Internet Interventions 2 (2015) 183–188

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

allocated a set of 1000 randomly selected tweets to a pair of human coders so that the same tweet would be judged by two coders. Tweets were presented in the manner in which they are published by Twitter, including the profile name and picture. To ensure the presentation of tweets was uniform among the coders, coding was completed using the Mozilla Firefox web-browser. Each tweet was coded only once and individual decisions were final. A progress bar indicated the completion status through the allocated quota.

3.3. Machine classification

Using the human coded data, machine learning methods were applied to develop a text classifier that could automatically distinguish tweets into the three categories of concern. The Scikit-Learn toolkit (Pedregosa et al., 2011) was used to implement the various machine learning methods. Using the toolkit, each tweet was first represented as a vector of features for use with each machine learning method. As the aim was to examine the feasibility of automatic classification, basic features of word frequencies, or unigrams, were utilised in the first instance. In this representation, all words present in the observed data set became features, resulting in a high-dimensional feature representation. A word was defined as any series of characters separated by a whitespace. A number of variants for this feature representation were then explored. The weighting Term Frequency weighted by Inverse Document Frequency (TFIDF) used in Information Retrieval (Salton and McGill, 1986) was used instead of the simple frequency. This weighting encapsulates the amount of information inherent in a word, based on a linguistic observation that, for a language such as English, words that occur in many statements often represent little meaning. For example, the words 'and' and 'of' occur frequently across tweets but do not add meaningful semantic content. In contrast, a noun or a verb may represent greater meaning but occur less often when compared to function words. In this variant of the feature space, a weighting based on the document frequency (the number of tweets containing the word) is multiplied by the frequency of the word in the tweet. The equation for this weighting is as follows:

$$\mathrm{tfidf}(\mathrm{t}) = \mathrm{freq}(\mathrm{t}) \times ln \frac{N}{|\{d \in D : t \in d\}|}.$$

In this equation, t is the word feature, N is the number of document data items (tweets), and d is a document in the document set D. In the analyses reported below, the original unweighted feature representation is referred to as "freq" and the weighted variant as "tfidf". Another feature representation variant, based on document frequency which attempted to remove words with little information, was also considered. Instead of using the TFIDF equation above, this variant removed all words that occurred above a threshold for document frequency. As such, words like 'the' and 'of' would be removed from the feature space. The threshold examined in this study used a document frequency >0.7. This variant is referred to as 'filter'.

To determine how well the derived classifiers were performing, the data set was separated into randomly selected training and testing portions. Specifically, after random shuffling, the first 90% of data points were used as the training set, and the last 10% of data points were kept as a testing set. Two machine learning algorithms for text classification were tested: Support Vector Machines (SVMs) (Joachims, 1999) and Logistic Regression (Berger et al., 1996). These methods were tested with each variant of the feature space ("freq", "tfidf", and "filter"). Using cross-validation methods, the average accuracy when the training set is divided into 10 "folds" or subsets was assessed, with each fold being used as an intermediate testing subset for the other 9 folds. Additional experiments were conducted using the best performing algorithm as determined by the 10-fold cross-validation results. These experiments were conducted on a held-out data set. Performance, in terms of total accuracy as well as the precision, recall and F1 metric for each of the categories, was examined. Precision is defined as the percentage of items correctly classified into a particular category by the algorithm. The category selected was considered correct if it coincided with the human coding. Recall indicates the percentage of the category that was successfully classified. F1 is the harmonic mean of the two and represents a balance between the two. The range of the precision, recall and the F1 metrics are all bounded between 0 and 1, of which a higher value indicates better performance. These metrics are defined as:

$$\mathrm{Precision(c)} = \frac{|\text{correctly identified items of type } c|}{|\text{Suggested items of type c}|}$$

$$\mathrm{Recall(c)} = \frac{|\text{correctly identified items of type } c|}{|\text{Actual items of type c}|}$$

$$\mathrm{F1(c)} = 2 \times \frac{\mathrm{precision}(c) \times \mathrm{recall}(c)}{\mathrm{precision}(c) + \mathrm{recall}(c)}.$$

In these equations, c indicates the category (one of 'strongly concerning', 'possibly concerning' or 'safe to ignore'). Ideally, a value close to 1 for the F1 score for each category should be observed. In the current study, the precision of the two extreme categories, 'strongly concerning' and 'safe to ignore' was of primary interest as the 'possibly concerning' category was to be used if the human coder was in any doubt. Similarly, the algorithm needed to discard tweets that were 'safe to ignore' without incorrectly discarding any concerning tweets.

