Supplementary Appendix to

Association of 7 million+ tweets featuring suicide-related content with daily calls to the Suicide Prevention Lifeline and with suicides, USA 2016-2018

Thomas Niederkrotenthaler, Ulrich S. Tran, Hubert Baginski, Mark Sinyor, Markus J. Strauss, Steven A. Sumner, Martin Voracek, Benedikt Till, Sean Murphy, Frances Gonzalez, Madelyn Gould, David Garcia, John Draper, Hannah Metzler

Table of Contents

Supplemental Text S1. Key words used to identify suicide-related tweets

Twitter data. In order to develop the machine learning approach described in detail in Metzler et al., 2021, we used the analytics platform Crimson Hexagon (now named Brandwatch; https://forsight.crimsonhexagon.com, https://brandwatch.com), to download the full daily volume of original English language tweets from users in the US posted between January 1, 2013, and May 31, 2020. Each post included at least one of the following keywords: suicide, suicidal, killed himself, killed herself, kill himself, kill herself, hung himself, hung herself, took his life, took her life, take his life, take her life, end his own life, end her own life, ended his own life, ended her own life, end his life, end her life, ended his life, ended her life, ends his life, ends her life. We excluded tweets about suicide bombing, and tweets containing keywords indicating that the term suicide was not used to refer to someone ending their life. These exclusion terms were: suicide squad (a movie), suicidechrist, SuicideGirl* (a website featuring pin-up photography of models), SuicideBoy* (male models), suicideleopard (a frequently mentioned Twitter user), suicidexjockey* (a Twitter user), suicidal grind (a music album), Epstein (speculation on the death of Jeffrey Epstein), political suicide (tweets using suicide as a metaphor for political failure), and leading political figure names including Trump, Clinton*, Hillary, Biden, Sanders (for whom the term suicide was used in political contexts).

Tweets about Epstein's suicide were excluded because many of them were about speculations about a possible suicide rather than the actual suicide, and Epstein was portrayed as a villain. Both of these characteristics mean that any effects on suicides are unlikely (see Niederkrotenthaler et al., 2009).

References

Metzler H, Baginski H, Niederkrotenthaler T, Garcia D. Detecting Potentially Harmful and Protective Suicide-related Content on Twitter: A Machine Learning Approach. ArXiv 2021: 2112.04796 [Cs]. http://arxiv.org/abs/2112.04796 [in press at JMIR- Journal of Medical Internet Research].

Niederkrotenthaler T, Till B, Voracek M, Dervic K, Kapusta ND, Sonneck G. Copycat effects after media reports on suicide: A population-based ecologic study. Social Science & Medicine. 2009;69:1085–1090.

Supplemental Text S2. Outliers identified in the Lifeline call time series

In total, there were 20 outliers in the Lifeline call time series suggesting call volumes above or below expected numbers on the respective dates. Some of the respective days were known to impact on suicides based on the epidemiological literature. In particular, eleven outliers were related to suicides by celebrities, public holidays and suicide-prevention related media events. These outliers were not modelled because suicide-related tweets on the given days were assumed to help explain the unusual call volume to the Suicide Prevention Lifeline. Particularly the suicides of US icons Kate Spade and Anthony Bourdain (June 5 and June 8, 2017, respectively); and Logic’s hip-hop song 1-800-273-8255 about the Lifeline number sparked strong media interest over a couple of weeks. The suicides were associated with increases in tweets, calls and suicides (Sinyor et al., 2021) whereas the song was related to an increase in calls and a decrease in suicides (Niederkrotenthaler et al., 2021). It is also known that public holidays including Thanksgiving impact on suicides (Phillips & Wills, 1987).

For the remaining nine outliers, there were no events known to be associated with suicides. Most of them were probably related to technical glitches in call registration. These outliers were modelled in the analysis.

Date	Type of outlier	Possible reason	Modelled
November 9 2016	additive	US General Elections	yes
November 17 2016	innovational	possible technical glitch	yes
January 6 2017	additive	possible technical glitch	yes
January 16 2017	additive	possible technical glitch	yes
February 4 2017	additive	possible technical glitch	yes
February 20 2017	additive	possible technical glitch	yes
May 6 2017	additive	possible technical glitch	yes
May 30 2017	additive	possible technical glitch	yes
August 28 2017	level shift	MTV Music Awards—Logic Song	no
November 23 2017	additive	Thanksgiving weekend	no
November 26 2017	additive	Thanksgiving weekend	no
November 27 2017	additive	possible technical glitch	yes
January 29 2018	innovational	Grammy Awards—Logic Song	no
June 6 2018	innovational	suicide Kate Spade (KS)	no
June 8 2018	level shift	suicide KS / suicide Anthony Bourdain (AB)	no
June 9 2018	innovational	suicides KS & AB	no
June 12 2018	level shift	suicides KS & AB	no
June 14 2018	level shift	suicides KS & AB	no

October 10 2018	innovational	World Mental Health Day	no
November 22 2018	additive	Thanksgiving weekend	no

Refererences

Niederkrotenthaler T, Tran U, Gould M, Sinyor M, Sumner S, Strauss MJ, Voracek M, Till B, Murphy S, Gonzalez F, Spittal MJ, Draper J. Association of Logic’s Hip Hop Song 1-800-273-8255 with Lifeline Calls and Suicides in the United States: Interrupted Time-Series Analysis. BMJ. 2021; 375:e067726.

