SAD: A Stress Annotated Dataset for Recognizing Everyday Stressors in SMS-like Conversational Systems

Matthew Louis Mauriello
mlm@udel.edu
University of Delaware
Newark, Delaware, USA

Emmanuel Thierry Lincoln
lincolnt@stanford.edu
Stanford School of Medicine
Stanford, California, USA

Grace Hon
gracehon@stanford.edu
Stanford School of Medicine
Stanford, California, USA

Dorien Simon
dksimon@stanford.edu
Stanford University
Stanford, California, USA

Dan Jurafsky
jurafsky@stanford.edu
Stanford University
Stanford, California, USA

Pablo E. Paredes
pparedes@stanford.edu
Stanford School of Medicine
Stanford, California, USA

ABSTRACT

There is limited infrastructure for providing stress management services to those in need. To address this problem, chatbots are viewed as a scalable solution. However, one limiting factor is having clear definitions and examples of daily stress on which to build models and methods for routing appropriate advice during conversations. We developed a dataset of 6850 SMS-like sentences that can be used to classify input using a scheme of 9 stressor categories derived from: stress management literature, live conversations from a prototype chatbot system, crowdsourcing, and targeted web scraping from an online repository. In addition to releasing this dataset, we show results that are promising for classification purposes. Our contributions include: (i) a categorization of daily stressors, (ii) a dataset of SMS-like sentences, (iii) an analysis of this dataset that demonstrates its potential efficacy, and (iv) a demonstration of its utility for implementation via a simulation of model response times.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**; User studies.

KEYWORDS

Conversational Agents; Stress Management; Daily Stress; Stressors; Datasets; Classification

ACM Reference Format:
Matthew Louis Mauriello, Emmanuel Thierry Lincoln, Grace Hon, Dorien Simon, Dan Jurafsky, and Pablo E. Paredes. 2021. SAD: A Stress Annotated Dataset for Recognizing Everyday Stressors in SMS-like Conversational Systems. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI '21 Extended Abstracts), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3411763.3451799

CHI '21 Extended Abstracts, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8095-9/21/05…$15.00
https://doi.org/10.1145/3411763.3451799

1 INTRODUCTION

There is currently limited infrastructure and available personnel for providing mental health services to those around the world who need it. To address this problem, chatbots—software systems that in lieu of humans provide support to users using conversations via text or text-to-speech—are increasingly being viewed as a scalable solution given the relative ubiquity of access to mobile and internet-based services. Early examples of chatbots for mental health include ELIZA [29] which delivered non-directive therapy mirroring Rogerian therapy by reflecting and rephrasing user input. More recent examples include Woebot [9], Wysa [14], and Tess [11] which focus on delivering Cognitive Behavior Therapy (CBT) to users to help reduce anxiety and depression symptoms. Chatbots exist on a simple continuum of conversational fluency. At one extreme are those that respond to any user input allowing for open-ended conversations. This is convenient for users as these systems mirror how they typically communicate online. However, as chatbots are still in early stages of development they can be clumsy at times, fail to recognize certain requests, and may not respond appropriately [18].

One common approach to building open-ended dialog systems is to train conversational models using large datasets of unlabeled data or input and output pairs generated from large-scale internet content or crowdsourcing methods. The third version of the Generative Pre-Trained Transformer (GPT-3) recently made a splash in the world of Natural Language Processing (NLP) and Machine Learning (ML) as it is able to take unlabeled internet content as data and use it to create a model that can generate human-like text based on an input prompt [3]. Although, its ability to create new text artifacts has been criticized for failing certain semantic and ethical tests [10]. For example, there has been reports of medical chatbots that use GPT-3 telling fictitious patients to commit suicide (*e.g.*, [5]). This is not a new issue, commercial assistants such as Alexa and others have been known to suggest murder and give other advice that presents safety risks [2, 4]. Moreover, open dialogue chatbot models have been found to amplify gender bias that exist in training dialogues [7, 16]. As a result, open-ended dialog remains unpredictable and problematic for mental health applications.

At the other extreme are chatbots that adhere to tightly scripted conversations. These yield predictable user interactions but are limited in their conversational scope resulting in low user adoption and high attrition. As a result, many of today's chatbots fall somewhere in the middle, incorporating both scripted and open-ended

conversations. Combinations of scripted conversation with classification models and keyword matching can direct conversational flows along reasonably predictable routes and allows designers to provide strategic diversification based on user utterances. However, there are few public datasets related to mental health issues available today that make some interactions hard to generate.