3.4. Statistical analysis

Statistical analysis was conducted to evaluate the quality of the data classification, including both the human coding and the machine learning. The first set of analyses focused on the rates of agreement among the human coders. The second set of statistical analyses examined whether or not a precise model of automatic detection could be derived from the human coded data. In both analyses, χ^2 tests with an alpha level of 0.05 were used to compare differences. Cohen's κ coefficient was calculated to measure the level of agreement between the two coders for each set of tweets (Hallgren, 2012). Percentage rates of agreement were also reported. Separate analyses are conducted for data sets A and B. Where appropriate, analyses on the combined data set are reported.

3.5. Ethics

This study was approved by the University of New South Wales Human Research Ethics Committee and the CSIRO Ethics Committee. Two main issues were addressed: i) individual consent from users was not sought as the data was publically available and attempts to contact the Twitter account holder for research participation could be deemed coercive and may change user behaviour; ii) psychological support was not offered to those Twitter users who appeared to be at-risk as intervention via Twitter as it may not be appropriate. This is due to several reasons: suicide risk fluctuates; uninvited contact with participants is an invasion of privacy; such contact could lead to unsolicited attention; and most importantly, the main aim of this study was to determine whether it was possible to categorise tweets in this way, rather than to immediately assume the coding was accurate. The research team accepts that Internet research for suicide prevention is complex, and in its infancy, and believes that there is a dire need for scientifically valid data before uninvited contact with individuals is made (Mishara and Weisstub, 2013). The CSIRO ethics committee required that the data be analysed at least three months after the collection date, and both committees required that names, usernames or any other identifiable information to be excluded from any research outputs.

4. Results

4.1. Human coding

During the time of data collection, 14,701 tweets matched the suicide-related search terms: 2000 (14%) were randomly selected for human coding. Table 1 presents the data set used in the major classification task including frequency distributions and rates of agreement. A total of 9% ($n = 178$) of data was discarded or known, and thus excluded. When data sets A and B were combined, 14% ($n = 258/1822$) were coded as 'strongly concerning', 57% ($n = 1030/1822$) 'possibly concerning', 29% ($n = 534/1822$) were coded as 'safe to ignore'. There were significantly more 'strongly concerning' tweets in Set A than Set B ($\chi^2 = 22.67$, $p < .001$) as coded by the researchers. No other significant differences between category distributions were found. Overall, the mean rate of agreement for the combined data set was 74% ($n = 1341/1822$, average $\kappa = 0.55$): 79% ($n = 809/1030$) for 'possibly concerning', 74% ($n = 192/258$) for 'strongly concerning' and 64% ($n = 340/534$) agreement for 'safe to ignore'. In Set A, human coders were in agreement 68% ($\kappa = 0.47$) of the time whereas in Set B, the agreement rate was 77% ($\kappa = 0.64$). There was a significantly higher rate of agreement among the 'safe to ignore' category in Set B (72%) than in Set A (52%, $\chi^2 = 25.79$, $p < .001$).

4.2. Prediction accuracy of machine classifier

The total number of tweets used in the training and testing was 1820: Set A = 829 (training: 746, testing: 83) and Set B = 991 (training: 891, testing: 100). Outlined in Table 2 is the performance of the classifiers indicating that there was variation in performance accuracy depending on the choice of algorithm and the data set used. The best performing algorithm was the "SVMs with TFIDF no-filter", and the additional tests were performed using this. The last section of the table outlines its performance on a held-out data set. There was a gain in performance accuracy when sets A and B were combined, with an overall accuracy score of 76%. This was significantly higher than chance in which the majority class would be selected, yielding a precision score of only 56% for the combined set. All precision scores for the categories were greater than 75%: In the combined set, a precision score of 80% was found for 'strongly concerning', 76% for 'possibly concerning' and 75% for 'safe to ignore'. For strongly concerning, recall was 53%. The highest precision score was 100% for the 'safe to ignore' category in Set B; however, the recall score was very low (36%).

