Social Science & Medicine 219 (2018) 19–29

Contents lists available at ScienceDirect

Social Science & Medicine

journal homepage: www.elsevier.com/locate/socscimed

Tracking the Werther Effect on social media: Emotional responses to prominent suicide deaths on twitter and subsequent increases in suicide

Robert A. Fahey[a], Tetsuya Matsubayashi[b], Michiko Ueda[c,*]

[a] Graduate School of Political Science, Waseda University, Building No.3 1-6-1 Nishiwaseda, Shinjuku-ku, Tokyo 169-8050, Japan
[b] Osaka School of International Public Policy, Osaka University, 1-31 Machikaneyama, Toyonaka, Osaka 560-0043, Japan
[c] Faculty of Political Science and Economics, Waseda University, Building No.3 1-6-1 Nishiwaseda, Shinjuku-ku, Tokyo 169-8050, Japan

ARTICLE INFO

Keywords:
Suicide
Twitter
Social media
Celebrity suicide
Werther effect
Imitation
Japan
Media

ABSTRACT

Rises in suicide rates following media reports of the deaths by suicide of public figures are a well-documented phenomenon. However, it remains unclear why, or by what exact mechanism, celebrity suicides act to increase suicidal risk in the wider public due to the lack of data showing how the public understands and reacts to the suicide of well-known figures. This study used a supervised machine learning approach to investigate the emotional content of almost 1 million messages sent on Twitter related to the suicides of 18 prominent individuals in Japan between 2010 and 2014. The results revealed that different demographic characteristics of the deceased person (age, gender, and occupation) resulted in significant differences in emotional response; notably that the suicides of younger people, of women and of people in entertainment careers created more emotional responses (measured as a ratio of emotionally-coded tweets within the overall volume of tweets for each case) than for older people, men, and those in other careers. Moreover, certain types of emotional response were shown to correlate to subsequent rises in the national suicide counts, with "surprised" reactions being positively correlated with the suicide counts, while a high proportion of polite messages of condolence were negatively correlated. The study demonstrates the importance of, and describes a methodology for, considering the content of social media messages, not just their volume, in research into the mechanism by which these widely-reported deaths increase suicide risk in the broader public.

1. Introduction

Research into causation and prevention of suicide deaths has identified a "copycat" phenomenon, dubbed the "Werther Effect" (Phillips, 1974), whereby media reporting of suicides by celebrities and well-known figures leads to an increase in suicide deaths in the general population. This effect has been studied and re-confirmed across multiple time periods and geographic regions in recent decades (Niederkrotenthaler et al., 2009; Stack, 1987; Ueda et al., 2014; Wasserman, 1984).

However, it remains unclear why, or by what exact mechanism, celebrity suicides act to increase suicidal risk in the wider public. A major obstacle to advancing that understanding has been the lack of data showing how the public understands and reacts to the suicide of well-known figures. Past studies (Niederkrotenthaler et al., 2009; Pirkis et al., 2006; Ueda et al., 2014) have attempted to capture public reactions to celebrity deaths by using the celebrity's status or the level of media coverage their death attracted as indicators, but these are rough measures at best and do not give a clear insight into how the public processes these events.

A promising new strand of research aims to overcome this challenge by using data from popular social media platforms (Dillman Carpentier and Parrott, 2016; Ueda et al., 2017). With the rapid rise in the popularity and societal impact of platforms like Facebook and Twitter, researchers have also begun to consider the potential impact these platforms may have on suicide (Colombo et al., 2016; Daine et al., 2013; Jashinsky et al., 2014). As a key source of both information and social interaction for many users, social media both mediates access to information about celebrity suicides and provides researchers with new sources of data on the public reaction to such events. As such, research into the propagation of and reaction to suicide-related news on social media can help us to understand the functioning of the Werther Effect and provide insights that allow us to devise new reporting guidelines or other interventions to reduce the potency of this effect.

This study builds upon existing work in relation to social media and the Werther Effect, particularly the confirmation that the degree of

* Corresponding author. Faculty of Political Science and Economics, Waseda University; Building No. 3, 1-6-1 Nishiwaseda, Shinjuku-ku, Tokyo 169-8050, Japan.
E-mail addresses: robfahey@fuji.waseda.jp (R.A. Fahey), matsubayashi@osipp.osaka-u.ac.jp (T. Matsubayashi), mueda@waseda.jp (M. Ueda).

https://doi.org/10.1016/j.socscimed.2018.10.004
Received 9 July 2018; Received in revised form 21 September 2018; Accepted 11 October 2018
Available online 12 October 2018

public reaction to a widely reported suicide can predict the extent to which suicide deaths increase in the wake of the event (Ueda et al., 2017). In that study, the reaction to suicides by 26 prominent individuals (identified as such by the reporting of their death in a national newspaper) in Japan between 2010 and 2014 was measured using the volume of posts about each death on Twitter, a popular micro-blogging site which is Japan's most widely used social network. It found that the volume of social media posts about a death was a much better indicator of the impact it would have on the actual suicide rate than the extent of newspaper or television coverage about a death. The present study replicates the data set employed by that paper (starting from the same 26 widely reported deaths, and the Twitter reactions to same) in order to investigate the actual content of the tweets send in relation to each death, categorised according to the types of emotion expressed in the text. This allows us to examine the relationship between different broad categories of celebrity or famous figure (gender, age and occupation) and the different types of emotional reaction expressed on social media, thus providing a significant new data point to literature attempting to refine the understanding of the Werther Effect through creating a taxonomy of famous figures (Stack, 1987). Moreover, studying and categorising the emotional content of tweets has the potential to improve the precision of the previous study's finding on the capacity of Twitter post volumes to prefigure suicide rate changes.

While Twitter posts are brief in nature (140 characters during the time period studied in this paper; increased to 280 characters in 2017), previous studies have shown significant success in identifying sentiments related to depression both on a micro (individual) level (Jashinsky et al., 2014; O'Dea et al., 2015) and on a macro (group) level (Cavazos-Rehg et al., 2016; Karamshuk et al., 2017; Woo et al., 2015). Our study employed a supervised machine learning approach to categorise the emotional content of c.950,000 tweets related to 18 high-response celebrity deaths. Using the resulting data, we were able to explore (1) which features of celebrity deaths were related to stronger or weaker emotional responses among social media users; and (2) which types of emotional response were associated with changes in population suicides.

This study makes three key contributions. Firstly, it moves the analysis of social media response to celebrity deaths beyond simply measuring volume and demonstrates a mechanism for studying the emotions being expressed by users instead. Secondly, it demonstrates a connection between known demographic features of the deceased person and the nature of the emotional response observed on social media. Thirdly, it shows a correlation between certain emotional responses and subsequent increases in the actual suicide counts – most notably that “surprised” reactions are strongly positively correlated with suicide count increases. This new evidence regarding the mechanism of the Werther Effect helps to fill in a gap in existing knowledge on this phenomenon.

2. Data and methods

2.1. Data

This study employed three data sets: the list of prominent suicide deaths, the archive of historical Twitter posts related to those deaths, and national records of deaths by suicide. We started from the list of 26 prominent figures whose suicides had been reported in the Japanese media between 2010 and 2014 that was originally compiled by Ueda et al. (2017) and accessed the Twitter data for each suicide through the historical archive of tweets maintained by data warehousing service Crimson Hexagon. Data were collected using the full name of the deceased person as a keyword along with, where appropriate, their common stage name or their common title (e.g. “Representative” + surname in the case of a political figure). Seven days of data prior to the reporting of the individual's suicide were gathered, along with 15 days after the initial report. It was important to gather some data from the period prior to the death report, as in several cases social media traffic around the individual began to peak ahead of the publication of mainstream media stories on the suicide, indicating that news had been distributed through other channels first; furthermore, a small number of cases had a multiple day lag between the reporting of an individual's suicide attempt and the reporting of their actual death.

