Contents lists available at ScienceDirect

Internet Interventions

journal homepage: www.invent-journal.com/

Detecting suicidality on Twitter

Bridianne O'Dea [a,*], Stephen Wan [b], Philip J. Batterham [c], Alison L. Calear [c], Cecile Paris [b], Helen Christensen [a]

[a] Black Dog Institute, The University of New South Wales, Hospital Road, Randwick, NSW 2031, Australia
[b] Commonwealth Scientific and Industrial Research Organisation (CSIRO) Information and Communication Technology Centre, Corner of Vimiera and Pembroke Roads, Marsfield, NSW 2122, Australia
[c] National Institute for Mental Health Research, Building 63, The Australian National University, Canberra ACT 2601, Australia

ARTICLE INFO

Article history:
Received 28 January 2015
Received in revised form 24 March 2015
Accepted 25 March 2015
Available online 7 April 2015

Keywords:
Twitter
Suicide
Social media
Machine learning
Prevention
Big data
Online

ABSTRACT

Twitter is increasingly investigated as a means of detecting mental health status, including depression and suicidality, in the population. However, validated and reliable methods are not yet fully established. This study aimed to examine whether the level of concern for a suicide-related post on Twitter could be determined based solely on the content of the post, as judged by human coders and then replicated by machine learning. From 18th February 2014 to 23rd April 2014, Twitter was monitored for a series of suicide-related phrases and terms using the public Application Program Interface (API). Matching tweets were stored in a data annotation tool developed by the Commonwealth Scientific and Industrial Research Organisation (CSIRO). During this time, 14,701 suicide-related tweets were collected: 14% were randomly ($n = 2000$) selected and divided into two equal sets (Set A and B) for coding by human researchers. Overall, 14% of suicide-related tweets were classified as 'strongly concerning', with the majority coded as 'possibly concerning' (56%) and the remainder (29%) considered 'safe to ignore'. The overall agreement rate among the human coders was 76% (average $\kappa = 0.55$). Machine learning processes were subsequently applied to assess whether a 'strongly concerning' tweet could be identified automatically. The computer classifier correctly identified 80% of 'strongly concerning' tweets and showed increasing gains in accuracy; however, future improvements are necessary as a plateau was not reached as the amount of data increased. The current study demonstrated that it is possible to distinguish the level of concern among suicide-related tweets, using both human coders and an automatic machine classifier. Importantly, the machine classifier replicated the accuracy of the human coders. The findings confirmed that Twitter is used by individuals to express suicidality and that such posts evoked a level of concern that warranted further investigation. However, the predictive power for actual suicidal behaviour is not yet known and the findings do not directly identify targets for intervention.

1. Introduction

The World Health Organization recently reported that on average, a suicide occurs every 40 s (World Health Organization, 2014). Worldwide, an estimated 804,000 suicide deaths occurred in 2012, representing an annual global age-standardised suicide rate of 11.4 per 100,000 population, 15.0 for males and 8.0 for females. Furthermore, there are up to 20 times as many adults who attempt suicide (World Health Organization, 2014). Suicide has a devastating impact on families (Cerel et al., 2008) and communities (Levine, 2008), and many suicide deaths are preventable (Bailey et al., 2011). Understanding the ways in which individuals communicate their suicidality is key to preventing such deaths. Suicidality is defined as any suicide-related behaviour, thoughts or intent, including completing or attempting suicide, suicidal ideation or communications (Goldsmith et al., 2002). Suicidal ideation is defined as thoughts about killing oneself, while suicidal behaviours involve acts of self-harm with the intention of causing death (Goldsmith et al., 2002). While not all individuals experiencing suicidal ideation will plan or make an attempt on their life, such ideation places individuals at increased risk of death by suicide (McAuliffe, 2002). In face-to-face settings, suicidality is usually uncovered by an outright disclosure of intent, or by asking an individual about their thoughts and actions. Some individuals have communicated their suicidal thoughts and plans to friends and family prior to suicide (Wasserman et al., 2008; Wolk-Wasserman, 1986); however, it is accepted that many do not disclose their intent. Recently, individuals have broadcast their suicidality on social media sites such as Twitter (Jashinsky et al., 2013), indicating that this social media site may have potential for use as a suicide prevention tool (Luxton et al., 2012).

Twitter is a free broadcast social media site that enables registered users to communicate with others in real-time using 140 character statements. Users create a network by following other accounts; although, the large majority of Twitter accounts are public which allows

[*] Corresponding author. Tel.: +61 2 9382 8509.
E-mail addresses: b.odea@blackdog.org.au (B. O'Dea), stephen.wan@csiro.au (S. Wan), philip.batterham@anu.edu.au (P.J. Batterham), alison.calear@anu.edu.au (A.L. Calear), cecile.paris@csiro.au (C. Paris), h.christensen@blackdog.org.au (H. Christensen).

http://dx.doi.org/10.1016/j.invent.2015.03.005

anyone to view their content. Twitter content can be posted via a web interface, SMS or a mobile device. It is available in almost all countries except China, Iran and North Korea, and has no minimum age requirement. Approximately 23% of online adults use Twitter and over 500 million tweets are sent per day (Duggan et al., 2015). Twitter has recognised that individuals express suicidality in their broadcasts and have created internal mechanisms that allow it to be reported (Twitter Inc, 2014). If deemed serious, Twitter can provide the account holder with crisis support services. This type of risk detection is not automatic, does not occur in real-time, and relies solely on the discretion of networked users, of whom many have difficulty determining genuine risk (Wolk-Wasserman, 1986). Similarly, clinicians have reported monitoring patients' mental health via social media and they too are uncertain about the sincerity of posts, their duty of care and the ethics of intervention (Lehavot et al., 2012). Given the large volume of Twitter data, it is not yet feasible or ethical to directly contact and survey every Twitter user who may be at risk. The parameters of this risk are yet to be determined. Previous studies have collected and classified suicide-related tweets (Jashinsky et al., 2013); however, data sets remain small and modelling for automatic detection is in its infancy. Although Twitter may provide an unprecedented opportunity to identify those at risk of suicide (Jashinsky et al., 2013) and a mechanism to intervene at both the individual and community level, valid, reliable and acceptable methods of online detection have not yet been fully established (Christensen et al., 2014). Best practice for suicide prevention using social media remains unclear.