4.3. Learning curves of machine classifier

Fig. 1 displays the learning curves for each data set and the change in performance accuracy that resulted from the addition of successively larger amounts of the coded data (starting at 0.1 and incrementing by a 0.1). Set A and Set B exhibited a decrease in performance when 0.3 of the data was added. Performance started to increase at approximately 0.4. When the entirety of data was reached, the performance accuracy of the algorithm was yet to plateau: the final addition of 0.2 of the data corresponded to a 6% increase in performance. As outlined, a plateau was not reached indicating that the model could still be improved using more data.

Table 1
Frequencies and rates of agreement after moderation (N = 1822).

	Set A $n = 830$	Set B $n = 992$
Category frequencies		
Safe to ignore	222 (27%)	312 (31%)
Possibly concerning	456 (55%)	574 (58%)
Strongly concerning	152 (18%)	106 (11%)
Rates of agreement		
Safe to ignore	116 (52%)	224 (72%)
Possibly concerning	344 (75%)	465 (81%)
Strongly concerning	109 (71%)	83 (78%)
Total agreement	569 (68%)	772 (78%)

Table 2
Properties, classifiers and metrics of the machine-learned Twitter data.

		Set A	Set B	Combined
Properties				
	Total word count	10526	12787	23321
	Unique words	2068	2482	3676
	Average word count per tweet	12	12	12
	Average character count per tweet	74	72	73
Classifier	Feature Space Variant	%	%	%
SVM — no filter	Frequency	**56**	64	60
	TFIDF	55	**67**	**63**
SVM — filter	Frequency	**56**	64	59
	TFIDF	54	**67**	62
LGR — no filter	Frequency	55	59	55
	TFIDF	55	61	58
LGR — filter	Frequency	55	59	55
	TFIDF	**56**	61	57
Metrics: SVM TFIDF no filter algorithm		Accuracy (%)		
	Overall accuracy	67	68	**76**
Strongly concerning	Precision	**88**	62	80
	Recall	64	43	53
	F1	74	51	64
Possibly concerning	Precision	62	68	76
	Recall	97	86	91
	F1	76	76	83
Safe to ignore	Precision	75	**100**	75
	Recall	14	36	53
	F1	24	53	62

Note: SVM = Support Vector Machine method, LGR = Logistic Regression method. Bolded values indicate the best performance result for that set.

5. Discussion

The aim of this study was to determine whether the level of concern for suicide-related tweets could be distinguished using human coders and a machine-learned classifier. When coded by humans, 14% of the suicide-related tweets were deemed 'strongly concerning' which on average, suggested that up to 32 tweets per day portrayed a level of suicidality that warranted further investigation. The largest majority of tweets were coded as 'possibly concerning'; however, this was not surprising as this was the default category. Interestingly, almost one third of the tweets (29%) were considered to be 'safe to ignore', despite the use of a suicide-related term or phrase. Overall, human coders were in agreement 76% of the time and the Kappa coefficients for each set indicated moderate (0.47) to good (0.69) agreement. The machine-learned classifier correctly identified 80% of 'strongly concerning' tweets and achieved an overall agreement rate of 76%. These findings illustrate a significant advancement in our ability to reliably detect suicide risk in social media data. In earlier work, Huang et al. (2007) achieved a 14% automated identification rate of suicide risk in My Space bloggers. More recently, Jashinsky et al. (2013) reported an agreement rate of 80% ($\kappa = 0.48$) among human coders of 1000 suicide-related tweets, but this was not replicated using machine learning and only 0.13% of the collected tweets were examined. In the current study, 14% of the collected tweets were examined and the machine classifier achieved the same level of accuracy as the human coders.

Importantly though, all human coders reported that the coding task was difficult. Although the agreement rates among human coders and accuracy rates of the machine-learned classifier were satisfactory, concordance was by no means perfect. In the early stages of human coding, differences between raters' level of concern were observed, evidenced by the need to continually refine the classification scheme. It is possible