Next, the Twitter data were filtered to ensure that our keyword searches had not picked up irrelevant material. Posts in languages other than Japanese were removed from the data set; although these were relatively few, the keyword searches for some Japanese celebrities who used English-language stage names had gathered several thousand irrelevant English-language tweets. After removing these tweets, we were left with 974,891 tweets across the 26 events. Next, we adjusted the tweet volume figure for each event to allow for the significant underlying growth in Twitter usage over the period under examination. This period (2010–2014) saw rapid growth in the usage of Twitter in Japan, meaning that growth in the number of users and frequency of usage of the platform which might bias our sample selection towards events later in the period. To avoid such bias, each event's tweet count was normalised using a measure of underlying platform growth (relative to the mid-point of the data in January 2013) calculated using a monthly mean of daily totals of tweets sent in Japan from the start of 2011 to the end of 2014 (provided to us by Crimson Hexagon). Since normalisation data was not available for 2010, events in that year were adjusted with the January 2011 value - given some anecdotal evidence suggesting that Twitter's growth in Japan was especially rapid after the Great East Japan Earthquake of March 2011, we felt that assuming a continuation of the trend line back into 2010 risked over-estimating events in that year, and opted instead for the most conservative estimation.

Using these adjusted Tweet volume figures, we proceeded to divide the sample into “high-response” and “low-response” groups. We adapted the methodology employed in Ueda et al. (2017), which had set the cut-off point between the two levels at 10,000 tweets over 21 days (including pre-reporting days). As our volume adjustment had increased the Tweet counts for several events, we chose a slightly stricter definition - events which had more than 10,000 tweets (adjusted) in the 14 days after news of the suicide was released were considered to be “high-response”, while those which did not meet this benchmark were treated as “low-response” events. Using this approach resulted in a data set with 18 high-response events and 8 low-response events, as against 14 high-response and 12 low-response in the study by Ueda et al. (2017). The 18 high-response events formed the basis of our investigation and a summary of the data related to them can be found in Table 1.

The national records of death by suicide used in this study were obtained from death records in the Vital Statistics Report compiled by Japan's Ministry of Health, Labour and Welfare. The Vital Statistics data used in this study were collected for administrative purposes and were anonymised and de-identified by the Ministry prior to analysis. Individual death records between July 2010 and December 2014 were made available for research purposes by approval of the Ministry. The data in the Vital Statistics Report are based on death certificates issued by physicians and subsequently reported to the local government by family members of the deceased. Deaths coded as X60 to X84 under the ICD10 standard were classified as deaths by suicide and were thus included in this study's dataset.

2.2. Method

2.2.1. Classifying emotions in tweets

We designed a content analysis codebook with five categories - *surprised*, *sadness*, *condolences*, *anger* and *no_emotion*. These categories are briefly explained in Table 2. The codebook was devised based on the results of a test-coding experiment in which human coders were asked to categorise tweets using a broader set of emotional categories drawn from prior literature on emotional categorisation of texts (Plutchik,

Table 1
Summary of case data.

Profession	Age	Gender	Year of Death	Tweets (Raw)	Tweets (Adjusted)
Entertainment	34	Female	2010	6376	25504
Critic	66	Male	2010	6471	25884
Critic	56	Male	2010	4755	19020
Critic	52	Male	2010	20152	80608
Entertainment	42	Female	2010	9085	36340
Entertainment	44	Male	2011	10162	28737
Entertainment	24	Female	2011	168104	435873
Business	61	Male	2011	6874	17823
Entertainment	45	Male	2011	57689	128620
Critic	79	Male	2011	15646	34785
Athlete	42	Male	2011	34653	77044
Business	64	Male	2011	15205	29588
Politics	73	Male	2012	74259	82644
Entertainment	78	Male	2012	39504	35179
Entertainment	34	Male	2013	26194	20936
Entertainment	62	Female	2013	135202	108063
Science	52	Male	2014	195696	119634
Entertainment	35	Male	2014	122260	73304
Total Tweets				*948287*	

Note: "Actual" tweet counts are the number of tweets collected from the Crimson Hexagon service related to this person's death. "Adjusted" tweets are normalised to account for the growth in overall usage of Twitter during the timeframe of this study.

1980; Sykora et al., 2013). We initially used the emotional ontology from Sykora et al., (2013) (*anger, confusion, disgust, fear, happiness, sadness, shame, surprise*), and used the experimental test-coding to identify emotions that did not appear often enough to be used reliably in analysis (e.g. *happiness*, *disgust*, *confusion* etc.). Feedback from the coders also prompted us to divide "condolences" from the more general category of "sadness".

In order to classify the emotional content of the full set of tweets, we employed a supervised machine learning approach – first creating a training set of data by using human coders to classify the dominant emotion of a random sample of the data, and then using these to train a machine learning algorithm which would classify the remainder of the tweets. To create the training set, four native Japanese speakers were employed to categorise a randomly selected sample of 10,000 tweets drawn from the full data set according to the codebook shown in Table 2. To ensure inter-coder reliability (ICR), each of the four coders completed an overlapping set of 1000 tweets drawn from the 10,000-tweet training corpus; the remaining 9000 tweets were each coded by a single coder. Tweets from the ICR set appeared at random during their work on the overall set. The coded data was adjusted slightly to improve inter-coder reliability; in line with the recommendations in Krippendorff (2006), p.430, unreliable multi-coded categories were merged and the work of a single coder with poor agreement scores was dropped, leaving us with 7741 coded tweets. The final Krippendorff's alpha score was 0.714, above the suggested cut-off point of 0.667 for this measurement. Further details of this process can be found in the Technical Appendix.

2.2.2. Categorising and representing the full data-set

We next trained a machine learning algorithm to categorise unseen tweets. For this purpose, the Twitter text data needed to be tokenised - i.e. divided into individual lexical tokens. This was done using the MeCab tokenising software (version 0.98pre1) (Kudo, 2016), supplemented with the *mecab-ipadic-neologd* dictionary of Japanese language neologisms (Sato et al., 2017). The *ja_tokeniser* Python package (Fahey, 2017) was used to identify and tokenise text features unique to social media data, such as "emoji" symbols, "kaomoji" (expressive faces made up from punctuation symbols), hashtags and username mentions.

Using the *scikit-learn* package (version 0.19.1, Pedregosa et al., 2011) running on Python 3.5.1, a number of different machine learning models were trained and tested against each other in order to select the most effective one. Each algorithm was trained and tested repeatedly on randomly selected sub-sets of the data – in each pass the algorithm would be trained on one sub-set and then tested against the held-out set, i.e. the unseen remaining data. The algorithms tested were two variants of Naive Bayes (Multinomial and Bernoulli), a standard and linear Support Vector Machine, Stochastic Gradient Descent, a Random Forest with ten decision trees, and a Neural Network (multi-layer Perceptron). The most effective algorithm was the Linear Support Vector Machine, which after training had an accuracy rating of 0.805, indicating over 80% agreement with human coders' decisions – similar to the level of agreement between individual coders themselves. Finally, this model was applied to the full set of tweets (n = 974,891), yielding a single classification from our five-category codebook for each tweet in the dataset.

The data were then aggregated in different ways to view the different emotional composition of Twitter users' reactions to different events or types of events. We first graphed the full data set by emotional category (ignoring the different volumes attributable to different events), allowing us to observe the relative frequency of different emotional responses over time in the full data set. For all subsequent analyses, we controlled for the different response volumes by normalising the events. To prevent minor changes in small events from skewing these graphs and narrow the focus to events shown by Ueda et al. (2017) to be relevant to the suicide rate, we included only large-response events. Graphs of the data were generated showing the different response patterns for every large response-event, and the underlying factors driving different types of response were investigated by aggregating response patterns according to three key demographic factors of the deceased person – gender, occupation (defined as "Entertainer" and "Other" following the finding that the deaths of entertainers generate a stronger Werther Effect impact (Stack, 2011)) and age (defined as above and below 50).

2.2.3. Linking tweet emotions and suicides

Our final analysis step examined which types of emotional responses found in Tweets were associated with changes in the national suicide counts. The effect of the overall volume of posts, which was the relationship already confirmed in Ueda et al. (2017), was controlled by our selection of only high-response events. This allowed us to control for variation between response rates to our 18 target events by using the ratio of tweets of different categories rather than the overall count of tweets in each category. We calculated the proportion of the number of tweets in each emotional category to the total number of tweets in

Table 2
Content analysis codebook.

surprised	Expressions of surprise or shock; emotional response to the unexpected nature of the person's suicide.
sadness	Expressions of personal sadness; implications that the news of the suicide has made the tweet's author feel sad or depressed.
condolences	Polite expressions of condolence or prayers for the deceased; references to the sadness of the deceased's loved ones, but not to personal sadness on the part of the author.
anger	Expressions of anger directed either at the deceased (e.g. condemnation of suicide as "selfish") or at another person/society in general for their perceived role in the suicide.
no_emotion	Tweets which showed no notable emotional content, including sharing media reports, factual statements about the death or dispassionate discussion.