In our work, we focus on daily stress and stress management—an area of mental health which impacts many and is often a prodromal symptom of other mental health conditions. Prior work has developed several chatbots for proactive stress management including Tess [5] and the Popbots [21] which users tend to find fairly scripted after prolonged use. Our aim is to enable more dynamic chatbot designs able to deliver appropriate interventions by taking advantage of NLP and ML techniques to direct conversations in ways that are predictable, generate empathy by acknowledging stressful events, and increase user adherence to their use. As a initial step, we developed a simple categorization scheme for daily stressors and acquired training data that can be used to classify user input. Given the sensitive nature and ethical concerns related to mental health, we argue that such training data should be made public and evaluated for use in mental health applications rather than siloed away from such review. Toward that goal, we present the Stress Annotated Dataset (SAD)—a dataset of 6850 high-quality SMS-like sentences (*e.g.,* "my lease is ending soon") that can be used to categorize daily stressors into multiple topics. Our classification scheme contains 9 categories and was derived from stress management literature, live conversations from a prototype chatbot system, and further extended and iteratively refined by US-based crowd workers and targeted web scraping from a repository of emotionally charged LiveJournal data [20]. Motivated to contribute a useful dataset for mental health research, we aim to address the following research questions: *How do users of a mobile chatbot system describe daily stress? Might this data generate a dataset that results in models that can classify stressors into multiple topics from short SMS-like message snippets? And, might these models be viable in mental health applications?*

To answer our research questions, we iteratively generated and refined the first version of SAD. To assess classification performances, we performed a N=20 bootstrap experiment and achieved a mean overall F1-score of 0.809 (SD=0.010, SE=0.002) using a pretrained BERT model that we fine-tuned. We then deployed our resulting model as a web-based API and simulated an experimental user load demonstrating that end-to-end communication times would meet user expectations for conversations in future applications. As a result, the contributions of our work include: (i) a simple categorization scheme for daily stressors derived from live and simulated conversation, (ii) a dataset of SMS-like sentences describing these stressors, (iii) an analysis of this dataset that demonstrates its potential efficacy for topic classification, and (iv) a demonstration of the utility of resulting models for implementation in mental health applications through a simulation of response times via a web-based API.

2 RELATED WORK

Here we provide a brief background on stress, describe how prior work informs our classification scheme, and discuss the importance of developing datasets for mental health applications.

2.1 Defining Daily Stress

The stress response is an evolutionary mechanism that mobilizes bodily resources to help humans cope with daily challenges as well as life-threatening situations. Stress has two components, a stressor and a stress response. The former could be linked to sources of uncertainty, complexity, cognitive loads, or emotional distress. The latter is the mental and physical reaction to such stimuli. Daily stressors are defined as the routine challenges of day-to-day living. The challenges can either be predictable (*e.g.* daily commutes) or unpredictable (*e.g.* an unexpected work deadline) and occurs in 40% of all days. Unlike chronic stress, these stressors are relatively short-lived and do not persist from day to day [1, 24]. However, daily stress has been shown to cause psychological distress and exacerbate symptoms of existing physical health conditions [1]. Repeated triggering of daily stress can also lead to chronic stress, which has been associated with a variety of patho-physiological risks—conditions that impair quality of life and shorten life expectancy [8, 17]. Given the increasing use of digital communication, having data that enables effective identification and classification of stressors in text would provide designers of mental health applications with new opportunities for creating positive interactions. However, what constitutes a daily stressor is subjective though there are several inventories. For example, the Holmes and Rahe Stress Scale [19] lists 43 specific stressors that include going on vacation and the death of a spouse. Other inventories are more categorical and include financial and family issues as significant sources. In our work, we propose a topic-based scheme of 9 categories for classification informed by literature, user conversations, and crowd workers.

2.2 Datasets for Mental Health

As noted in the introduction, recent work creating datasets for chatbots focuses on open-domain models. Moreover, there are numerous tools for training customer support chatbots (*e.g.,* Collect.chat) using example conversations, crowdsourced data, and public FAQs. One limitation is that the resulting models produce chatbots that can have reasonably correct conversations but are often not suitable for mental health applications because they do not empathize as an engaged human conversational partner might. To address this issue, Raskin *et al.* [25] recently developed an open-domain conversational model for generating empathetic responses to user input using emotionally charged content as training data. While such datasets and models generate conversations that are perceived as being more empathetic compared to those trained on large-scale internet data alone, open-dialogue systems still may generate unpredictable responses. Another limitation is that mental health issues require specialized and specific data. As a result, there is interest in generating new datasets for such applications. One area of focus has been detecting and predicting risk of suicide using data flagged in online posts. For example, the UMD Reddit Suicidality Dataset [27] was generated as part of CLPsych 2019 Shared Task [31] to detect risk of suicide from Reddit posts and such efforts may result in a successful screening procedure that could be widely used in mental health applications—particularly to escalate conversations from chatbots to human support personnel. Closer to our work are those datasets that focus on topic classification. For example, the Dreaddit dataset [28] harvested 190K posts from ten subreddits expected to

produce content in five stress domains (*e.g.,* r/relationships, r/ptsd). Crowd workers labeled 3K of these posts and classification results demonstrate lexical similarities that enable the origin (subreddit) of a post to be identified and that high accuracy can be achieved when identifying stressful posts. Our work also uses crowdsourcing and web scraping to generate a dataset but with more categories and using input similar to SMS-like conversations.