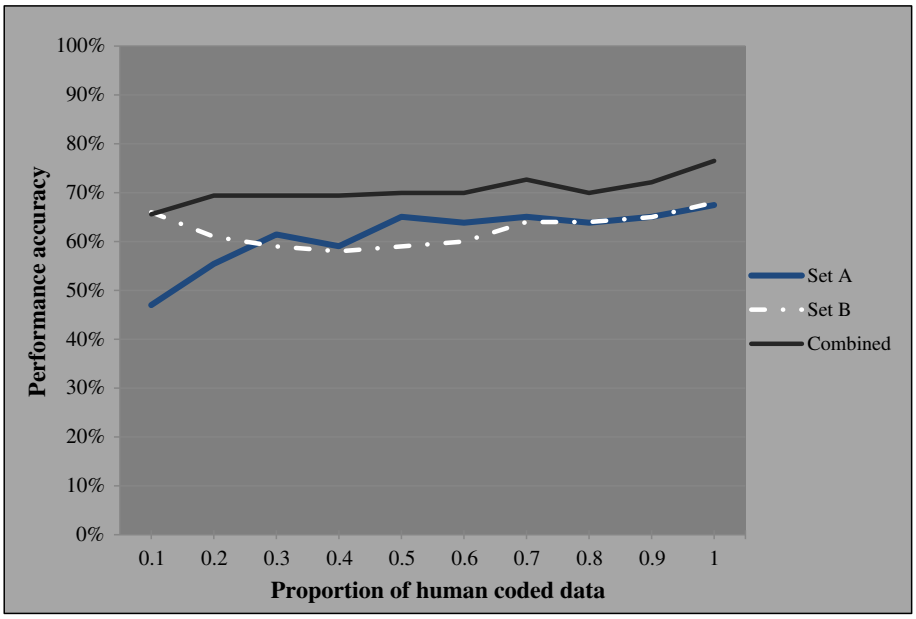

Fig. 1. The change in performance accuracy (%) of the machine classifier based on the successive additions of human coded data.

that similar contrasts in risk assessment would occur in the general user population of Twitter. It is likely that some users would express greater concern than others when exposed to a suicide-related tweet. This can be attributed to the different social networks, cultures and backgrounds of Twitter users. For example, some of the suicide-related tweets appeared to include flippant use of the phrases "I'll kill myself…" or "I want to die…" which may be a reflection of the vernaculars of different ages and cultures, rather than a direct expression of suicidality. Twitter users discuss a broad range of issues on this site and such discussion may not always be an honest indication of their emotions and behaviour. It cannot be definitively stated that all 'strongly concerning' suicide-related tweets were genuine statements of suicidality or that the suicide-related tweets collected were truly indicative of suicide. Nonetheless, these statements evoked of a strong level of concern among coders and were believed to warrant further investigation.

Unsurprisingly, it became increasingly apparent that a greater understanding of the context of the tweet, beyond the simple text expression, was essential. Context can take many forms: within the Twitter post (e.g., images, emoticons, hashtags, retweets); within the Twitter account (e.g., previous tweets, replies to tweets, followers, Twitter network) and external to Twitter (e.g., current events and emotional state of the account holder, language spoken, offline social network). It is difficult to obtain the context beyond the Twitter post automatically and without direct contact with the account holder. While obtaining this information may clarify the nature of the tweet, the volume of Twitter data makes individual contact almost impossible and as suicidal ideation fluctuates (Matsuishi et al., 2005; Prinstein et al., 2008; Williams et al., 2006), a time delay from posting to verification could be expected. It could be hypothesised that users who reply (i.e., the reply network) may know the account holder personally, allowing their reply content to contextualise the sincerity of the original tweet (Fu et al., 2013). Future analysis of the replies to the suicide-related tweets may help to clarify the level of risk.

The ethics of this type of suicide detection remain difficult to navigate (Lee, 2014). Although this study was not an intervention project, understanding the acceptability of online suicide detection among Twitter users is critical. Privacy issues are pertinent. Twitter users were not contacted in the current study, and for a tool such as this to be utilised by the public, users must consent to their tweets being monitored by an organisation or an individual, and permission to be contacted if a 'strongly concerning' tweet is detected. While it may appear that an individual expressing suicidality in a public forum such as Twitter may welcome intervention, this assumption cannot be made. Furthermore, the expectations and responsibilities of other Twitter users remain unclear. As it was beyond the scope of this study, the most appropriate action for when a 'strongly concerning' suicide-related tweet has been detected remains unknown. Even with automatic responses, crisis support services may not have the capacity to intervene in every instance: Twitter is global and many locations do not have established crisis services. Given that some may find this type of monitoring invasive and inappropriate, future research must involve consultation with Twitter users, consumers and mental health professionals before such a tool could be considered for public use (Schroeder, 2014).