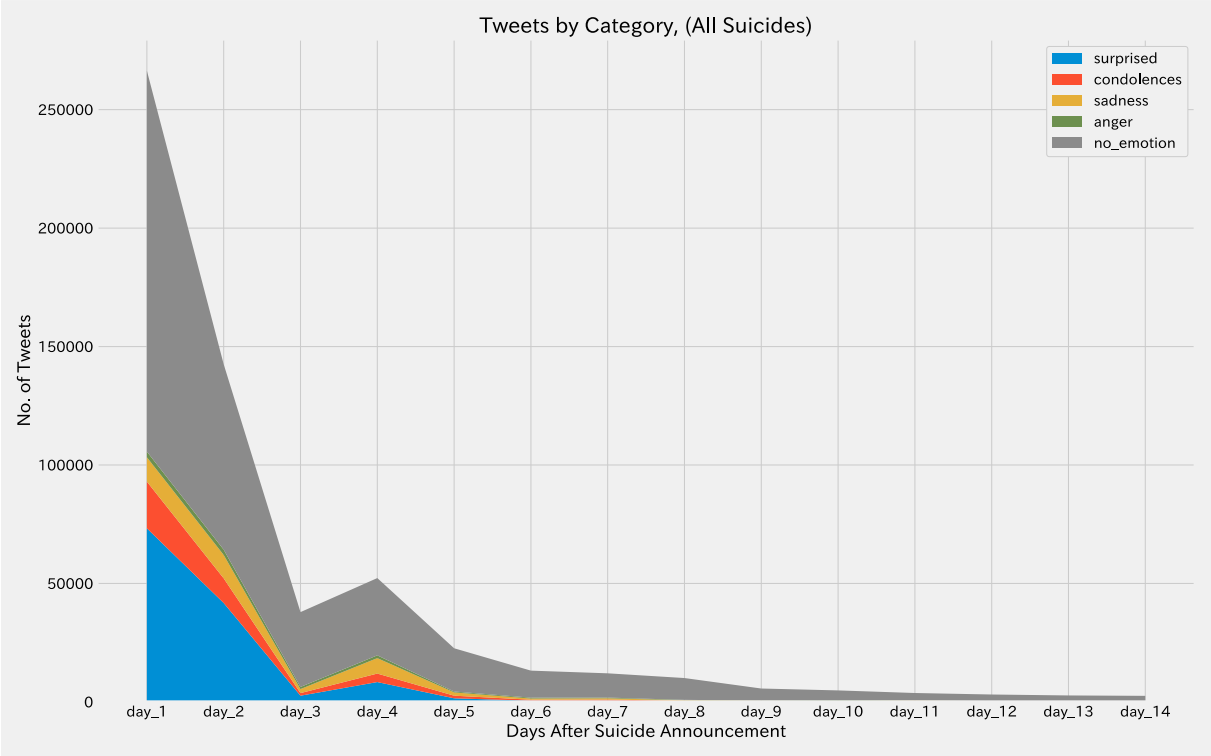

Fig. 1. Tweets by emotional category, all events aggregated.

the post-reporting period; similarly, we calculated the total number of suicides during the same period. Focusing on the 18 celebrities whose deaths had led to greater than 10,000 tweets in the 14-day post-reporting period, we then examined the relationship between these two variables using a bivariate scatter plot and a multiple-regression model. In our regression analysis, we included the total count of tweets, the career type of the celebrity (i.e. whether they were an Entertainer or another class of celebrity), and the gender and age of the celebrity as control variables. This analysis was conducted using Stata/MP software (version 15; StataCorp, College Station, TX).

3. Results

The most straightforward result from the analysis is shown in Fig. 1, which aggregates the tweets for all 26 suicides (18 high-response and 8 low-response events) and plots the emotional categories of those tweets on a time series. This figure shows a secondary peak on Day 4 of the data set; this is entirely caused by a double-peaked response pattern for a single case, in which a celebrity's suicide attempt was widely reported a few days prior to confirmation of his death in hospital. While setting its starting point of this at the reporting of the suicide attempt risks conflating responses to suicide attempts to responses to suicide deaths, in light of both a large proportion of the emotional response to the case coming after the reported attempt and the case itself being one of the most high-profile celebrity suicide deaths during the period, we chose to retain the case in our data set and to maximise the proportion of emotional tweets available for analysis in this case. To confirm the validity of this approach, we later reproduced our full analysis without this case, and saw no substantive difference in any results.

Fig. 2 controls for the variation in response volumes seen for different events by normalising the emotional content for each event between 0.0 and 1.0, restricting the data set to the 18 high-response events in order to prevent undue bias arising from small changes in the absolute numbers of low-response events. Both figures demonstrate that the Twitter data is dominated by "*no_emotion*" tweets – tweets which showed no clear emotional content, including sharing of newspaper articles or non-emotive discussion or reporting of a suicide death.

Fig. 1 also reveals the rapid drop-off in discussion after the first few days post-report; most suicides fell out of discussion almost entirely by the ninth day following the death, and emotional tweets were almost entirely gone from the data by the sixth day. By presenting the data proportionally, Fig. 2 allows us to see that this drop-off in emotional tweets was not uniform across emotional categories. While the number of *surprised* tweets drops off rapidly in the first few days post-reporting (as we would expect, since most users will be aware of the news by then), the drop-off in *condolences* tweets is much slower, and the proportion of *sadness* and *anger* tweets actually increases in the days after the suicide report. The implication is that, in general (bearing in mind that this is an aggregate of 18 different events), surprised reactions and messages of condolence are the immediate reaction to suicide reports, with emotions like sadness and anger becoming more prevalent in subsequent days.

The right-hand side of Fig. 2 suggests an upwelling in emotional tweets in the second week after a death is reported. However, taken in context with the graph of absolute tweet volumes in Fig. 1, it can be seen that this change is being caused by very small numbers of tweets and is not relevant to our study. In fact, emotional response approached zero by day six in the time series; therefore, all subsequent figures exclude bias resulting from small absolute changes in the low-volume later days by displaying only the six days post-report.

Fig. 3 shows the emotional response patterns for six days post-report for the 18 high-response events. This reveals a wide variety of different patterns; while certain basic features remain consistent throughout every event in the data – *no_emotion* is the dominant category for almost all events and *surprised* almost always peaks at the beginning before dropping off – there are significant differences between individual events. One group of events is dominated by *no_emotion* tweets right from the outset, implying that there was almost no emotional reaction from Twitter users despite interest in the suicide being high enough to create a high volume of tweets. Other deaths show a high volume of *surprised* tweets, often combined with a large number of *condolences* tweets, but almost no emotional tweets in the *sadness* or *anger*

