Algorithms, social media and mental health

Heike Felzmann and Rónán Kennedy, NUI Galway

Computers & Law, October/November 2016, p.31-34, http://www.scl.org/site.aspx?i=ed48960

The Samaritans recently decided to harness the power of social media to support persons suffering from mental distress and suicidality. They developed the "Samaritans Radar" app that was designed to identify people who had posted phrases indicating suicidality on social media (Samaritans 2016). The app sent alerts to its users if anyone they were connected to on social media used high risk phrases. The intention of the app designers was to facilitate the timely provision of support to those in need. However, the app was criticised despite its initial popularity, due to concerns that the app could be used in a predatory or exploitative way to identify vulnerable persons. The Samaritans suspended the app shortly after its introduction and ultimately deleted it. Even though it merely re-used social media information that was already available to users, it became apparent that this technology had wider and more troubling applications than originally envisaged.

Mental health information has also been gleaned from much more innocuous seeming information. Recently, Reece and Danforth (2016) identified a number of predictive markers of depression from pictures posted on Instagram, including in particular preferred content, colour schemes and the use of specific filters for image processing. According to the authors, their statistical model was able to identify individuals before they had received a formal diagnosis of depression and outperformed general practitioners' with regard to accuracy of diagnosis.

As these examples show, algorithmic identification of mental health characteristics is feasible on the basis of easily available information on social media. An extensive knowledge base regarding characteristics of persons with certain mental health conditions is already available from research in psychology, public health and psychiatry. Such information can be extracted by algorithmic methods from publicly available information. The use of certain diagnostic terms or critical phrases in a person's online activity (as for example in the Samaritans Radar app), the connections of this person with other users who show similar characteristics, or the use of online fora or groups dedicated to mental health issues are the most obvious cases. However, by means of machine learning on identified individuals with a verified diagnosis it is also possible to identify additional, much less obvious characteristics that are related to the diagnosis (as the Instagram case illustrates). It is increasingly possible with public social media data to identify a multitude of patterns associated with certain conditions, especially characteristics with regard to social activity and connectedness, affective states, or use of language. For example, a social media research group around De Choudhury has performed comprehensive analyses with regard to patterns associated with depression (e.g. De Choudhury et al. 2013) and they are expanding their analysis to a number of additional mental health conditions. A group around Coppersmith (2015) have focused specifically on the differential analysis of language patterns of persons with different mental health diagnoses. With regard to depression it has been identified, for example, that metadata such as the number of social connections, as well as number and timing of posts can provide indications of a diagnosis (De Choudhury et al. 2013). Other characteristics that are independent from provided content, like

natural typing patterns which could be captured in real time by any computer that a person uses, have also been identified as candidates for the early identification of subtle psychomotor impairment (Giancardo et al. 2015), with potential applications for early diagnosis of Parkinson's disease or dementia. It is highly likely that these analyses will be further refined and additional, as yet unidentified parameters will be identified in due course.

While capturing and tracking affective states and attitudes has long been pursued by companies due to their commercial significance for consumers' engagement with products and services, the algorithmic assessment of mental health states on social media represents a further development that raises significant legal and ethical concerns. This type of second-order processing of data which is 'public' in a wide sense raises interesting data protection issues. Initially, of course, much of this is personal data (and therefore subject to the basic protections of the law), but once it is processed in such a way as to enable predictions or conclusions (however preliminary) about an individual's mental health, it becomes sensitive personal data and the data controller must comply with a more demanding set of conditions.