2. Aims

This study aimed to establish the feasibility of consistently detecting the level of concern for individuals' Twitter posts, colloquially referred to as 'tweets', which made direct or indirect textual or audio-visual references to suicidality. Using a set of instructions and categories, human coders aimed to do this using only the content of the tweet itself. Following this process, this study aimed to design and implement an automated computer classifier that could replicate the accuracy of the human coders. The feasibility of this automated prediction was to be examined using recall and precision metrics.

3. Methods

The method of the current study consisted of three main steps: i) data collection, ii) human coding and iii) machine classification. Section 3.1. outlines the collection of suicide-related tweets using Twitter's public Application Program Interface (API). Tweets were stored in a data coding tool developed by the Commonwealth Scientific and Industrial Research Organisation (CSIRO). Section 3.2. outlines the human coding conducted by the researchers. The machine learning processes that were applied to acquire a predictive model for the automatic identification of 'strongly concerning' tweets are outlined in Section 3.3.

3.1. Data collection

Twitter offers a public API which enables programmatic collection of tweets as they occur, filtered by specific criteria. From 18th February 2014 to 23rd April 2014, this API was used within a tool developed by the CSIRO to monitor Twitter for any of the following English words or phrases that are consistent with the vernacular of suicidal ideation (Jashinsky et al., 2013):

> "suicidal; suicide; kill myself; my suicide note; my suicide letter; end my life; never wake up; can't go on; not worth living; ready to jump; sleep forever; want to die; be dead; better off without me; better off dead; suicide plan; suicide pact; tired of living; don't want to be here; die alone; go to sleep forever".

When a tweet matching any of the above terms was identified by this tool, it was stored in this tool alongside the Twitter profile name and picture.

3.2. Human coding

Human coding was used to determine the level of concern within the suicide-related tweets, as judged from the perspective of the coding team which consisted of three mental health researchers and two computer scientists. The mental health researchers specialised in suicide prevention and possessed training in the detection of suicide risk although they are not practicing clinicians. The computer scientists had no formal or informal training on suicide prevention but are researchers with expertise in social computing. The coders were asked to conceptualise the task as the level of concern one would have when seeing such a post from within their own online social network and whether they considered the post to warrant further investigation from a friend, family member or third party. Tweets were examined individually and coded according to a classification system reiterated by the research team. In the first instance, five researchers (three mental health researchers and two computer scientists) classified a small random set of tweets ($n = 100$) using only two levels, 'concerning' and 'not concerning'. It was immediately recognised that a simple dichotomy did not provide enough variance. As a result, three levels were devised: 'strongly concerning', 'concerning', and 'safe to ignore'. Another small coding task was conducted on a random set of tweets ($n = 100$) using the same five researchers. In this instance, the instructions for the task were considered too ambiguous and allowances for any references to song lyrics, popular music videos and colloquial vernacular had not been factored in. Thus, three levels with detailed definitions and specific instructions were created:

1) *Strongly concerning*: a convincing display of serious suicidal ideation; the author conveys a serious and personal desire to complete suicide, e.g., "I want to die" or "I want to kill myself" in contrast to "I might just kill myself" or "when you call me that name, it makes me want to kill myself"; suicide risk is not conditional on some event occurring, unless that event is a clear risk factor for suicide, e.g., bullying, substance use; the risk of suicide appears imminent, e.g., "I am going to kill myself" versus "If this happens, I will kill myself"; a suicide plan and/or previous attempts are disclosed; little evidence to suggest that the tweet is flippant, e.g., tweets with "lol" or other forms of downplaying are not necessarily flippant and may still be included in this category;
2) *Possibly concerning*: the default category for all tweets; to be removed from this category, the tweet must be able to be classified as 'strongly concerning' or 'safe to ignore';
3) *Safe to ignore*: no reasonable evidence to suggest that the risk of suicide is present.

Coders were instructed to select only one of the following levels and to select the default level if in doubt. An additional two options were created: 'data known' (the Twitter account holder is known to the research team) and 'data discard' (the tweet cannot be understood; used sparingly and does not include cases where the context is simply ambiguous). It was estimated that a minimum of 2000 tweets would need to be coded for a data-driven model to be derived (Jashinsky et al., 2013). As such, the human coders completed the final coding task on a large sample of tweets ($n = 2000$) which was divided equally into two subsets ($n = 1000$). Each pair of researchers (one mental health researcher – PB or AC – and one computer scientist — SW or CP) were assigned one subset to classify: 1000 items were classified in one hour blocks (e.g., 100 tweets per hour) to avoid annotation fatigue. Disagreements were arbitrated by a third independent mental health researcher (BO). A secure CSIRO web-based interface was used to perform the human coding. This interface