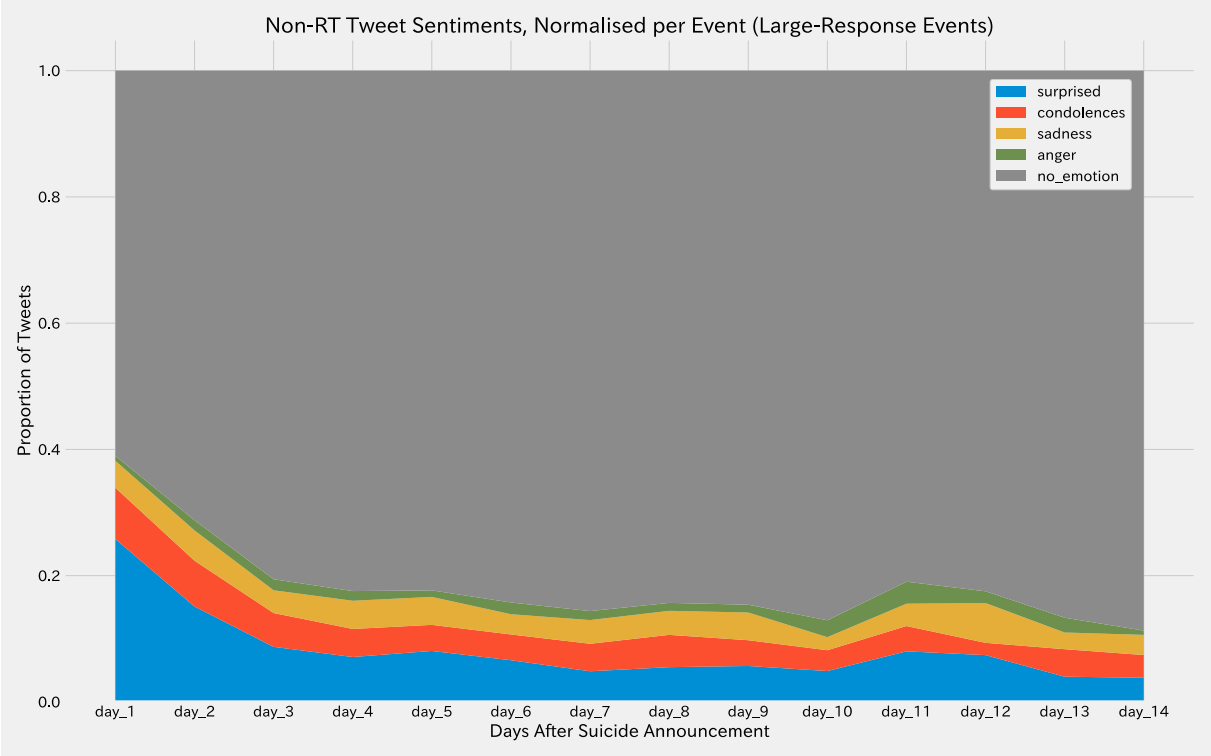

Fig. 2. Tweets by emotional category (proportional), "large" events (n = 18).

categories; while yet others have thick bands of *sadness* tweets which often peak in size a few days after the death.

One major outlier is Male/45/Entertainment (fourth from the top in the right-hand column of Fig. 3), whose emotional response graph has the largest proportion of *sadness* tweets of any case in the data set, peaking several days into the time series. This case is the one discussed above in which the suicide attempt and hospitalisation of a popular and well-known musician was reported in the press (the start of our time series) several days before confirmation of his death (the peak in emotional data).

Next, we re-aggregated the data from each case according to three primary factors related to the deceased person – their gender, their age, and their occupation – and generated graphs showing the aggregated emotional responses according to each category. Gender was divided into male (n = 14) and female (n = 4), while age was divided into those aged over 50 (n = 10) and those aged 49 and under (n = 8). For occupation, there were several different types of occupation represented in our data set, including entertainers, critics (i.e. professional reviewers and commentators), politicians, businesspeople, a scientist and a former athlete. However, as previous studies have shown a clear link between the intensity of the Werther Effect and the deceased being a celebrity entertainer (Stack, 2011), we chose to group the occupations as “entertainer” (n = 9) and “other” (n = 9).

Fig. 4 shows the emotional response patterns for male and female suicides respectively, while the patterns of emotional response for different age groups are shown in Fig. 5. Finally, Fig. 6 shows the division of emotional response patterns between “entertainer” and “other” occupations. Some difference in the response pattern can be seen between the two categories in each case. Female suicides initially provoke a higher rate of surprise than male suicides (17.0% of responses on day one, compared to 14.57% for men), though this evens out slightly over the six-day period – surprise comprises 13.18% of responses to female suicides over the measurement period, and 11.54% of responses to male suicides. Female suicides are also associated with a notably higher degree of anger – 2.45% of responses, compared to 1.14% for men. The emotional response to suicides of older people is depressed overall compared to that for younger people – *condolences* messages are consistent between the two groups (5.0% for older people, 6.27% for younger people), but both surprise (8.50% for older people, 16.15% for younger people) and sadness (1.80% for older people, 7.04% for younger) are significantly lower in responses to older people. A similar, though less distinct, pattern of response can also be viewed for the career-based Entertainer/Other categorisation; surprise (14.84% for entertainers, 8.97% for others) and sadness (5.80% for entertainers, 2.46% for others) are more elevated in responses to the deaths of entertainers, while condolence messages (entertainers 5.69%, others 5.43%) remain mostly consistent across both groups.

3.1. Tweet emotional responses and suicides

The final step in our study was to analyse the relationship between these broad emotional categories and the actual national suicide counts. Fig. 7 shows the bivariate scatter plots between the percent of tweets in each category for each celebrity and the suicide counts during the 14-day window after each celebrity's death. These scatter plots are based on the 18 high-response events. Among the four emotional categories studied, *surprised* had a strong positive correlation with the suicide counts, while positive correlations were also seen for the percentages of *sadness* and *anger* tweets, although these were significantly smaller. The percentage of *condolences* tweets showed no relationship with the suicide counts in this analysis.

We next regressed suicide counts during the 14-day window after each celebrity's death on all of the four emotional categories, allowing us to isolate the influence of each category. The results of this analysis, which also included a log of total number of tweets on each celebrity, age, and dummy variables for gender and occupation category, can be viewed in Table 3. The percent of *surprised* tweets was shown to have a positive relationship [Coef. 12.67, 95% CI: 7.94, 17.40] with the suicide counts. The coefficient suggests that the suicide count increases by about 13 as the percent of *surprised* tweets increases by one percentage point. On the other hand, the percent of condolence tweets was shown to have a negative correlation with the suicide counts [Coef. −14.28,

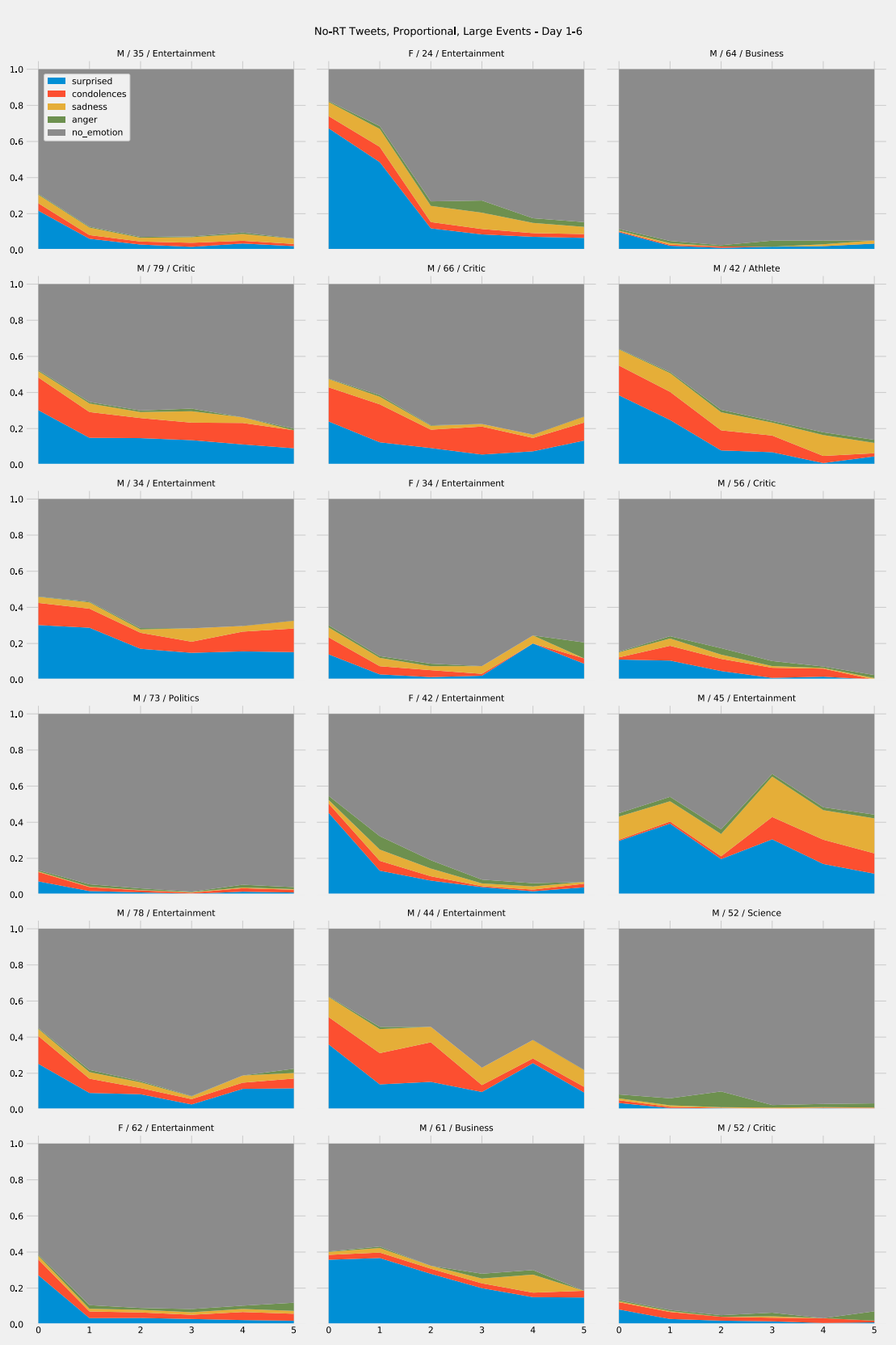

Fig. 3. Emotional response patterns (proportional), "large" events.