3 DATASET GENERATION

Here we describe our process for generating our dataset including the use of user data, crowdsourcing, and web scraping.

3.1 Development of the Classification Scheme

To begin generating our dataset of daily stressors, we first compiled a list of 185 unique sentences obtained directly from user conversations in a prototype chatbot system [21]. During each conversation, users were asked to describe a recent event that was stressful for them given the prompt: *"What is stressing you out right now?"* The research team then reviewed these conversations, extracted the responses, and assigned an appropriate stressor category based on simplifying the Holmes and Rahe Stress Scale [19]. As noted earlier, the Holmes and Rahe Scale contains 43 specific items related to major stressors making it difficult to use directly for labeling and topic classification because the items were both too specific and too numerous; however, many of these items can be clustered into a few general topics such as problems related to finance (*e.g.,* taking out a loan, foreclosure on a loan). Three researchers iteratively discussed each item to derive the topic list used to create our initial simplified categorization. This resulted in 10 stressor categories including: family, commute, exhaustion, financial problem, personal life, physical health, social relationships, travel, work, and school.

3.2 Content Generation and Curation

In the next phase, we developed two Human Intelligence Tasks (HITs) for Amazon Mechanical Turk (AMT). Using these HITs, crowd workers on the platform labeled and conducted quality assurance activities as described here.

3.2.1 Stressor Generation. In the first HIT, individual workers were asked to supply new stressors answering a similar question as users in the prototype chatbot system (Figure 1a) but with additional helpful context and input length counters (though input was unlimited). These workers then labeled their stressors using our initial categories (Figure 1b) and assigned a severity using a 10-pt Likert scale (rated *not stressful* to *extremely stressful*). The research team then reviewed the submitted data and observed that: (i) workers interpreted stressor categories differently than expected, (ii) certain categories were rarely used, and (iii) submitted stressors often contained two stressors. The research team used these observations to collapse and modify categories as well as definitions toward improving consistency resulting in Table 1. For example, "exhaustion" was rarely used in favor of "physical health" thus the merged category became "Health, Fatigue, or Physical Pain" to better reflect the way stressors were being categorized by crowd workers.

3.2.2 ality Assurance. In the second HIT, five workers were recruited to review individual sentences. These workers determined whether a sentence was stressful using a binary (*i.e.,* yes/no) question, assigned up to two categories, and rated the severity. We then aggregated this data, calculated the percent agreement workers had on each question, and used a majority vote to determine the final labels. Thus, the data used in the proceeding analysis represents not the intent of the original author but the interpretation of the sentence by human raters. Additionally, some data was generated

Table 1: Final iterated version of our stressor categories, and definitions with representative examples.

Category	Abbr.	Definition	Examples
Work	W	Stress resulting from the person's job or commute to/from their place of employment.	*"I just started working a new job and I'm worried about messing up."*
School	S	Stress resulting from the person's schoolwork or experience in school.	*"I need to get my homework done but I don't have a lot of time."*
Financial Problem	FP	Stress resulting from money related issues.	*"My lease is ending soon."*
Emotional Turmoil	ET	Stress resulting from the person's inner perceptions, emotional distress, or anxiety.	*"I have been feeling kind of lonely lately."*
Social Relationships	SR	Stress resulting from the person's friends, romantic companions, coworkers, schoolmates or other acquaintances.	*"I found out my ex has a new girlfriend."*
Family Issues	FI	Stress related to anyone who would be generally considered a family member.	*"My baby is learning to climb out of her crib."*
Health, Fatigue, or Physical Pain	H	Stress resulting from a person's physical or mental condition including health problems, injury, or tiredness.	*"I have a terrible headache."*
Everyday Decision Making	ED	Stress resulting from small problems or decisions that people face on a daily basis.	*"Don't know what to cook for dinner."."*
Other	O	Stress resulting from a stressor that does not fit any definitions above.	*"Politics"*

