Peer to Peer Support

Using machine learning to identify individuals who are at risk of harm

The Alan Turing Institute

Overview

A for-profit company has developed a peer-to-peer support app. Their aim is to help users by allowing them to share experiences and seek advice from others who have experienced mental health issues.

Use of the service is offered free of charge to users by many schools, universities, and workplaces who subscribe to the service as a way to improve the productivity and wellbeing of staff, students or pupils.

The idea is for users to feel comfortable sharing personal stories and as a result all users are kept anonymous while using this tool. They can freely chat to one another without revealing their identity.

In most cases, users can message each other without external intervention. But, if the technology detects the use of certain combinations of 'trigger words', the incident is escalated. Anonymity is then broken so that their institution is notified and the user can be given personal help from a practitioner.

When consenting to use this platform, users are told that anonymity may be breached in extreme cases. But specific details about why and when anonymity is broken are not provided.

Key Consideration

An organisation may have multiple reasons for offering this system.

On the one hand, they genuinely want to improve wellbeing. On the other hand, the system also boosts productivity.

Deliberative prompts

1. **When do you think it is and is not appropriate for this service to break a user's anonymity in order to give them personal help?**
2. **Would it be better if there was no monitoring of the chats taking place on this system?**
3. **Do the power imbalances between the user and the institution offering this service affect your views on whether breaking anonymity is justified?**

Datasheet

This datasheet gives details on the information collected by this service and available to the developers.

Category	Details
Available Data	• Transcripts of the conversations between users • Metadata taken from the app, including timestamps for messages and the IP address of the user • Survey results from users of the service (e.g. how useful they find the app)
Type of Technology	An algorithmic technique known as 'natural language processing' is used by the system to automatically monitor the conversations of the users. These conversations remain anonymous unless the algorithm detects a phrase that indicates a possible risk of harm. A trained human professional then reviews the conversation before the user is identified and their organisation is contacted.

Groups, Organisations and Affected Individuals

1. **Users of the support forum**
2. **For-profit company**
3. **Universities, schools and employers who may buy this service**