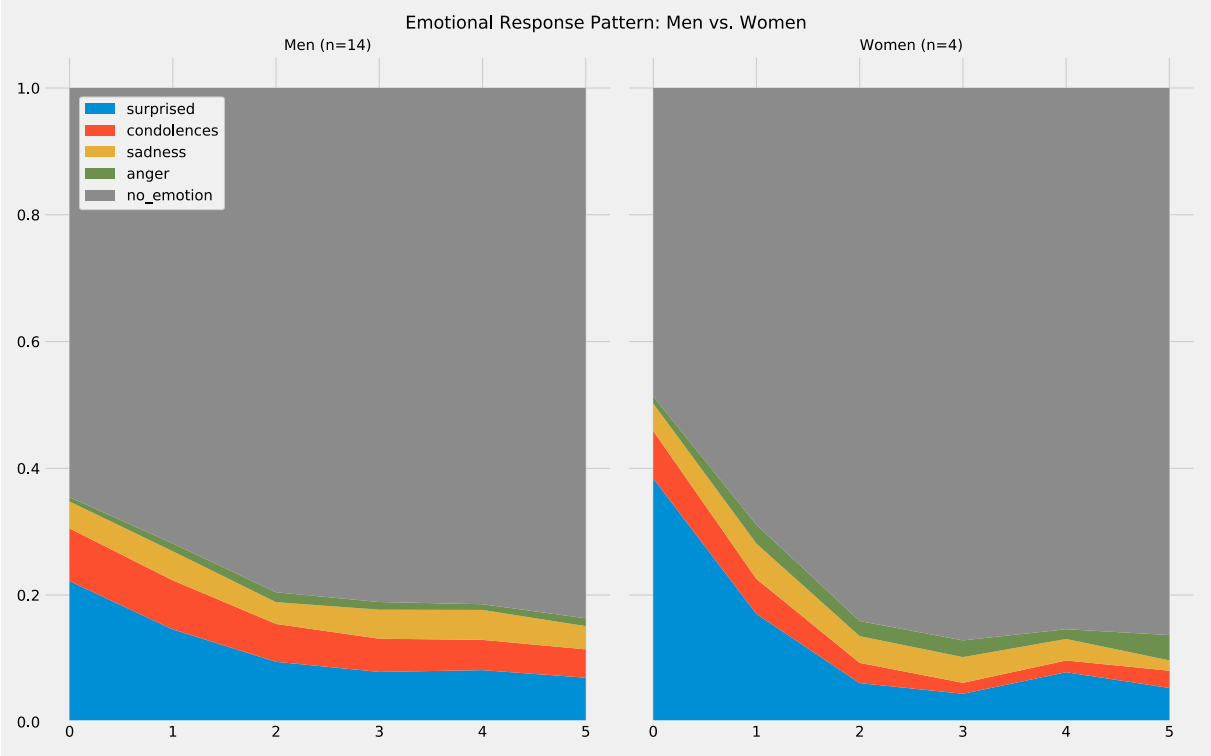

Fig. 4. Emotional response pattern - by gender of the deceased.

95% CI: 24.28, −4.29]. The percentages of *anger* and *sadness* tweets, along with the other control variables, were not shown to be significant in this analysis.

4. Discussion

The primary aim of this study was to understand how social network users respond to reporting of celebrity suicides, and how those responses drive or moderate the functioning of the Werther Effect. Since Ueda et al. (2017) had already established that the volume of posts on

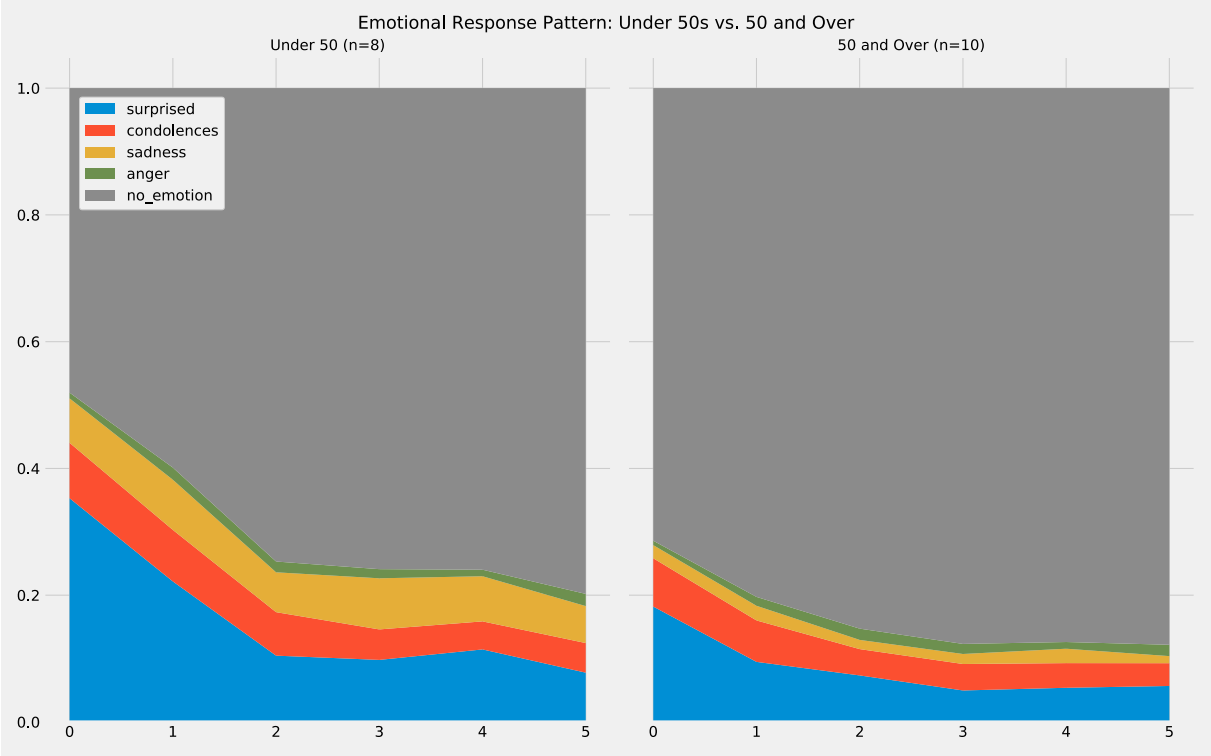

Fig. 5. Emotional response pattern - by age of the deceased.