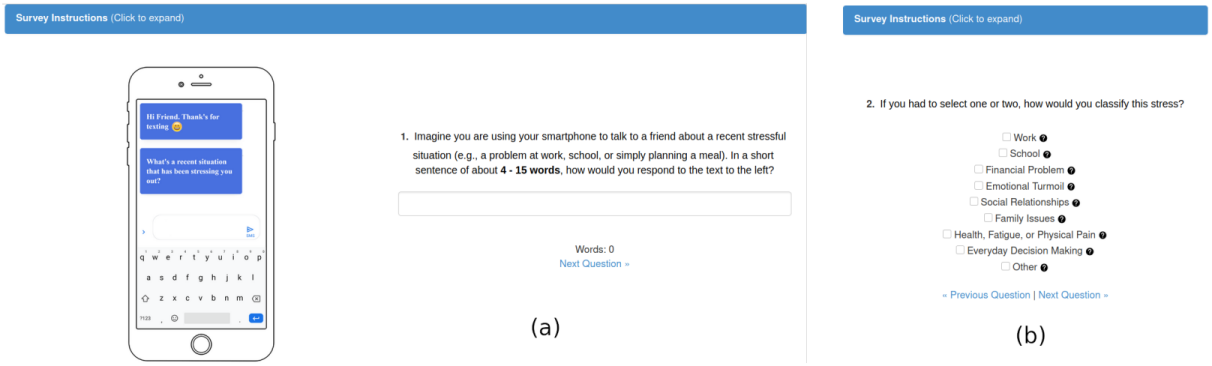

Figure 1: Example of Human Intelligence Tasks: (a) worker is prompted to describe a recent stressful event in a few words—similar to a text message to a friend (screen capture from Generation HIT), (b) in both HITs workers labeled the stressful event (screen capture from final Quality Assurance HIT).

during the start of the COVID-19 pandemic, resulting in perturbation in the distribution of generated stressors as evidenced by a decrease in *"Work"* as well as other categories and a rise in *"Financial Problems"* and *"Health, Fatigue, and Physical Pain"* (Figure 2, left). As a result, workers also rated, on a binary question, whether a sentence was related to this pandemic.

3.2.3 Deploying Tasks. After piloting our HITs, we then recruited crowd workers to label our data. To ensure both quality of work and familiarity with the content/context, we used qualifying criteria that included: (i) being located in the US, (ii) fluency in English, and (iii) a HIT Approval Rate greater than 80. We gathered an additional 3119 unique stressful sentences and then ran these sentences through our quality assurance HIT. In total, 3578 unique workers participated in our tasks—recall that individuals workers submitted stressful sentences while five reviewed them. Procedures were reviewed and approved by our university's Institutional Review Board (IRB). Workers could complete tasks multiple times and were paid $13.00 per hour pro-rated based on each task's estimated completion time (~30 seconds) in accordance with state minimum wage.

3.3 Targeted Web-Scraping

Ultimately, data generation via AMT resulted in several low cardinality categories. To address this issue and further enrich our dataset, we scraped 3546 similar sentences to increase the examples in these low cardinality categories using cosine-similarity and an available repository of emotionally charged LiveJournal data [20]. To achieve this, we first improved the original GloVe [23] sentence index used by the tool by re-indexing the sentence corpus with SBERT embeddings [26]. We then performed the dataset enrichment via two distinct approaches: (i) picking ten random sentences from low cardinality categories in our dataset and searching for an estimated number of similar sentences (with greater than ~73% similarity) in the LiveJournal corpus before (ii) taking five distinct sentences from the non-enriched data, computing the average embedding vector of those sentences, and using the new vector to find the ten most similar sentences inside the LiveJournal corpus. We then passed this new data through our quality assurance HIT.

4 DATASET SCHEMA AND ANALYSIS

In total, the first version of our dataset contains 6850 example sentences across 9 stressful categories (Figure 2, right) that can be used to identify stressful topics in short SMS-like sentences.

4.1 Schema

Each stressor in our dataset is associated with several elements of metadata that will allow researchers to experiment. This metadata includes a stressor ID, the sentence text, and the source of the data (*i.e.,* whether the data came from live users, crowd workers, or web scraping). Each sentence is labeled three ways: with the original author's label (permuted with category updates) as well as the top and second label based on the majority vote of five crowd workers. We also include the full distribution of selected labels which we have found useful in some experiments such as evaluating single-label versus multi-label input during model training. Descriptive statistics (*i.e.,* mean, median, and standard deviation) of the severity rating provided by the crowd workers are also provided. Metadata also includes two binary fields, *isStressor* and *isCovid*, with percent agreement fields for those ratings. Finally, an *isSeed* binary field indicates whether an example sentence was used to seed our web scraping efforts. While the analysis that follows uses the full set of data, we plan to use this metadata to further evaluate, experiment, and maintain quality as we iterate on future versions of the dataset.