6. Limitations

In order to improve the reliability and accuracy of the automatic classifier, future efforts would benefit from expanding the range of suicide-related search terms to ensure that more expressions of suicidality are included. Although a model for accurate classification could be derived from the human coded data, the analyses used to extract this model were rudimentary and primarily based on single words. Future modelling should attempt to normalise words by removing prefixes and suffixes, and by analysing adjacent sequences of words as well as the words themselves, since as rare words or expletives may represent greater risk of suicide. As stated, the current methods were unable to clearly discern, beyond a required level of accuracy, those who were experiencing passive suicidal ideation from those who were in immediate danger of taking action. The current study did not attempt to validate the risk of suicide with offline measures, such as family and friends, standardised questionnaires or clinical consultation. The Twitter data used in this study was also unable to provide sample characteristics, such as age and gender, which limits the generalisability of results. However, given the link between suicidal ideation and suicide (Large and Nielssen, 2012; Posner et al., 2011), serious consideration should be given to individuals who use Twitter to communicate thoughts of suicide. The results of this study indicate that these short messages have the capacity to capture user attention and evoke a strong level of concern. Future research may benefit from adopting a mixed-methods approach using qualitative techniques to explore how Twitter users react, both internally and externally, to such messages both immediately and over time. Future research could also attempt to

clarify genuine risk by a retrospective analysis of the Twitter content of those who are known to have died by suicide. Alternatively, conducting a prospective study in which users provide consent to having both their suicide risk and Twitter posts monitored, with appropriate procedures for adverse events, would help to better understand the nature of Twitter behaviour among those who experience suicidal ideation.

7. Implications and conclusion

The current study confirms that Twitter is used by individuals to express suicidality and that it is possible to distinguish the level of concern among suicide-related tweets, using both human coders and a machine classifier. However, the inability to determine the external context of such tweets and the unachieved plateau in the learning curves means that further work is needed to improve the reliability and validity of this method. With improvements, this method may become an additional modelling variable that is optimised for public health surveillance of suicide risk: identifying the locations or time zones that have elevated rates of 'strongly concerning' tweets, in addition to monitoring suicide contagion (Luxton et al., 2012), which can then be combined with other population measures (Won et al., 2013) to give a real-time and automated overview of suicide risk. Gaining insight into how individuals and their communities react to suicide-related tweets may help to outline new policies and practices for the prevention of suicide at the community level. Such analyses may help to form the quantitative rationale for innovative public health campaigns aiming to reduce suicide and its associated stigma. As outlined by the limitations of the current study, the predictive power for actual suicidal behaviour using Twitter is not yet known and the findings do not directly identify targets for intervention. Overall, the results of this project are encouraging and suggest that future work will yield promise for social media to identify and potentially respond to suicide risk among individuals and the community.

Funding

This project was supported in part by funding from the NSW Mental Health Commission and the NHMRC John Cade Fellowship 1056964. PJB and ALC are supported by the NHMRC Early Career Fellowships 1035262 and 1013199.

Conflict of interest

The authors declare that there is no conflict of interest.

Acknowledgements

We acknowledge the efforts of Dr David Milne, previously of the CSIRO, who helped to establish the beginnings of this project and outlined the initial ideas for search terms and classification.

