Review of: "Exploring patient experiences and concerns in the online Cochlear Implant community: a natural language processing approach"

Johannes Feldhege

Potential competing interests: The author(s) declared that no potential competing interests exist.

The paper presents an analysis of the language in a community about Cochlear Implants (CI) on the social media website Reddit. The authors employed topic modelling using BERTopic to identify the prevalent topics in this community. Their results show that experiences and concerns surrounding in the community can differ from those identified using other methods such as surveys or interviews. The paper is therefore an informative and useful addition to the literature on CIs in the field of Otolaryngology.

I have some comments concerning the methods and results presented in the paper:

- It is unclear whether the text of comments was included in the topic modelling alongside the text of the posts the comments were made to. In the section Data Collection the authors state that they preprocessed the text of comments but since the text of comments is seemingly not used by any of the subsequent analyses
- It would improve the paper if the authors could briefly state the reason and intended use for the collection of comments and post metadata in the section Data Methods. The reader is left wondering for a long time why the authors collected them.
- It should be mentioned that the authors were using a pre-trained model to generate document embeddings.
- The authors might want to state the total number of posts that were assigned to a different topic from their initial topic (with and without the posts in the “uncategorized” topic). The number of reassigned posts for each topic can be found in the Supplement but it would be helpful to have a total to judge the quality of the BERTopic algorithm and the required amount of manual work.
- It would be helpful if the authors could explain what the score of a Reddit post is as readers unfamiliar with Reddit will not know what it is.
- It would be beneficial if the authors could show with a statistical test whether there is a significant difference in number of comments and score between posts from different topics. Additionally, it would be helpful if the authors could provide some measure of variance in number of comments and score.
- It might be worthwhile to discuss the possibility of users sharing false information about CI in the community, whether that is done accidentally or intentionally and whether clinicians should take a role of correcting information in a “non-clinician” controlled space.
- It might be worth it to mention as a limitation that the authors were only able to access posts that had not been deleted by the post author or a moderator.

- In Table 1, it might be informative for the reader to also have the most frequent words for the manually curated topics. I was wondering if it is possible to identify the most frequent words from TF-IDF of the manually grouped posts in each manually curated topic.
- Figure 2 might be more informative if it showed the percentage of each topic of the total number of posts per time period. The total number of posts in the community has obviously grown a lot leading to increases in posts per topic. However, that could obscure changes in importance of topics over time.