4.2 Influence of Quality Assurance & Scraping

Combing through large volume of data to generate datasets appropriate for machine learning applications is quite tedious, however, crowdsourcing and web scraping have been standard techniques for overcoming this obstacle. For example, Hara *et al.* used crowd workers to evaluate accessibility issues in images of sidewalks collected from Google Street View and evaluated different numbers of raters and voting mechanisms [13]. Based on these results, we selected five workers to evaluate our stressor sentences and used majority voting as our aggregation method given the perceived lower complexity of our task. We evaluated the impact of using the original labels versus one, three, and five raters majority vote

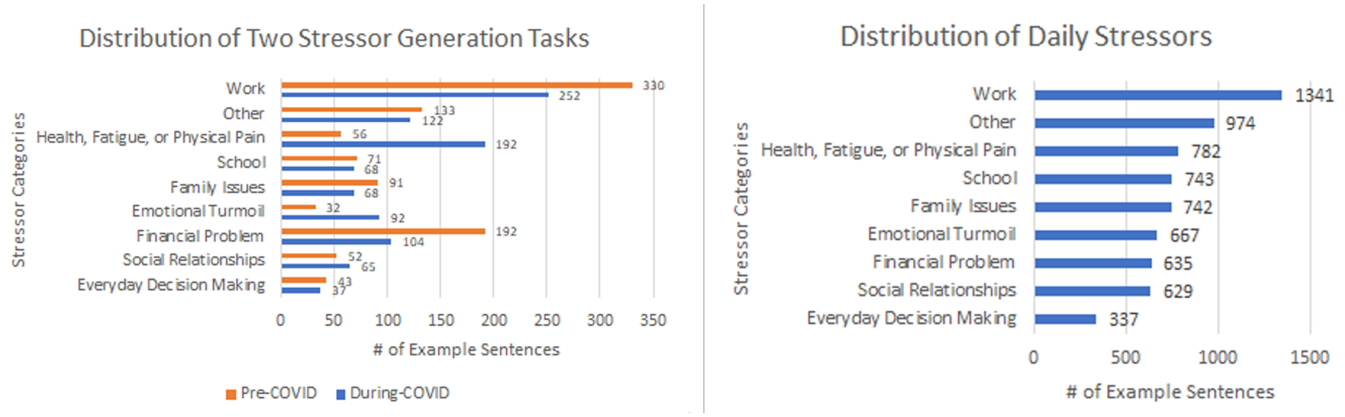

Figure 2: Two example distributions from stressor generation tasks (*i.e.,* 1000 stressors generated per) pre- and during- the COVID-19 pandemic (left). The final distribution of stressors across our 9 iterated categories after crowdsourcing and web scraping (right).

(Figure 3) using the default Support Vector Classification algorithm provided by scikit-learn [22]. As the original labels suffered from clarity issues, we see a considerable improvement in scores as we improve our labeling scheme and increase the number of raters. We also see slight improvement by including data from the LiveJournal corpus in terms of overall averages; however, earlier experiments demonstrated that classification error resulting from class imbalance was decreased, contributing to the results presented in Table 2.

4.3 Overall and Multi-Topic Classification

For this evaluation, our goal was not a deep linguistic analysis of the sentences but rather to understand whether the lexical patterns differed enough for relatively accurate classification. Based on the overview of the topic distribution (Figure 2), we decided to incorporate data that had at least 600 examples into our analysis and merged the *"Everyday Decision Making"* category into *"Other"*. For the classification task we chose to fine-tune a pre-trained *BERT*$_{(base)}$ model. We trained the model for three epochs (similarly to [6]), the batch size was sixteen and all other hyperparameters where left at their default values. To assess classification performances, we performed a N=20 boostrap experiment where at each run we sampled by category 80% of the dataset as training data, leaving 20% as the test set, and fine-tuned the model for each iteration. As a result, we achieved a mean overall F1-score of 0.809 (*SD*=0.010, *SE*=0.002). Detailed F1-score measurements per category are shown in Table 2. On review, these results seem reasonable compared to other datasets on nascent topics (*e.g.*, Dreddit [28], Fake News vs Satire [12]). This high-performing model suggests that our short SMS-like sentences are able to be identified by topic though limitations discussed below.

5 SYSTEM INTEGRATION

To evaluate the utility of our model for mental health applications, we performed a simulation of response times for a web-based API.

5.1 Simulation

Users typically expect that chatbots respond in 2-3 seconds which is problematic for BERT models because they suffer from high inference times [30]. To make the classification process viable in the context of a chatbot system, it was crucial that the deployed model delivers low inference times at scale. To evaluate this inference time, we set up a Flask REST API alongside a TensorFlow Serving Docker instance which performs respectively pre-processing and inferencing. A simulation was performed with N={1, 10, and 50} users simultaneously querying the model deployed on an AWS EC2 c5.large instance (N=5 bootstrap), this resulted in average end-to-end response time of 0.16s (*SD*=0.01, *Min*=0.15, *Max*=0.21), 1.00s (*SD*=0.09,

Table 2: Based on our bootstrap approach, we report the overall average precision, recall, and f1-measure scores as well as their mean, standard deviation, and standard error in the first three columns followed by the f1-measures for the individual categories.