References

Bailey, R.K., Patel, T.C., Avenido, J., Patel, M., Jaleel, M., Barker, N.C., Khan, J.A., Ali, S., Jabeen, S., 2011. Suicide: current trends. J. Natl. Med. Assoc. 103, 614–617.
Berger, A.L., Pietra, S.D., Pietra, V.J.D., 1996. A maximum entropy approach to natural language processing. Comput. Linguist. 22, 39–71.
Cerel, J., Jordan, J.R., Duberstein, P.R., 2008. The impact of suicide on the family. Crisis 29, 38–44.
Christensen, H., Batterham, P., O'Dea, B., 2014. E-health interventions for suicide prevention. Int. J. Environ. Res. Public Health 11, 8193–8212.
Duggan, M., Ellison, N.B., Lampe, C., Madden, M., 2015. Social Media Update 2014. Pew Research Center.
Fu, K.-W., Cheng, Q., Wong, P.W.C., Yip, P.S.F., 2013. Responses to a self-presented suicide attempt in social media: a social network analysis. Crisis 34, 406–412.
Goldsmith, K., Pellmar, T.C., Kleinman, A.M., Bunney, W.E., 2002. Reducing Suicide: A National Imperative. The National Academies Press, Washington D.C.
Hallgren, K.A., 2012. Computing inter-rater reliability for observational data: an overview and tutorial. Tutor. Quant. Methods Psychol. 8, 23–34.
Huang, Y., Goh, T., Liew, C.L., 2007. Hunting suicide notes in web 2.0 — preliminary findings. Ninth IEEE International Symposium on Multimedia. IEEE Computer Society, Los Almitos, CA.
Jashinsky, J., Burton, S.H., Hanson, C.L., West, J., Giraud-Carrier, C., Barnes, M.D., Argyle, T., 2013. Tracking suicide risk factors through Twitter in the US. Crisis 35, 51–59.
Joachims, T., 1999. Making Large-scale Support Vector Machine Learning Practical. MIT Press, Cambridge MA.
Large, M.M.B.M.F., Nielssen, O.M.P.F., 2012. Suicidal ideation and later suicide. Am. J. Psychiatr. 169, 662.
Lee, N., 2014. Trouble on the radar. Lancet 384, 1917.
Lehavot, K., Ben-Zeev, D., Neville, R.E., 2012. Ethical considerations and social media: a case of suicidal postings on facebook. J. Dual Diagn. 8, 341–346.
Levine, H., 2008. Suicide and its impact on campus. New Dir. Stud. Serv. 2008, 63–76.
Luxton, D.D., June, J.D., Fairall, J.M., 2012. Social media and suicide: a public health perspective. Am. J. Public Health 102 (Suppl. 2), S195–S200.
Matsuishi, K., Kitamura, N., Sato, M., Nagai, K., Huh, S.-Y., Ariyoshi, K., Sato, S., Mita, T., 2005. Change of suicidal ideation induced by suicide attempt. Psychiatry Clin. Neurosci. 59, 599–604.
McAuliffe, C.M., 2002. Suicidal ideation as an articulation of intent: a focus for suicide prevention? Arch. Suicide Res. 6, 325–338.
Mishara, B.L., Weisstub, D.N., 2013. Challenges in the control and regulation of suicide promotion and assistance over the Internet. In: Mishara, B.L., Kerkhof, A.J. (Eds.), Suicide Prevention and New Technologies. Palgrave Macmillan, Hampshire, UK, pp. 63–75.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.
Posner, K., Shen, S., Mann, J.J., Brown, G.K., Stanley, B., Brent, D.A., Yershova, K.V., Oquendo, M.A., Currier, G.W., Melvin, G.A., Greenhill, L., 2011. The Columbia-Suicide Severity Rating Scale: initial validity and internal consistency findings from three multisite studies with adolescents and adults. Am. J. Psychiatr. 168, 1266–1277.
Prinstein, M.J., Nock, M.K., Simon, V., Aikins, J.W., Cheah, C.S.L., Spirito, A., 2008. Longitudinal trajectories and predictors of adolescent suicidal ideation and attempts following inpatient hospitalization. J. Consult. Clin. Psychol. 76, 92–103.
Salton, G., McGill, M., 1986. Introduction to Modern Information Retrieval. McGraw Hill, New York, NY, USA.
Schroeder, R., 2014. Big Data and the brave new world of social media research. Big Data Soc. 1, 1–11.
Twitter Inc, 2014. Dealing With Self-harm and Suicide.
Wasserman, D., Tran Thi Thanh, H., Pham Thi Minh, D., Goldstein, M., Nordenskiöld, A., Wasserman, C., 2008. Suicidal process, suicidal communication and psychosocial situation of young suicide attempters in a rural Vietnamese community. World Psychiatry 7, 47–53.
Williams, J.M.G., Crane, C., Barnhofer, T., Van der Does, A.J.W., Segal, Z.V., 2006. Recurrence of suicidal ideation across depressive episodes. J. Affect. Disord. 91, 189–194.
Wolk-Wasserman, D., 1986. Suicidal communication of persons attempting suicide and responses of significant others. Acta Psychiatr. Scand. 73, 481–499.
Won, H.H., Myung, W., Song, G.Y., Lee, W.H., Kim, J.W., Carroll, B.J., Kim, D.K., 2013. Predicting national suicide numbers with social media data. PLoS One 8, e61809.
World Health Organization, 2014. Preventing Suicide: A Global Imperative. World Health Organisation, Geneva, Switzerland.