Review 1 - Ana Sabina Uban

The core review

This paper presents an analysis of mental health support on social media, comparing the support offered by licensed mental health professionals (MHP) with support offered by regular members of the platform, from different perspectives.

Strengths: The proposed topic is novel and very interesting, and it could stimulate future research into this problem; answering the research questions proposed it could also have valuable social impact. The authors perform several analyses to understand the behavior of social media users offering mental health support, comparing professionals and non-professionals. The analyses cover a wide scope, from classifying between the two groups, to looking into the differences in the language used from both a semantic and stylistic perspective (based on LIWC categories complemented by other methods such as a language model for finding word importance, and a linguistic style matching score metric), to looking further into what kind of language elicits replies received from the original support seekers. The analyses conducted confirm some hypotheses posed by the authors, which are novel findings on the issue of online mental health support: such as the finding that MHPs elicit more responses from the support seeker, the similar distinctive features of MHP responses and those of high-scoring responses in general, or the difference between the pronoun and verb usage in MHPs vs peers (with MHPs focusing more on the second person and the future, and peers focusing more on the first person and the past). For each kind of analysis, a comprehensive picture of existing work that puts the motivation for the current study into context is given. The authors also publish the dataset used, which is the first of its kind, and has the advantage of covering many different disorders. The annotation process for finding MHP, while it is still semi-automatic (based on reddit flairs), is quite reliable compared to other ways of "soft labelling", since it involves the checking of credentials by the members of the community to confirm the MHP status. The paper is very well written and clear.

Weaknesses: One limitation I see in the methods is the reliance on LIWC categories for all the performed experiments, as well as the fact that all the analyses are performed at the word level. The Naive Bayes/SVM classifiers with unigram features/LIWC categories cannot capture information beyond unigram distribution; more sophisticated models could capture additional interesting linguistic phenomena. Similarly, while LIWC is one of the most reliable lexicons for this field, the reliance entirely on LIWC categories might miss some linguistic phenomena going beyond the word level, or including words not covered by the dictionary, or some word-topic correlations not covered in LIWC. While the paper includes a more sophisticated analysis using language models to compare word entropy, the results are still interpreted through the lens on LIWC categories. I think there could be more interesting findings from a linguistic and NLP perspective if it also leveraged methods that can go deeper into the semantic phenomenon like topic modelling, more sophisticated semantic representations, or at least word n-grams (instead of unigrams). One other methodological weakness I see is the lack of statistical significance tests for some of the experiments comparing LIWC category prevalences across the groups.

Reasons to Accept

Reasons to accept:

- novel interesting and task on particularities of MHP responses in mental health support, along with interesting and possibly impactful findings on the topic
- dataset released on social media mental health support covering many disorders
- wide range of analyses (on 4 different directions)

- extensive comparison with previous literature and motivation
- well written

Reasons to Reject

Reasons to reject:

- all the experiments rely on LIWC which might limit the depth of linguistic findings
- missing statistical significance for some comparisons

Reproducibility: 4
Reproducibility Checklist Feedback: Somewhat useful
Ethical Concerns: No
Author Identity: 1
Overall Recommendation: 3.5
Reviewer Confidence: 4
Recommendation for Best Paper Award: No

Checkbox: Defining “NLP for Social Good” and in which ways NLP can improve people's lives in various dimensions
Answer: Yes

Checkbox: Known issues, unaddressed harms, and potential damages that NLP can cause on society (e.g., political polarization, privacy issues)
Answer: Yes

Checkbox: Discussions of how NLP research can make both positive and negative impacts and novel approaches to foster the former while mitigating the latter
Answer: No

Checkbox: Quantitative and qualitative methods to assess the social impact of NLP research
Answer: No

Checkbox: Ways in which NLP practitioners can partner with practitioners from other fields to develop socially impactful research and applications
Answer: No

Checkbox: Applications of NLP for addressing socially relevant problems such as health, education, and other areas covered by (and not limited to) the UN Sustainable Development Goals. **Please note that the paper should provide evidence of the methodology applied to a real-world setting.** At the least, one should simulate the real-world setting that the system is meant to improve.
Answer: No

Checkbox: Reflecting on the NLP community's current progress for solving real-world, socially impactful problems and how to make meaningful changes toward that goal over the foreseeable future
Answer: No

Checkbox: Discussions of how heavily aligned NLP research should be with real-world topics regarding social good
Answer: No

Checkbox: Other
Answer: No

Checkbox: This paper is not related to the Theme.
Answer: Yes

Questions for the Author(s)

I was a bit confused regarding the stylistic analysis: the categories compared are stated to be function word categories (verbs, pronouns, ..etc), but then some results on LMS are reported for other categories (in the same section), such as "social" - which categories are considered in computing the LMS

Typos, Grammar, Style, and Presentation Improvements

Well written, only a couple of minor typos:

- "Does the users' behaviors reflect"
- "bacground"
- "by human annotation based the community"

In Table 2, I think a different label for the second column might be more informative and more clear (e.g. "Overall group", instead of "Comparison")

Have you read the author response?: 2
Review Update: 1

Reason for Review Update

I appreciate the responses of the authors to my questions which are now clarified. The main concerns, regarding reliance on LIWC, were commented upon in the rebuttal, and the authors brought up a fair point about the shown limitations of LIWC being related mainly to emotion expression (which the present study goes beyond). However, as the authors also mention, a deeper analysis using methods beyond the word level and beyond this lexicon might still be interesting and provide new findings, so I maintain my position on this point. The paper is still novel and the findings are interesting so I reccommend acceptance.

Reviewer read author response: 7 Apr 2021 19:31:16 GMT

Review 2

The core review

This works analyzes text from self-identified mental health professionals on Reddit in order to answer two research questions: (1) do experts have influences compared to non-experts and (2) do experts behavior reflect known counseling principles? Analyses are broken into three main tasks: (1) can experts and non-experts be automatically detected (classification task), (2) what are the linguistic differences between experts and non-experts, and (3) what language leads to further engagement with support-seekers? Additionally, a data set is constructed from mental health conversations from Reddit with self-identified mental health professionals.

The paper is addresses a novel and important task. It is well-written, easy to follow, is well grounded in the literature, and is honest about its limitations. The discussion section nicely frames the results. The authors clearly outline their research goals/questions and follow up with a related series of analyses, tying results back to the original questions at each point. I appreciate the significance tests and CIs in Figure 4. The data set also seems particularly useful for a number of mental health or CSS tasks and would probably use this in my own work.

Unfortunately, the paper uses a very limited set of linguistic features (LIWC and WordNet, plus unigrams in the classification task). I appreciate the inclusion of both