	Precision	Recall	F1	O	W	SR	FP	ET	HF	S	FI
Mean	0.814	0.807	0.809	0.667	0.905	0.779	0.869	0.636	0.837	0.920	0.861
SD	0.012	0.009	0.010	0.036	0.011	0.033	0.018	0.033	0.025	0.011	0.018
SE	0.003	0.002	0.002	0.008	0.002	0.007	0.004	0.007	0.006	0.002	0.004
Support	1043	1043	1043	152	238	99	101	103	101	123	126

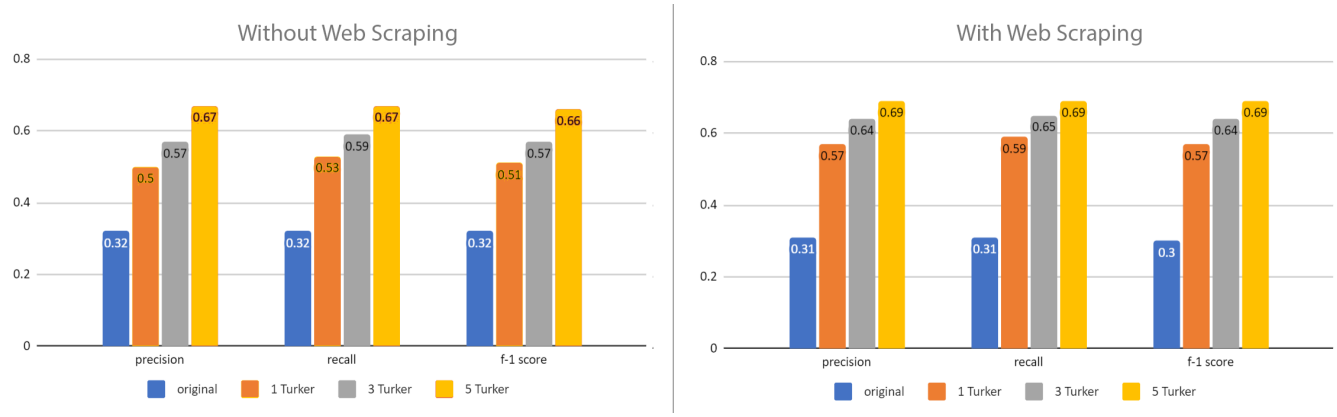

Figure 3: While models generated from the original label of sentences submitted by worker-authors performs quite poorly due to iterating on the labeling scheme, the influence of multiple workers labeling stressful sentences with the final scheme demonstrates relative improvement—particularly moving from one worker to five—when evaluating on the prediction of the top label generated by a simple majority vote (left). Moreover, we observe additional improvement in overall scores after adding the web scraping data to balance the number of examples in each class (right).

Min=0.65, *Max*=1.10), 3.41s (*SD*=1.36, *Min*=1.22, *Max*=5.28) respectively. For 50 simultaneous users, the performance time remains within an acceptable range but may begin to exceed user expectation. However, that load is unlikely for experimental scales (*i.e.,* 0-10K users). Greater scale can be achieved by deploying the model and API infrastructure on a more powerful machine or using multiple machines with load balancing.

6 DISCUSSION

We created a dataset of 6850 SMS-like sentences with high-quality labels across 9 stressor categories with associated metadata that we call SAD—the first version of our daily stressor dataset. This data has been anonmyized and, we believe, it will be helpful for researchers and application designers. Thus, we have made this dataset available at the following address: https://github.com/PervasiveWellbeingTech/Stress-Annotated-Dataset-SAD

Our initial analysis suggests that topic classification of different stressors is possible using our training data and our simulation results suggest that resulting models are viable for deployment in conversational systems. However, there are several important limitations and areas for future work. First, as our data comes from live users and crowd workers it is important to continue to monitor the results of our labeling process and the category distribution to mitigate the impact of, for example, topic drift that may result from major events that impact users lived experience. While our classification was robust to changes in distribution that resulted from the COVID-19 pandemic (*i.e.,* possibly because health was already a category), future events and their impact are hard to predict. Moreover, further clarification of definitions could improve label quality (*e.g.,* category ET and H could be confused). Second, crowdsourcing data to create enough examples for training proved difficult and costly, thus, we turned to web scraping based on cosine-similarity. This may have artificially improved our classification results requiring further evaluation though our manual review of this data suggests that these sentences reasonably approximate what we received through other sourcing mechanisms. Finally, as our criteria for our HITs restricted contributions to US-based crowd worker it is likely that our annotations are biased towards US perceptions of daily stressors. Other cultures could view daily stressors differently.