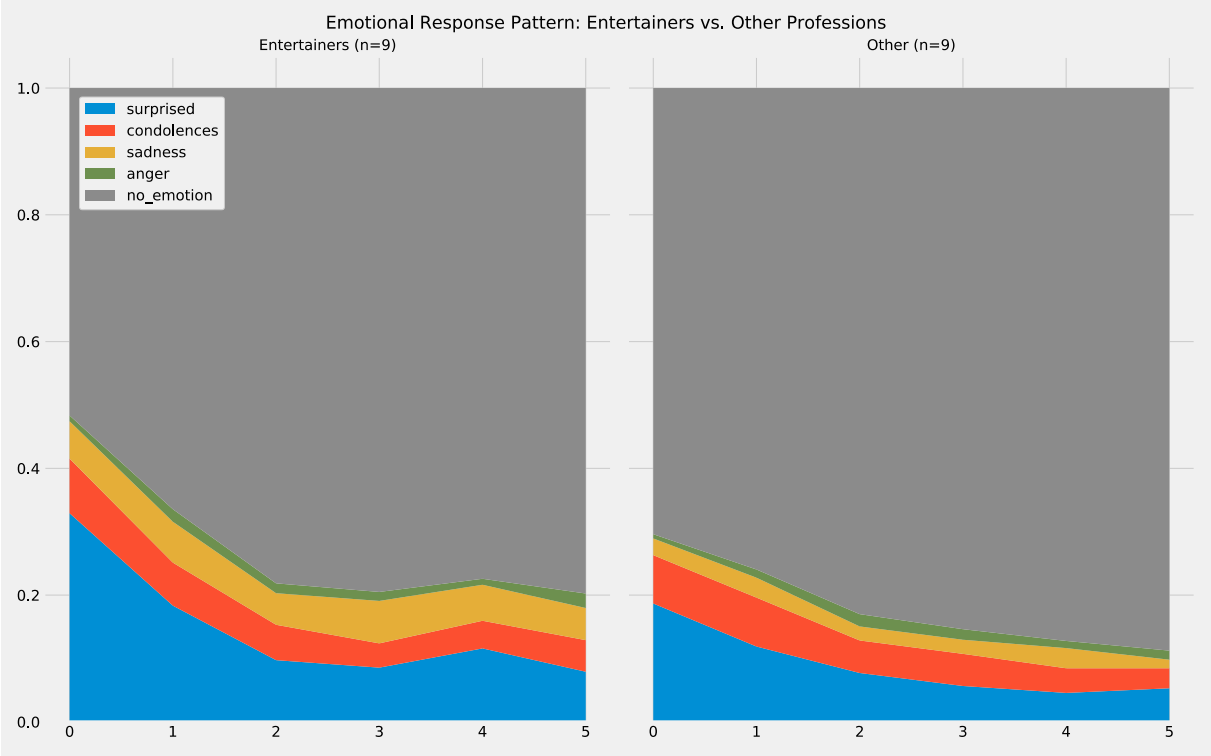

Fig. 6. Emotional response pattern - by occupation of the deceased.

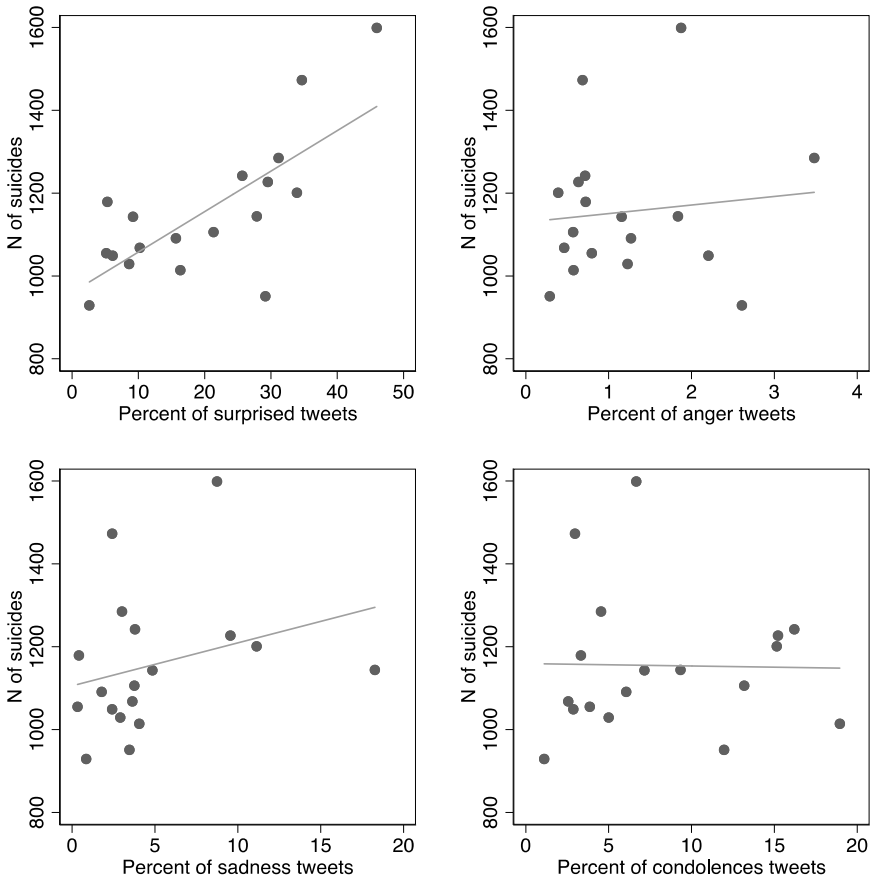

Fig. 7. Scatterplots of emotional categories to suicide counts.

Table 3
The relationship between four emotional categories and suicide counts.

	Coef	95%CI
Percent of surprised tweets	12.67	[7.94,17.40]
Percent of anger tweets	−58.72	[-135.85,18.42]
Percent of sadness tweets	3.57	[-17.36,24.50]
Percent of condolence	−14.28	[-24.28,-4.29]
Log of total number of tweets	34.13	[-48.01,116.27]
Entertainer	−106.53	[-236.40,23.35]
Female	163.97	[-134.52,462.45]
Age	2.77	[-3.63,9.18]
Constant	594.99	[-458.03,1648.01]
N of observations	18	
Adjusted R^2	0.65	

Note: This table reports regression coefficients and the 95% confidence intervals in square blankets. The total number of suicides during the 14-day window is the outcome variable.

Twitter is related to the change in suicide rates following a celebrity's death – much more strongly so than the volume of media reporting in newspapers or on TV – our study effectively serves to "lift the lid" on those high-volume cases and look more closely at the types of responses seen in different cases, giving a clearer view of the mechanism through which dissemination of news about and discussion of a celebrity's suicide over social media can serve to increase the risk of suicide among some individuals or groups.

While the actual mechanism of the Werther Effect remains unknown (although competing hypotheses such as the social learning hypothesis and the grief hypothesis have been proposed), our results provide two major findings which help to fill in this gap in existing knowledge about the impact of high profile suicides. First, we were able to show how different characteristics of the celebrity death create radically different patterns of emotional response among Twitter users. Next, we were able to uncover the connection between those emotional responses and the actual suicide counts – the end-point of the Werther Effect mechanism.

Our first set of analyses showed that the majority of discussion of celebrity suicides on Twitter is among users who are not strongly emotionally invested in, or affected by, the event, as indicated by the high volume of *no_emotion* tweets in Figs. 1 and 2. However, we also observed a large volume of emotional responses to prominent suicide deaths, especially in the first few days after the announcement of their deaths. Among the tweets categorised as containing emotional responses, those in the *surprised* category constituted a vast majority, followed by those in the *condolences* and *sadness* categories.

We also found that public responses to celebrity suicide vary greatly by the characteristics of the deceased. We found that Twitter users were more likely to exhibit emotional reactions if the deceased was female, younger than 50 years old, and an entertainer. Our analysis also showed that suicides by female prominent figures tend to generate more *surprised* and *angry* reactions by social media users. The higher degree of surprise seen for female suicides is likely be explained by the relatively small number of female suicides in the sample (n = 4, vs. n = 14 for male suicides), reflecting both the significantly lower suicide rate among women (the ratio of male to female suicides in Japan was 2.7 in 2009 (Värnik, 2012)) and lack of women in high-profile public roles outside of entertainment (e.g. politics or business) in Japan; the lower rate of female suicide makes these events more uncommon and surprising.

As for the more pronounced level of *angry* tweets for female suicides, it may reflect particular circumstances surrounding the death of individuals in our sample. We read a random sample of the tweets in this category and concluded that it comprises a mixture of anger at third parties who are blamed for the person's suicide, and anger at the deceased themselves – this was particularly notable in a single case where the deceased was the mother of a young child.

With regard to age grouping, suicides by people over 50 provoked a lower emotional response from Twitter users than suicides of younger people (Fig. 5); a very similar, albeit less pronounced, pattern of differences was seen in our aggregation by occupation (Fig. 6). We also noted that *sadness* tweets actually rise proportionally over time for both entertainers and for people under 50, whereas they remain consistent at a low level for non-entertainers and people over 50. We note that there is some debate around when copycat suicides attributed to the Werther Effect actually peak; some studies suggest a peak around three days after a high-profile suicide is reported followed by a gradual decline (Pirkis et al., 2006), which may fit this observation in our data, while other studies have found an immediate onset and sustained effect over 10 days post-reporting (Ueda et al., 2014).