An obvious preliminary question is whether this data is fairly obtained in the first place. While many think of social networking services (SNS) as 'public spaces' (and posts and comments there can often be reported in the media as if they were published into the public domain), these are in fact legally walled gardens to which one can have varying levels of access depending on whether or not one has an account with a given service, and what type of account it is. It is certainly arguable that anything posted on the general Internet is public, as is information which is posted on the public elements of one's profile on Facebook, Twitter, and so on. However, much of (for example) Facebook content is only visible to those who have an account with that service, and some of this may only be visible to those whom a particular user acknowledges as a 'friend'. Data controllers who are taking information from within those services for processing in order to identify those with mental health issues may be in breach of the terms of service. Facebook's current *Statement of Rights and Responsibilities* (dated 30 January 2015) contains a number of provisions that seem relevant here, such as "You will not collect users' content or information, or otherwise access Facebook, using automated means (such as harvesting bots, robots, spiders, or scrapers) without our prior permission.", "You will not use Facebook to do anything unlawful, misleading, malicious, or discriminatory.", You will not post content or take any action on Facebook that infringes or violates someone else's rights or otherwise violates the law.", "If you collect information from users, you will: obtain their consent, make it clear you (and not Facebook) are the one collecting their information, and post a privacy policy explaining what information you collect and how you will use it." Instagram's terms of service are more ambiguous, saying only "We prohibit crawling, scraping, caching or otherwise accessing any content on the Service via automated means, including but not limited to, user profiles and photos (except as may be the result of standard search engine protocols or technologies used by a search engine with Instagram's express consent)." Other services may have different prohibitions; the analysis would need to be conducted on a case-by-case basis. What is important to highlight is the extent to which third parties may not be able to legally collect data for algorithmic mental health analysis.

However, this preliminary hurdle can, it seems, be easily overcome by the SNS itself. It is the gatekeeper and has access to all of its users' data, including words typed, typing patterns, and colour preferences. However, other data protections will come in play. The data must be obtained for a

specified, explicit, and lawful purpose. Facebook's Data Policy says that it collects data so that it can "deliver our Services, personalize content, and make suggestions for you by using this information to understand how you use and interact with our Services and the people or things you're connected to and interested in on and off our Services." and "to improve our advertising and measurement systems so we can show you relevant ads on and off our Services and measure the effectiveness and reach of ads and services." This is a broad statement of purpose; it may not be explicit enough to include "analysing the state of your mental health". Details of one's typing patterns may also be excessive for this purpose.

Of course, it is possible to argue that this information has been made public as a result of steps deliberately taken by the data subject. While bearing in mind the caveat noted above, that SNSs are not truly 'public', this may be true for certain types of content. If I post about how much I dislike going to work, or the dreary weather, or over-crowding on trains, and particularly if I do this on a regular basis, then it is difficult for me to argue that this is not information about my internal state which I have consciously decided to reveal to the world at large. However, some of the algorithmic studies of mental health rely on metadata, such as timing, number and length of postings or the number of social connections, or on hidden information which is buried in the content produced, such as choice of colour filters, key phrases that seem innocuous unless tied to other indicators of depression or typing patterns. As Reece and Danforth (2016) highlight, many of the features identified by their algorithmic analysis did not correspond to lay assumptions about characteristics associated with depression. With such features, it is much more difficult to claim that the data has been deliberately revealed.

Both the existing Data Protection Directive (DPD) and the incoming General Data Protection Regulation (GDPR) control the taking of automated decisions; however, this prohibition may not apply to this particular type of profiling, at least for now. Under the DPD, the decision has to have a significant effect on the individual concerned, and an indication of possible poor mental health may not meet this threshold. In the future, this may change. The GDPR requires a data protection impact assessment before undertaking a profiling exercise (Recital 91), but this may only apply where the decision has "consequences" for the individuals concerned (see Recitals 60 and 63).There seems to be a distinction between consequences broadly defined, and consequences "which produces legal effects concerning [the data subject] or similarly significantly affects [the data subject]" (Article 22). Once the GDPR enters into force in May 2018, these types of algorithm-based mental health predictions may become more difficult to undertake, at least in Europe (Heimes, 2016).

