generic loss for robust curriculum learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Ernie Chang, Hui-Syuan Yeh, and Vera Demberg. 2021. Does the order of training samples matter? improving neural data-to-text generation with curriculum learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 727–733, Online. Association for Computational Linguistics.

Xiaohan Chen, Yu Cheng, Shuohang Wang, Zhe Gan, Zhangyang Wang, and Jingjing Liu. 2021. Early-BERT: Efficient BERT training via early-bird lottery tickets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2195–2207, Online. Association for Computational Linguistics.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *ICWSM*.

Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. 2018. Smhd: a large-scale resource for exploring online language usage for multiple mental health conditions. In *COLING*.

Arman Cohan and Nazli Goharian. 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).

Arman Cohan, Sydney Young, Andrew Yates, and Nazli Goharian. 2017. Triaging content severity in online mental health forums. *Journal of the Association for Information Science and Technology*, 68.

Glen A. Coppersmith, Craig Harman, and Mark Dredze. 2014. Measuring post traumatic stress disorder in twitter. In *ICWSM*.

J. Elman. 1993. Learning and development in neural networks: the importance of starting small. *Cognition*, 48:71–99.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 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.

Linyuan Gong, Di He, Zhuohan Li, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2019. Efficient training of BERT by progressively stacking. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2337–2346. PMLR.

Gholamreza Haffari, Maxim Roy, and Anoop Sarkar. 2009. Active learning for statistical phrase-based machine translation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 415–423, Boulder, Colorado. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 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.

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

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

Sean MacAvaney, Bart Desmet, Arman Cohan, Luca Soldaini, Andrew Yates, Ayah Zirikly, and Nazli Goharian. 2018. Rsdd-time: Temporal annotation of self-reported mental health diagnoses. In *CLPsych@NAACL-HTL*.

Sean MacAvaney, Franco Maria Nardini, Raffaele Perego, Nicola Tonellotto, Nazli Goharian, and Ophir Frieder. 2020. Training curricula for open domain answer re-ranking. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 529–538. ACM.

Sean MacAvaney, Sajad Sotudeh, Arman Cohan, Nazli Goharian, Ish A. Talati, and Ross W. Filice. 2019. Ontology-aware clinical abstractive summarization. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 1013–1016. ACM.

Jennifer Nicholas, Sandersan Onie, and Mark Erik Larsen. 2020. Ethics and privacy in social media research for mental health. *Current Psychiatry Reports*, 22:1–7.

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom Mitchell. 2019. Competence-based curriculum learning for neural machine translation. 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 1162–1172, Minneapolis, Minnesota. Association for Computational Linguistics.

Martin Popel and Ondrej Bojar. 2018. Training tips for the transformer model. *The Prague Bulletin of Mathematical Linguistics*, 110:43 – 70.

Philip Resnik, Anderson Garron, and Rebecca Resnik. 2013. Using topic modeling to improve prediction of neuroticism and depression in college students. In *EMNLP*.

Shreyas Saxena, Oncel Tuzel, and Dennis DeCoste. 2019. Data parameters: A new family of parameters for learning a differentiable curriculum. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 11093–11103.

Sajad Sotudeh, Nazli Goharian, and Ross Filice. 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.

Sajad Sotudeh, Nazli Goharian, and Zachary Young. 2022. Mentsum: A resource for exploring summarization of mental health online posts. In *Proceedings of the Language Resources and Evaluation Conference*, pages 2682–2692, Marseille, France. European Language Resources Association.

Yi Tay, Shuohang Wang, Anh Tuan Luu, Jie Fu, Minh C. Phan, Xingdi Yuan, Jinfeng Rao, Siu Cheung Hui, and Aston Zhang. 2019. Simple and effective curriculum pointer-generator networks for reading comprehension over long narratives. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4922–4931, Florence, Italy. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 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.

Benfeng Xu, Licheng Zhang, Zhendong Mao, Quan Wang, Hongtao Xie, and Yongdong Zhang. 2020. Curriculum learning for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6095–6104, Online. Association for Computational Linguistics.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. In *EMNLP*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.