Considering these aggregated demographic categories in context with the individual cases themselves (Fig. 3) lends support to interpreting these categories as intersectional; the death of celebrities belonging to multiple "high emotional response" categories created even stronger responses, while those belonging to multiple "low emotional response" groups largely provoked non-emotive discussion or reporting of the death. There are clear examples of this in the individual data; the single largest volume event in the data set, which also had by far the highest emotional peak (over 80% of tweets sent on the first day were categorised as emotional), was of a young female entertainer (F/24/Entertainment), while by far the lowest level of emotional response (despite a high volume of tweets, almost all of which were *no_emotion*) was for an older male scientist (M/52/Science).

These differences in emotional response occurred independently of the actual volume of tweets – some cases had a very high volume of tweets but almost no discernible emotional response. This confirmed our expectation that looking at the volume of tweets alone could not yield a satisfactory picture of the Werther Effect's functioning on social media. Moreover, these category analyses suggest that much of the emotional response pattern to a high-profile suicide can be predicted by the demographic characteristics of the deceased. There will of course be exceptions – the unexpected suicide of a very high profile, much-beloved celebrity would likely provoke a strong reaction regardless of their membership of any of the above classes – but these three features provide coherent explanations for the overall emotional trends of the majority of suicides in our data set. Moreover, the discovery of these different emotional patterns related to different characteristics of the deceased fits with existing findings regarding the impact of different types of suicide on copycat suicide rates – e.g. the aforementioned discovery of a stronger effect when the deceased was a celebrity entertainer (Stack, 2011). This strongly suggests that the emotional response patterns we uncovered are related to the functioning of the Werther Effect, and can be used to measure or predict the impact of a high-profile suicide on the broader suicide rate. As our analysis of the content of the tweets revealed differences in emotional response which we were able to link with the abovementioned specific factors about the deceased celebrity, we conclude that the emotional responses we were observing constitute a step in the mechanism.

The next step in the mechanism is the connection between those emotional responses and the actual suicide counts. Our analysis revealed that *surprised* reactions to a celebrity suicide are strongly associated with the actual suicide counts – while reactions expressing *sadness* or *anger* did not reveal a significant link (Fig. 7, Table 3). Reactions expressing *condolences* were negatively correlated, which makes intuitive sense; when a large proportion of responses to a death are polite messages of condolence (which, in Japanese as in English, are primarily drawn from a small number of set phrases used for such occasions), we can infer that while people may find the death tragic or regrettable, they do not find it personally upsetting or depressing.

The connection between *surprised* reactions and the suicide count also makes a great deal of sense when it is considered in context of our findings regarding the individual factors which drive strong emotional responses. In particular, we noted that social media users expressed more emotion (and particularly more surprise) in respect to the suicides

of women, young people and entertainers. Women are overall less likely to die by suicide than men, so surprise at the death of a high-profile woman is to be expected; similarly, the death of a young person is more likely to surprise people than that of an older person. In the case of entertainers, we noted that some of those in the "other career" categories (which encompassed businessmen, politicians and scientists, among others) had been embroiled in scandals prior to their deaths, which can be expected to moderate people's surprised reactions to those deaths.

In short, the *surprise* emotional category was effectively providing a measurement of the degree to which social network users found a death shocking – due to the factors we identified (gender, age, and occupation), and potentially a range of other factors we did not include in our analysis. This sense of shock and surprise then translates into a direct effect on the suicide count, suggesting that at the heart of the Werther Effect's mechanism an important role is played by a response to a death which is shocking and unexpected. This can be viewed as supporting evidence for the social learning hypothesis regarding the mechanism of the Werther Effect; surprised reactions show that a person's death by suicide was unexpected, meaning that for some individuals this death may change their perception of the act of suicide or of the circumstances in which it is possible or even "warranted". Tweets which expressed *sadness* or *anger*, by comparison, were not associated with any change in the suicide count – suggesting that these emotional responses do not play a role in this specific part of the Werther Effect's functioning. The fact that *sadness* did not prove relevant in our analysis supports prior research dismissing the grief hypothesis as it relates to the Werther Effect (Blood and Pirkis, 2001).

4.1. Limitations

The approach we employed in this study has a number of important limitations. Firstly, while the range of emotions that can be expressed in a tweet is quite significant – arguably especially so in Japanese, where the use of *kanji* ideographs permits the expression of complex concepts in just one or two characters – the study had to limit itself to a small set of broadly defined emotional responses for practical purposes. More fine-grained analysis would encounter two key problems; firstly, the degree of subjectivity involved in judging the dominant emotion of a tweet would become increasingly problematic as the number of categories (and thus potential overlap between them) increased, and secondly, the difficulty of training an effective classification algorithm (during the machine learning step of the methodology) would also grow rapidly as the categories became more fine-grained. The impact of this problem was minimised by running a test-coding experiment which identified emotional categories that appeared often enough to be coded reliably (and also highlighted the issue of standard condolence messages being categorised as *sadness*, leading us to create a separate category for them). However, the category schema remains quite coarse, so future studies may find value in judicious expansion of the emotional categories used.

A further limitation of the approach is that it only allows us to study high-level relationships between emotional responses and suicide – looking at almost a million tweets spread across 18 celebrity suicide deaths gives us a data set which is very amenable to empirical investigation and machine learning approaches but permits only conjecture regarding the micro-level of these relationships and mechanisms. We have, for example, no way of identifying tweets from users who are at-risk for suicide (or who subsequently died by suicide) and looking at their reactions specifically to see how they responded to celebrity deaths, or how those responses differed from those of other groups.

4.2. Concluding remarks

One point we consider very important for future research is the question of whether social networks are simply usurping the role of traditional media in reporting on celebrity suicide, or whether the discursive and interactive nature of social media also plays a role in shaping the functioning of the Werther Effect across this medium. If social networks are simply acting as a reporting channel, then the finding of Ueda et al., (2017) reflects at-risk groups being more likely to use social media than to watch TV or read newspapers, and the findings of the present study can be interpreted as a potential mechanism for using social media analytics to anticipate danger to at-risk groups and individuals in the wake of a high-profile suicide. If, however, social networks are also functioning to modulate the functioning of the Werther Effect – in other words, if the tenor and emotional content of social media discourse, not just the reporting of the death itself - has a direct impact on at-risk groups and individuals, then the emotional responses reflected in this study are not merely a way of measuring risk but are active participants in the causal mechanism. Without micro-level analysis, however, it is impossible to say which of these scenarios is accurate.

In either scenario, careful consideration must also be made of how the impact of the Werther Effect might be minimised on social network platforms. The Japanese government recently called on Twitter to block specific suicide-related keywords from its discussions – while the effectiveness of this kind of measure is untested, it demonstrates an interest on the part of legislators and regulators in moderating social networks in such a way as to minimise the Werther Effect and the spread of suicidal ideation more generally. We note that while newspapers and other traditional media outlets in many jurisdictions have guidelines aimed at reducing the extent to which consumers will be exposed to news of a celebrity suicide, social networks have no such controls. Guidelines for the use of social media by organisations or individuals engaging with mental health issues do exist (e.g. Entertainment Industries Council, 2014), but social media platforms themselves are generally not subject to any such code or guidelines; hence, when a celebrity suicide becomes a major topic of social media discussion, a user (at-risk or otherwise) might be exposed and re-exposed to news and discussions of that suicide dozens or even hundreds of times per day. Policies aimed at decreasing the extent of exposure to such reports should be researched and tested to see if they can be effective in diminishing the Werther Effect. The links between emotional responses and suicide counts shown in our study also suggest that monitoring of the emotional tenor of discussions around such topics could help to guide social network operators to make responsible and effective decisions about how to moderate the flow of information related to them.

Funding

This work was financially supported by JSPS Grants-in-Aid for Scientific Research (Grant Numbers: 17H02541, 26870326) and the Innovative Research Program on Suicide Countermeasures Research Grant, and the Telecommunications Advancement Foundation Research Grant. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgement

The Twitter data used in this paper were originally collected with Yasuyuki Sawada and Kota Mori.

Technical Appendix. Machine learning technical details

A number of steps were taken to create the supervised machine learning system which categorised our data. First, it was necessary to perform some filtering steps on the coded training data created by our human coders. Coders had been permitted to select either one or two categories for each tweet; we judged this multi-coding approach to be

appropriate in this instance due to the nature of the Japanese language, which is much more semantically dense than European languages, permitting the expression of multiple sentiments within the 140-character limit. In practice, this multi-coding approach created difficulties with ICR measurements, with a measurement on the "raw" coded data yielding an average observed agreement between coders of 0.760; Cohen's Kappa was 0.606, while Krippendorff's Alpha was 0.598 – somewhat below the range considered acceptable for use in research.

