A collection of real-life examples of the risks that are faced when using data in advocacy work, along with mitigation strategies to overcome these challenges.

Mental health-related alerts from Twitter

The case of the Samaritans[1].

Context

Samaritans' work happens primarily through people in need contacting them, via email, telephone, or in person, in addition to their outreach work including workplace training, and work in schools. Samaritans launched the first 24-hour telephone helpline in the UK, and they train volunteers to be "listeners", helping people work out their own way forward.

They identified a number of cases of individuals sharing their thoughts about suicide on Twitter, not receiving a response and tragically going on to take their own life; so, they started a new project to try and prevent this from happening.

1 **Samaritans** are a non profit organisation based in the UK, with 201 branches across the UK and the Republic of Ireland, and a volunteer community of approximately 20,000 people. They provide emotional support to people in times of need.

Their aim:
moving the Samaritans' "listening ethos" into the realm of social media.

On 29 October 2014, Samaritans launched the Samaritans Radar, an "online app designed to offer people a second chance to see a tweet from someone they know who might be struggling to cope."

The app, Samaritans Radar, monitored the twitter feed of app users to see if anybody followed by the user had tweeted specific keywords or phrases that had been identified as being more used by people who are struggling to cope, such as "I hate myself". If a tweet was found with those keywords, the user would receive an email with a link to that tweet, along with suggested guidance of actions they could carry out.

The response to the application was mixed; some lauded its innovative approach to using social media, but others–notably, many from the mental health community in the UK–reacted strongly against the app.

In response to the negative reaction to Radar, the Samaritans suspended the app just nine days after the launch, and in March 2015, they announced that the app would be permanently closed, and all associated data would be deleted.[2]

How it worked

By default, searches for keywords/phrases were carried out on the tweets of everyone followed by users who signed up to the app. The keywords and phrases were based on research undertaken by Jonathan Scourfield at Cardiff University as part of the **COSMOS project**,[3] looking at **the relationship between social media and suicide**.[4]

If a certain user signed up, then all of the tweets from anybody that they followed were then included within the app and scanned for the designated keywords; and based on the use of certain words which they had identified as commonly being used to indicate "suicidal intent", the tweets were included in notifications to the app user.

As a result, a person's tweets could have been included within the app without that person having any idea, and people following them could have been notified if the Twitter user used a predefined "trigger phrase" unknowingly. At the start the Samaritans site made this very obvious, saying that the people a user follows on Twitter will not be notified that a user has signed up to the app, and all alerts would be sent directly to the user's email address. The Samaritans justified this decision by saying:

> *The app works in such a way that the alerts sent out are only seen by the subscriber, who would have sight of the information anyway. Samaritans does not monitor the tweets or view them – we're just giving people who have signed up to Radar a second chance to see a call for help, which they might*

2 **http://www.samaritans.org/how-we-can-help-you/supporting-someone-online/samaritans-radar#10mar**
3 **https://www.cs.cf.ac.uk/cosmos/**
4 **https://www.cs.cf.ac.uk/cosmos/research-on-social-media-and-suicide/**

have initially missed, from a friend that is in need of support."
Sophie Borromeo, director of communications at the Samaritans,
quoted in the Guardian on Nov 3rd 2014[5]

Essentially, what the app did was make very visible and explicit anything that someone tweeted publicly with predefined "keywords". Technically, anything tweeted from an open account (ie. as opposed to a "closed" account where the user approves individually anyone who wants to follow their tweets) is indeed public, but often it is thought of by users as 'private', especially by users who have very few followers, or who envision that very few people see their tweets.

What's different between this and how Twitter usually works?
Given the fleeting nature of Twitter–that what you see changes rapidly, that it's easy to miss tweets that are sent out by people at certain times of day, tweets are understood to have different levels of visibility. If you tweet at another person without putting a "." before it, for example, only you, that person, and people who follow both of you will see it in their timeline. So, it's a reasonable assumption for a Twitter user with few followers, tweeting when the majority of those followers are asleep, that very few (if any) people will actually see that tweet. What the app essentially did was change that level of visibility.

Mixed responses

Changing that level of visibility, and flagging up every occasion when someone tweeted with those keywords was not seen as helpful by some members of the mental health community.

As written by a former Samaritans volunteer,

"How likely are you to tweet about your mental health problems if you know some of your followers would be alerted every time you did? Do you know all your followers? Personally? Are they all friends? What if your stalker was a follower? How would you feel knowing your every 3am mental health crisis tweet was being flagged to people who really don't have your best interests at heart, to put it mildly? In this respect, this app is dangerous. It is terrifying to think that anyone can monitor your tweets, especially the ones that disclose you may be very vulnerable at that time"–**blog post, Oct 29, 2014**[6]
by @elphiemcdork

5 www.theguardian.com/voluntary-sector-network/2014/nov/03/samaritans-radar-twitter-mission-charity
6 emsyblog.wordpress.com/2014/10/29/the-samaritans-radar-app-the-problem-is-right-there-in-the-name/

But others working in the charity sector praised the Samaritans for leading the way in moving their work into the digital realm, **saying that controversy was inevitable with such innovation**.[7]

Challenges faced

As highlighted by those within the mental health community themselves, people suffering from mental health issues felt like the app meant they might have had to self-censor on Twitter. Some felt like their privacy was invaded by the app changing that level of visibility; essentially meaning that they couldn't casually tweet something without others being notified.

For those facing online or offline violence, the app provided potentially more information to allow stalkers or bullies to target people when they were at their most vulnerable. By making it easier for those concerned to see "worrying" tweets, it also made it easier for others–perhaps those with malicious intent–to see the same tweets, and thus identify times of particular vulnerability.

It seems as though during development of the app, primarily positive and useful uses of the data and the app itself were envisioned. Additionally, it was assumed that writing anything negative on Twitter was an explicit cry for attention and cry for help, rather than simply an expression online of one's feelings at that time. As it turned out, many people in the mental health community use Twitter as a way of expressing themselves in a more intimate way than had been envisioned given the technically "public" nature of tweets.

Mitigation strategy

In terms of the technical functionalities, the app included a "whitelist" function; initially, if organisations didn't want their tweets to ever be included within the tweets that were scanned by the app, they could send a direct message to @Samaritans. This was designed for organisations who tweeted regularly with the designated "trigger phrases". Following the negative feedback, this functionality was extended to individuals who didn't want their tweets to be included; however, by default the app was 'opt-out' rather than 'opt-in', meaning that users had to write to the Samaritans to get on the whitelist.

The app was suspended nine days after launch, and permanently closed five months after it began. During the nine days after launch, the Samaritans team put out a number of updates and press releases, all of which are still public and available online.

7 www.theguardian.com/voluntary-sector-network/2014/nov/03/samaritans-radar-twitter-mission-charity

Following the app's suspension, the organisation issued an apology which stated:

> *"We've learned that we must consult even more widely than we have done in the development of Samaritans Radar and we will continue to respect and better understand the diversity of existing communities and users. To this end, we will be holding a series of consultation events as well as continuing to gather views via an online survey from as wide a range of people as possible."*

Importantly, following the app's closure, Samaritans carried out a 6-month knowledge and learning project entitled Digital Futures, with the stated aim of "finding out people's views on the opportunities and challenges for emotional support and suicide reduction presented by the online environment." Their comprehensive findings were all published online in November 2015, and made very clear that they had carried out a broad public consultation, talking to privacy and data experts. These findings outlined areas that the Samaritans could improve in for future digital engagement, suggested by their users, and showed clearly that they had learned from their experience with Radar.