Next steps for our work is to implement our models in an active chatbot system and automatically send new stressor sentences from users to AMT for labeling and incorporation into the dataset. This will necessitate additional quality assurance methods such as on-and-off testing against a gold standard test set as has been done in other active learning systems (*e.g.,* [15]). Finally, since the inception of this work, new and improved approaches for generating data and deploying models have become readily available. For example, cosine-similarity search using embedding indexing can now be achieved with Elasticsearch (Elastic.co) instead of relying on [20] and BERT model training as well as implementation can be done using Rasa NLU (Rasa.com). We aim to explore these advances and improve our SAD dataset for future use in research and mental health applications.

7 CONCLUSIONS

In this paper, we built a dataset of 6850 example SMS-like sentences representing 9 categories of daily stressors. In addition to releasing this dataset, we analyzed it and show results that are promising for stressful topic classification. Moreover, we demonstrate the utility of our dataset by training a model and exploring its practical implementation. As a result, our contributions include: (i) a simple categorization scheme for daily stress, (ii) a dataset of SMS-like sentences describing these stressors, (iii) an analysis of this dataset that demonstrates its potential efficacy for topic classification, and (iv) a demonstration of the utility of resulting models for implementation in mental health applications through a simulation of response times via a web-based API.

ACKNOWLEDGMENTS

Work supported by the Stanford Institute for Human-Centered Artificial Intelligence. Contributions by M.L.M. made while in transition from Stanford School of Medicine to University of Delaware.

REFERENCES

[1] David M Almeida. 2005. Resilience and Vulnerability to Daily Stressors Assessed via Diary Methods. *Current Directions in Psychological Science* 14, 2 (2005), 64–68. https://doi.org/10.1111/j.0963-7214.2005.00336.x

[2] Timothy W Bickmore, Ha Trinh, Stefan Olafsson, Teresa K O'Leary, Reza Asadi, Nathaniel M Rickles, and Ricardo Cruz. 2018. Patient and consumer safety risks when using conversational assistants for medical information: an observational study of Siri, Alexa, and Google Assistant. *Journal of medical Internet research* 20, 9 (2018), e11510.

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL]

[4] Jeffrey Dastin. 2018. 'Kill your foster parents': Amazon's Alexa talks murder, sex in AI experiment. *Reuters* (Dec 2018). https://www.reuters.com/article/us-amazon-com-alexa-insight/kill-your-foster-parents-amazons-alexa-talks-murder-sex-in-ai-experiment-idUSKCN1OK1AJ

[5] Ryan Daws. 2020. Medical chatbot using OpenAI's GPT-3 told a fake patient to kill themselves. *AI News* (Oct 2020). https://artificialintelligence-news.com/2020/10/28/medical-chatbot-openai-gpt3-patient-kill-themselves/

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[7] Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. Queens are powerful too: Mitigating gender bias in dialogue generation.

[8] Elissa S Epel, Elizabeth H Blackburn, Jue Lin, Firdaus S Dhabhar, Nancy E Adler, Jason D Morrow, and Richard M Cawthon. 2004. Accelerated telomere shortening in response to life stress. *Proceedings of the National Academy of Sciences of the United States of America* 101, 49 (2004), 17312–17315. https://doi.org/10.1073/pnas.0407162101

[9] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR Mental Health* 4, 2 (2017). https://doi.org/10.2196/mental.7785

[10] Luciano Floridi and Massimo Chiriatti. 2020. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines* (2020), 1–14.

[11] Russell Fulmer, Angela Joerin, Breanna Gentile, Lysanne Lakerink, and Michiel Rauws. 2018. Using Psychological Artificial Intelligence (Tess) to Relieve Symptoms of Depression and Anxiety: Randomized Controlled Trial. *JMIR Ment Health* 5, 4 (dec 2018), e64. https://doi.org/10.2196/mental.9782

[12] Jennifer Golbeck, Matthew Mauriello, Brooke Auxier, Keval H. Bhanushali, Christopher Bonk, Mohamed Amine Bouzaghrane, Cody Buntain, Riya Chanduka, Paul Cheakalos, Jennine B. Everett, Waleed Falak, Carl Gieringer, Jack Graney, Kelly M. Hoffman, Lindsay Huth, Zhenya Ma, Mayanka Jha, Misbah Khan, Varsha Kori, Elo Lewis, George Mirano, William T. Mohn IV, Sean Mussenden, Tammie M. Nelson, Sean Mcwillie, Akshat Pant, Priya Shetye, Rusha Shrestha, Alexandra Steinheimer, Aditya Subramanian, and Gina Visnansky. 2018. Fake News vs Satire: A Dataset and Analysis. In *Proceedings of the 10th ACM Conference on Web Science* (Amsterdam, Netherlands) *(WebSci '18)*. Association for Computing Machinery, New York, NY, USA, 17–21. https://doi.org/10.1145/3201064.3201100

[13] Kotaro Hara, Vicki Le, and Jon Froehlich. 2013. Combining Crowdsourcing and Google Street View to Identify Street-Level Accessibility Problems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. Association for Computing Machinery, New York, NY, USA, 631–640. https://doi.org/10.1145/2470654.2470744

[14] Becky Inkster, Shubhankar Sarda, and Vinod Subramanian. 2018. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. *JMIR mHealth and uHealth* 6, 11 (2018), e12106.

