CL4Health 2024 Reviews for Submission #13

Title: Unveiling Voices: Identification of Concerns in a Social Media Breast Cancer Cohort via Natural Language Processing
Authors: Swati Rajwal, Avinash Kumar Pandey, Zhishuo Han and Abeed Sarker

REVIEWER #1

Detailed Comments

The authors explore breast cancer patients' concerns and possible reasons for treatment discontinuation in a Twitter dataset via a combination of sentiment analysis and topic modelling.
The main contribution of the paper does not consist in suggestiong new NLP methods, but rather applyng the existing tools for an important practical task. The importance of the task is well-justified, and I would suggest that the findings present interest for the clinical community.

One question to the methodology used is the choice of the sentiment analyser. Though NLTK is perhaps the most accessible tool for this task, it may not always be the best one. It would be interesting to know whether authors tried any other tools (from what I remember, StandordNLP stanzapackage has a sentiment analysis tool, as does AllenNLP; there are probably available solutions within spacy universe, and there are definitely models on Huggingface, there may even be some trained specifically on medical data). Having an idea of how good/bad the sentiment analyser is would help make the conclusions of the paper better grounded. It would also be interesting to see what topics occur in the 'positive'/'neutral' part of the dataset, and in the dataset as a whole.

Response: Thank you for your feedback. We implemented a script (and uploaded to the github repository) to test if changing the sentiment analyzer (keeping everything constant) affected the topics identified.For alternative sentiment analyzers, we used Stanford’s Stanza package and Huggingface Sentiment analyzers to test the robustness of our results. We found that, similar topics/concerns were identified across all the sentiment analyzers. This gives us more confidence in our results.

==

REVIEWER #2

==

Detailed Comments

## Missing Data Related Information:

- It would greatly enhance the paper's credibility to provide detailed statistics regarding the dataset, such as the source of the ~1.5 million tweets and any preprocessing steps applied. It helps to get a clearer understanding of the data's origins and any potential biases.
- I found myself wondering about the connection between the manually annotated data from Al-Garadi et al. (2020) and the larger dataset of ~1.5 million tweets.

Response: Thank you for your thoughtful questions and suggestions. We have added dataset statistics such as size, unique ids, train-test-dev data sizes used for training classifiers in the updated manuscript under section 1.2.

The 1.5 million dataset is used for topic modeling/ concerns identification while the manually annotated data is used for Roberta (and other traditional ML) classifier train and evaluation. Afterwards, we employ the Roberta Classifier to identify self reported tweets on the 1.5 million tweets dataset.

- It would be helpful to clarify how the classifier was evaluated and whether the results can be compared with the original dataset paper, to be able to assess the validity of the findings.

Response: The classifier was evaluated on a held out (never used in training process) test dataset.

All the classifiers were trained on a training dataset, with hyper-parameter tuning using the validation dataset and the final testing on a held-out dataset. Our train-validation-test dataset construction closely follows the original dataset paper to maintain the validity of our self-report tweets classifier. We then compare all our classification model’s performance (reported in the paper) and selected the best performing classifier for our primary objective of identifying concerns amongst the Breast Cancer cohort tweets.

- When applying the trained classifier on the entire dataset, it would be valuable to know the number of instances classified as self-reports.

Response: Thank you for the feedback. We have added this detail in section 3.1 in our updated manuscript.

## NLP Pipeline Explanation:

- While the NLP pipeline diagram provides an overview, it seems like quite a lot of information is missing from the depiction itself.

Response: Thank you for your feedback. We have now updated our pipeline diagram to be more thorough with step-by-step marked workflow depicted in the diagram (Figure 1). We have also the added Roberta Classifier and Sentiment filtering in our new pipeline diagram along with step-wise numeric markings for better reader interpretability and understanding.

- Could you explain where the self-report classifier and the sentiment classifier would fit into the pipeline?

Response: We have added this detail in figure 1. To make it more easier to follow, we have added numbering to highlight the sequence of processing that takes place. But to explain here, the self-report classifier is used to classify self-reported tweets from the 1.5 million tweet dataset. After this step, we perform sentiment analysis and retain the negative sentiment post for further analysis in our topic modeling framework. Furthermore, to establish the robustness of our results, we used different packages of sentiment analyzers (additional scripts added in github repository)

## Improvement in Presentation and Clarity:

- Overall, the paper could use improvement in terms of presentation and clarity. Especially in terms of data and the NLP pipeline.
- It's important to clearly communicate the contributions of the study beyond its stated objectives.

Response: Thank you for suggestion. The main contribution of this work is to to identify the major concerns that self-reported breast cancer cohort experiences by leveraging NLP toolkits and social media dataset. These concerns are not usually captured in traditional EHR reports.We believe that the insights derived from this study could be used towards motivating public health policy making and breast cancer awareness programs

----

Questions for Authors

- What was the motivation for setting negative sentiment threshold set to less than 0.5 (Section 2.3). Did you consider neutral label as well ?

Response: We captured all the negative and near-neutral posts by filtering the sentiment score with <= 0.5. The goal was to capture the broader self-reported breast cancer tweets that indicate a wide range of negative sentiment.

- In Figure two, I get the overall idea is related to the contribution of certain words to the cluster. What does the numbers in the x-axis refer to.

Response: The X-axis represents the scores/weights of each term (on y-axis) within a specific topic. These scores show how relevant a term is to its corresponding topic. We have also mentioned this in the manuscript section 3.2.

Also, why are there two different clusters (blue and pink) for "Covid Concerns" ?

Response: The different colors for each topic in the barchart by BERTopic (python package) is used for visual clarity. The colors are automatically selected by the package and does not convey any additional meaning beyond separation of the topics. We have also added this note in the manuscript section 3.2.