In order to improve this score and the reliability of the data, two measures were taken based on accepted post-coding data refinement measures - "after data have been generated, reliability may be improved by removing unreliable distinctions from the data, recoding or lumping categories, or dropping variables that do not meet the required level of reliability" (Krippendorff, 2006, p. 430). First, the multi-categories (which accounted for a very small number of tweets) were merged back into their most appropriate parent categories (i.e. tweets belonging to two categories were merged with both in turn and the merge which yielded the highest agreement between coders was retained). Secondly, the work of a single coder whose agreement scores with the other three coders were low was dropped entirely, leaving us with a total of 7741 categorised tweets. By doing this, the inter-coder reliability score was raised to an average observed agreement of 0.843, Cohen's Kappa of 0.715, and Krippendorff's Alpha of 0.714. This is above the 0.667 score generally considered to be the cut-off point for drawing tentative conclusions from the data; while it remains below the 0.80 "gold standard" for inter-coder reliability, we judged that given the constraints imposed by both the slightly subjective nature of the task at hand (judging emotional content rather than, for example, identifying the topic of a piece of text) and of the data being studied (Twitter data with many thousands of different authors, writing styles and so on), 0.714 is an acceptable Alpha score in this context.

The 7741 coded tweets were used as the development set, which was repeatedly shuffled and divided up into *training* and *test* sets - allowing each model to be tested on unseen data, with the results cross-validated by repeated training and testing on different samples of the development set. A grid search approach was used to establish the most effective combination of hyper-parameters (priors used to guide the training of the model) for each model. Ultimately, the best model and combination of parameters which emerged from the search was the Linear Support Vector Classifier using a squared-hinge loss function, l2 penalty and a penalty parameter (C) of 0.1. This model achieved precision of 0.805 on the test data set, marginally higher than the best scores from other competitive algorithms such as Stochastic Gradient Descent (0.79) and Support Vector Classification with a Radial Basis Function (RBF) kernel (0.78).

In practice, this means that the trained Linear SVC model can be expected to classify unseen Tweets in a manner that agrees with our human coders slightly over 80% of the time. This figure is close to the average observed agreement of our human coders themselves (0.843), which has two implications. Firstly, it suggests that the model is approaching the upper bound of its hypothetical reliability, since it cannot reasonably be expected to outperform human coders' agreement given that it was trained on data they created. Secondly, it means that the model is performing almost as effectively as our human coders; while the 0.805 accuracy figure is important to bear in mind in interpreting our results, overall its conclusions may be treated as if they had been achieved by one of the human coders, which is the ultimate aim of training this kind of model.

References

Blood, R.W., Pirkis, J., 2001. Suicide and the media. Part III: theoretical issues. Crisis 22, 163–169. https://doi.org/10.1027//0227-5910.22.4.163.

Cavazos-Rehg, P.A., Krauss, M.J., Sowles, S., Connolly, S., Rosas, C., Bharadwaj, M., Bierut, L.J., 2016. A content analysis of depression-related Tweets. Comput. Hum. Behav. 54, 351–357. https://doi.org/10.1016/j.chb.2015.08.023.

Colombo, G.B., Burnap, P., Hodorog, A., Scourfield, J., 2016. Analysing the connectivity and communication of suicidal users on twitter. Comput. Commun. 73, 291–300. https://doi.org/10.1016/j.comcom.2015.07.018.

Daine, K., Hawton, K., Singaravelu, V., Stewart, A., Simkin, S., Montgomery, P., 2013. The power of the web: a systematic review of studies of the influence of the internet on self-harm and suicide in young people. PloS One 8, e77555. https://doi.org/10.1371/journal.pone.0077555.

Dillman Carpentier, F.R., Parrott, M.S., 2016. Young adults' information seeking following celebrity suicide: considering involvement with the celebrity and emotional distress in health communication strategies. Health Commun. 31, 1334–1344. https://doi.org/10.1080/10410236.2015.1056329.

Entertainment Industries Council, 2014. Social Media Guidelines for Mental Health Promotion and Suicide Prevention.

Fahey, R.A., 2017. ja_tokeniser: MeCab-based Japanese Language Tokeniser Optimised for Twitter Data. GitHub.

Jashinsky, J., Burton, S.H., Hanson, C.L., West, J., Giraud-Carrier, C., Barnes, M.D., Argyle, T., 2014. Tracking suicide risk factors through Twitter in the US. Crisis 35, 51–59. https://doi.org/10.1027/0227-5910/a000234.

Karamshuk, D., Shaw, F., Brownlie, J., Sastry, N., 2017. Bridging big data and qualitative methods in the social sciences: a case study of Twitter responses to high profile deaths by suicide. Online Soc. Netw. Media 1, 33–43. https://doi.org/10.1016/j.osnem.2017.01.002.

Krippendorff, K., 2006. Reliability in content analysis. Hum. Commun. Res. 30, 411–433. https://doi.org/10.1111/j.1468-2958.2004.tb00738.x.

Kudo, T., 2016. Mecab: yet Another Japanese Morphological Analyzer. GitHub.

Niederkrotenthaler, T., Till, B., Kapusta, N.D., Voracek, M., Dervic, K., Sonneck, G., 2009. Copycat effects after media reports on suicide: a population-based ecologic study. Soc. Sci. Med. 69, 1085–1090. https://doi.org/10.1016/j.socscimed.2009.07.041.

O'Dea, B., Wan, S., Batterham, P.J., Calear, A.L., Paris, C., Christensen, H., 2015. Detecting suicidality on twitter. Internet Interv 2, 183–188. https://doi.org/10.1016/j.invent.2015.03.005.

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, É., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Phillips, D.P., 1974. The influence of suggestion on suicide: substantive and theoretical implications of the werther effect. Am. Sociol. Rev. 39, 340–354. https://doi.org/10.2307/2094294.

Pirkis, J.E., Burgess, P.M., Francis, C., Blood, R.W., Jolley, D.J., 2006. The relationship between media reporting of suicide and actual suicide in Australia. Soc. Sci. Med. 62, 2874–2886. https://doi.org/10.1016/j.socscimed.2005.11.033.

Plutchik, R., 1980. Emotion: a Psychoevolutionary Synthesis. Harper and Row.

Sato, T., Hashimoto, T., Okumura, M., 2017. Implementation of a word segmentation dictionary called mecab-ipadic-NEologd and study on how to use it effectively for information retrieval (in Japanese). In: Proceedings of the Twenty-three Annual Meeting of the Association for Natural Language Processing. The Association for Natural Language Processing pp. NLP2017-B6-1.

Stack, S., 1987. Celebrities and suicide: a taxonomy and analysis, 1948-1983. Am. Sociol. Rev. 52, 401–412. https://doi.org/10.2307/2095359.

Stack, S., 2011. Suicide in the media: a quantitative review of studies based on nonfictional stories. Suicide Life-Threatening Behav. 35, 121–133. https://doi.org/10.1521/suli.35.2.121.62877.

Sykora, M.D., Jackson, T.W., O'Brien, A., Elayan, S., 2013. Emotive Ontology: extracting fine-grained emotions from terse, informal messages. IADIS Int. J. Comput. Sci. Inf. Syst. 8, 106–118.

Ueda, M., Mori, K., Matsubayashi, T., 2014. The effects of media reports of suicides by well-known figures between 1989 and 2010 in Japan. Int. J. Epidemiol. 43, 623–629. https://doi.org/10.1093/ije/dyu056.

Ueda, M., Mori, K., Matsubayashi, T., Sawada, Y., 2017. Tweeting celebrity suicides: users' reaction to prominent suicide deaths on Twitter and subsequent increases in actual suicides. Soc. Sci. Med. 189, 158–166. https://doi.org/10.1016/j.socscimed.2017.06.032.

Värnik, P., 2012. Suicide in the world. Int. J. Environ. Res. Publ. Health 9, 760–771. https://doi.org/10.3390/ijerph9030760.

Wasserman, I.M., 1984. Imitation and suicide: a reexamination of the werther effect. Am. Sociol. Rev. 49, 427–436. https://doi.org/10.2307/2095285.

Woo, H., Cho, Y., Shim, E., Lee, K., Song, G., 2015. Public trauma after the sewol ferry disaster: the role of social media in understanding the public mood. Int. J. Environ. Res. Publ. Health 12, 10974–10983. https://doi.org/10.3390/ijerph120910974.