[15] Gierad Laput, Walter S. Lasecki, Jason Wiese, Robert Xiao, Jeffrey P. Bigham, and Chris Harrison. 2015. Zensors: Adaptive, Rapidly Deployable, Human-Intelligent Sensor Feeds. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 1935–1944. https://doi.org/10.1145/2702123.2702416

[16] Haochen Liu, Jamell Dacon, Wenqi Fan, Hui Liu, Zitao Liu, and Jiliang Tang. 2020. Does Gender Matter? Towards Fairness in Dialogue Systems. In *Proceedings of the 28th International Conference on Computational Linguistics*. 4403–4416.

[17] B S McEwen. 1998. Stress, adaptation, and disease. Allostasis and allostatic load. *Annals of the New York Academy of Sciences* 840 (1998), 33–44. https://doi.org/10.1111/j.1749-6632.1998.tb09546.x

[18] Adam S Miner, Arnold Milstein, Stephen Schueller, Roshini Hegde, Christina Mangurian, and Eleni Linos. 2016. Smartphone-Based Conversational Agents and Responses to Questions About Mental Health, Interpersonal Violence, and Physical Health. *JAMA Internal Medicine* 176, 5 (2016), 619. https://doi.org/10.1001/jamainternmed.2016.0400 arXiv:15334406

[19] Peter A Noone. 2017. The Holmes–Rahe Stress Inventory. *Occupational Medicine* 67, 7 (2017), 581–582.

[20] Pablo Paredes, Ana Rufino Ferreira, Cory Schillaci, Gene Yoo, Pierre Karashchuk, Dennis Xing, Coye Cheshire, and John Canny. 2017. Inquire: Large-Scale Early Insight Discovery for Qualitative Research. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 1562–1575. https://doi.org/10.1145/2998181.2998363

[21] Pablo E Paredes, Nantanick Tantivasadakarn, Grace Hon, Emmanuel Thierry Lincoln, Nikhil Gowda, Marco A Mora-Mendoza, and Matthew Louis Mauriello. 2020. Towards PopBots: A Suite of Conversational Agents for Daily Stress. *CH 2020 Workshop on Conversational Agents for Health and Wellbeing* (2020). https://www.researchgate.net/publication/344771066_Towards_PopBots_A_Suite_of_Conversational_Agents_for_Daily_Stress/

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[23] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1532–1543. https://doi.org/10.3115/v1/D14-1162

[24] Jennifer R Piazza, Susan T Charles, Martin J Sliwinski, Jacqueline Mogle, and David M Almeida. 2013. Affective reactivity to daily stressors and long-term risk of reporting a chronic physical health condition. *Annals of behavioral medicine : a publication of the Society of Behavioral Medicine* (feb 2013). https://doi.org/10.1007/s12160-012-9423-0

[25] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. *arXiv preprint arXiv:1811.00207* (2019), 5370–5381. https://www.aclweb.org/anthology/P19-1534

[26] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://arxiv.org/abs/1908.10084

[27] Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. 2018. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*. 25–36.

[28] Elsbeth Turcan and Kathleen McKeown. 2019. Dreaddit: A Reddit dataset for stress analysis in social media. *arXiv preprint arXiv:1911.00133* (2019).

[29] Joseph Weizenbaum. 1966. ELIZA—a Computer Program for the Study of Natural Language Communication between Man and Machine. *Commun. ACM* 9, 1 (jan 1966), 36–45. https://doi.org/10.1145/365153.365168

[30] Ze Yang, Linjun Shou, Ming Gong, Wutao Lin, and Daxin Jiang. 2020. Model Compression with Two-Stage Multi-Teacher Knowledge Distillation for Web Question Answering System. In *Proceedings of the 13th International Conference on Web Search and Data Mining* (Houston, TX, USA) *(WSDM '20)*. Association for Computing Machinery, New York, NY, USA, 690–698. https://doi.org/10.1145/3336191.3371792

[31] Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. 2019. CLPsych 2019 Shared Task: Predicting the Degree of Suicide Risk in Reddit Posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology* (Minneapolis).