Paper Review: How Do You Feel? Information Retrieval in Psychotherapy and Fair Ranking Assessment

BY: Vivek Kumar, Giacomo Medda, Diego Reforgiato Recupero, Daniel Riboni, Rim Helaoui & Gianni Fenu

2.1)

Context

Due to the onset of the COVID-19 pandemic, people were forced to isolate themselves, which resulted in increased mental health issues like depression and anxiety. Being unable to interact with others led to low access to professional help in such cases. Due to this, the authors felt that people should have reliable therapeutic content that is publicly accessible and readily available to allow people to get help for their individual psychological needs. The paper uses Information Retrieval (IR) methods to rank the content and provide the individual with the content best suited for their personal needs.

Problems

It has been found that high-quality, publicly accessible data is scarce in the healthcare domain and, thus, the psychological domain. For this reason, it is challenging to train IR models to provide aid in mental issues successfully. Since there isn't much data to go on for this task, fairness ranking is extremely important to ensure that there are no biases towards a particular psychological issue over others. This approach makes sure that the model is able to provide the required support to all individuals regardless of how many instances of their particular issue are available in the training dataset. It is also important to keep in mind the sensitive nature of the task at hand, suggesting the need for professional help for quality assessment and annotations since therapy sessions are highly subjective.

Contributions

The authors use the first publicly available psychological dataset, the AnnoMI dataset, which has professionally annotated Motivational Interviewing (MI) therapy sessions which are labeled as either high-quality or low-quality. They are successfully able to use state-of-the-art IR models on the AnnoMI dataset to rank the therapy sessions based on quality such that higher quality sessions are placed above the lower. The paper also evaluates the fairness of the ranking system for the distinct psychological topics. It finds that unfairness depends on the model used and is caused by the uneven distribution of high-quality vs low-quality content. One of the key contributions by the authors is the provision of a publicly accessible source code, which makes future research in the field easier and provides researchers with the opportunity to develop the application of IR models further towards mental health and psychological needs.

2.2) Critically evaluate

What is the type of paper?

This paper by Vivek Kumar et al. is an applied research paper that uses IR models to address the specific real-world problem of ranking therapeutic content based on psychological needs while maintaining the quality and fairness of each topic. While the paper provides sufficient theoretical evidence and reasoning, it also applies various IR models to the dataset.

How well are the research questions addressed?

The following research questions are addressed in the paper:

To what extent can the ranking utility be considered reliable for IR tasks in therapeutical settings? The authors used the AnnoMI dataset, which has professionally labeled annotations, which gives the authors a basis for a ranking based on quality. It is seen that the chosen IR models can successfully distinguish between high and low quality content, proving that the ranking utility used is reasonable and considered reliable for therapeutic sessions. Although, the usage of a single dataset limits how well the results can be generalized and adds slight doubt in the reliability of the ranking utility for other therapeutic content. Making use of a greater number of therapeutic data would solidify the reliability of the ranking utility.

How does the ranking utility variation across different psychological topics impact the fairness of ranking algorithms? The fairness is evaluated by checking if the IR models are able to provide rankings over the range of topics, with every topic getting a fair shot at being represented in the higher rankings. The paper would do better to address other aspects that could address fairness, such as gender or age, as well as specific issues within each topic, such as the severity of the symptoms.

Do the rankers systematically exhibit bias towards specific psychological diseases that result in negative impacts on the ranking utility? The authors evaluate if particular topics frequently rank either higher or lower, which helps identify potential bias within the IR model itself that may favour a particular psychological disease. In doing so, the authors are able to address this research question. Even so, they are unable to find the reasons why some topics may rank higher than others. Finding the reasons for this difference in ranking could include a more in depth evaluation of the models and include metrics for fairness consisting of statistical tests as well.

Is the approach realistic?

The approach is only realistic to an extent and has certain limitations. While the proposal to use IR models to rank therapeutic content according to quality is highly creative, it is only feasible for a dataset such as AnnoMI. Its reliability cannot be guaranteed for other datasets. Using this approach for real-world applications will be tricky as there is a scarcity of well-annotated and high-quality psychotherapeutic datasets. While the approach does lay a good foundation, real-world implementation would require addressing the limitations of the data, using more in-depth evaluation metrics and making sure that the approach can be adapted for different therapeutic settings.