Necessity can be a defence, particularly to prevent injury or other damage to the health of the data subject, or for medical purposes. Mental health disorders represent a significant public health concern. According to Kessler and others (2012), it has been estimated in a US context that major depression alone as the most common mental health condition is likely to affect close to 30% of the population during their lifetime (Lifetime Morbid Risk LMR of 29.9%). Anxiety disorders like social anxiety (LMR 13%), posttraumatic stress disorder (LMR 10.1%) or panic disorder (LMR 6.8%) similarly affect a significant proportion of the population. These conditions all have significant impact on sufferers' quality of life and are associated with additional risks and burdens. Algorithmic diagnosis could therefore be seen as an early detection method that could facilitate timely interventions with positive impact on the affected individual's mental health and public health more generally, as for example De Choudhoury et al. (2013) or Reece & Danforth (2016) suggest. However, although it may

seem helpful to monitor the mood and affect of users of SNSs, and intervene where it seems appropriate - perhaps through targeted advertising, a reminder to consider seeking help, or the presentation of advice - this falls short of being necessary. A desire to be of assistance, particularly proactively before a problem is identified by the data subject, is not the same as a requirement to act.

What further complicates the issue is that the algorithmic identification of such disorders is not just of interest for those who act in the interest of public health goals, but also for those with commercial interests. Companies may be interested in such information either to identify marketing targets (such as providers of potential remedies and therapies in search of susceptible customers) or as a negative or exclusion criterion (such as insurance providers or employers seeking to avoid high-risk customers or employees).The diagnosis of a mental health condition is highly sensitive and could have a range of negative impacts on those who are identified as having such condition. Information about a person's mental health is generally identified within healthcare relationships and enjoys strong protections with disclosure in most circumstances only at the individual's discretion. The use of algorithmic diagnosis allows going far beyond those traditional protected contexts of diagnosis and opens the door to new practices that may circumvent existing professional protections of confidentiality with regard to health information.

Once such algorithms have been established they can be applied widely by anybody with access to these tools and to relevant information, without awareness and consent of those affected. It could be argued that being the object of hidden algorithmic analysis is an unavoidable part of using major internet and social networking services and is the price that we pay for services that are provided free of charge. However, as indicated above, it is unlikely that persons using SNS have been aware (or could reasonably be expected to be aware) of the extent to which their seemingly innocuous contributions may implicitly contain identifiable features indicative of mental health and other sensitive characteristics, and may be used, by their own service provider or others who may gain access to their data, for purposes that may run counter to their interests. From an ethical perspective, there is something particularly problematic about using algorithmic analysis to identify implicit characteristics that carry a stigma and may have negative repercussions for persons, especially if this is done on the basis of characteristics that could not reasonably have been envisaged to carry this kind of significance. It is also of doubtful legality, particularly as European data protection law becomes more stringent.

References:

Coppersmith, G., Dredze, M., Harman, C., & Hollingshead, K. (2015). From ADHD to SAD: Analyzing the Language of Mental Health on Twitter through Self-Reported Diagnoses. *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, Denver, Colorado, USA.

De Choudhury, M., Gamon, M., Counts, S. & Horvitz, E. (2013). Predicting Depression via Social Media. *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media* (*ICWSM),* 128-137, https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/download/6124/6351

Facebook (2015). *Statement of Rights and Responsibilities*, January 30, 2015, https://www.facebook.com/terms [Accessed September 20, 2016]

Giancardo, L., Sánchez-Ferro, A., Butterworth, I., Mendoza, C. S., & Hooker, J. M. (2015). Psychomotor impairment detection via finger interactions with a computer keyboard during natural typing. *Scientific reports*, 5, 9678(2015), doi:10.1038/srep09678.

Heimes, R. (2016) Top 10 operational impacts of the GDPR: Part 5 - Profiling. *The Privacy Advisor*, https://iapp.org/news/a/top-10-operational-impacts-of-the-gdpr-part-5-profiling/ [Accessed September 20, 2016]

Kessler, R.C., Petukhova, M., Sampson, N.A., Zaslavsky, A.M., and Wittchen, H.U. (2012). Twelve-month and lifetime prevalence and lifetime morbid risk of anxiety and mood disorders in the United States. *International Journal of Methods in Psychiatric Research*, *21*(3), 169–184. http://doi.org/10.1002/mpr.1359

Reece, A. & Danforth, C. (2016). Instagram photos reveal predictive markers of depression, https://arxiv.org/ftp/arxiv/papers/1608/1608.03282.pdf

Samaritans (2016) *Samaritan's Radar updates*, http://www.samaritans.org/how-we-can-help-you/supporting-someone-online/samaritans-radar