Phillips DP, Wills JS. A drop in suicides around major national holidays. Suicide Life Threat Behav. Spring 1987;17(1):1-12doi: 10.1111/j.1943-278x.1987.tb00057.x

Sinyor M , Tran US, Garcia D, Till B, Voracek M, Niederkrotenthaler T. Suicide mortality in the United States following the suicides of Kate Spade and Anthony Bourdain. Aust N Z J Psychiatry. 2021;55:613-9. doi:10.1177/0004867420976844

Supplemental Text S3. Summary of machine learning analyses

Please see Metzler et al. for a detailed explanation of the machine-learning basis of this work (Metzler et al., 2021). We manually labelled 3,200 English language tweets, using an annotation scheme specifically developed for social media data.

Preprocessing. Before making predictions of tweet categories, we preprocessed the text of tweets using standard preprocessing strategies (Guogin, 2019). Specifically, we replaced all URLs with http, all user mentions with @user, and lower-cased all words. Emojis, stopwords (e.g., pronouns, articles, function words), and punctuation were retained and separated into single tokens (i.e., words), given that they can indicate the emotional connotation of a message (e.g., “!” expressing excitement or surprise, frequent singular pronouns indicating suicidal ideation; DeChoudhury, 2016).

Model predictions for the six categories. To automatically label the 7.15 million tweets retrieved for the time period January 1, 2016 to December 31, 2018, we applied the best performing machine-learning model for this task, developed in previous work (Metzler et al 2021). This model was based on a pretrained BERT (Bidirectional Encoder Representations from Transformers) base uncased model (https://huggingface.co/bert-base-uncased; Devlin et al., 2019) and fine-tuned to classify the six categories. BERT is a deep contextual language representation model developed by Google AI that learns to make predictions by the sequence of all words in the sentence. Across the six categories, the BERT model correctly classified 74% of tweets on average from novel data not used during model training. F1-scores, which are a standard evaluation metric in machine learning and defined as the harmonic mean between precision and recall of the respective model (see Metzler et al., 2021), were between 55% to 85% for all six content categories (above 70% for all but the suicidal ideation and attempts without coping category, which had a lower score in part due to problems in separating them from sarcastic, non-serious tweets, see Metzler et al., 2021). These classification performances are comparable to the state-of-the-art on similar tasks (e.g., Burnap et al. 2017). The model’s agreement with human annotations was comparable to the agreement between two human raters. For the five categories of interest (which were analysed using time-series models), human interrater reliability (Cohen’s kappa) was 0.85 between human raters, and 0.81/0.80 between the model and each of the two human raters.

References

Burnap, P., Colombo, G., Amery, R., Hodorog, A., & Scourfield, J. (2017). Multi-class machine classification of suicide-related communication on Twitter. Online Social Networks and Media, 2, 32–44. https://doi.org/10.1016/j.osnem.2017.08.001

De Choudhury, M., Kiciman, E., Dredze, M., Coppersmith, G., & Kumar, M. (2016). Discovering Shifts to Suicidal Ideation from Mental Health Content in Social

Media. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 2098–2110.

Guoqin Ma. (2019). Tweets Classification with BERT in the Field of Disaster Management (Tech. Rep No. 15785631). Department of Civil Engineering, Stanford University.

Metzler H, Baginski H, Niederkrotenthaler T, Garcia D. Detecting Potentially Harmful and Protective Suicide-related Content on Twitter: A Machine Learning Approach. ArXiv 2021: 2112.04796 [Cs]. http://arxiv.org/abs/2112.04796 [in press at JMIR- Journal of Medical Internet Research].

Supplemental Text S4. The Statistical Model

We used SARIMA (Seasonal Autoregressive Integrated Moving Average) to analyze the present daily time-series data (an introduction to this class of models and their application in epidemiological research is provided in Schaffer et al., 2021). (S)ARIMA models are standard in psychiatric and public health literature and we have used them in previous analyses for the same type of outcome data, i.e. suicides and Lifeline call data (e.g., Niederkrotenthaler et al., 2021; Niederkrotenthaler et al., 2019; Sinyor et al., 2019). Other modelling approaches such as Vector Autoregression (VAR) are also options for these data but we used SARIMA as a pretty robust and conservative approach that has been widely used in the field of psychiatric and suicide research. SARIMA modeling allows for a flexible approach to the modeling of dependent time series, of outliers, and of independent (time-varying) variables that might be associated with changes in the dependent time series. These considerations were all relevant in the present study, as we investigated the temporal associations between proportions of tweets and Lifeline call and suicide time series, respectively, which, themselves, were affected by background events, but also included outliers and technical anomalies (see Supplemental Text S2 in this supplement for an overview of outliers).

SARIMA models are characterized by six parameters, SARIMA(p,d,q)(P,D,Q). P denotes the position or number of time lags that autoregressively impact on current values, d denotes the number of times the differences between consecutive values were computed to remove trends and to reduce non-stationarity in the time series, and q denotes the position or number of current and past error terms which affect current values (random shocks); P, D, and Q are the respective parameters of a SARIMA model with a periodicity of 7 (i.e., conforming to the days of the week). In the present study, seasonality therefore related to weekly periodicities in the data.

Stationarity is a an important prerequisite of SARIMA modeling. This requires that the mean (and the variance/autocovariance) of the series remains constant over time. Stationarity demands the removal of time trends, which was accomplished via the differencing operation in the SARIMA model for the Lifeline time series. No seasonal (i.e., weekly) trends were observed in the present time series, which would have required seasonal differencing.

The selection of models was aided by the SPSS Expert Modeler function, version 26 (IBM), choosing models with the lowest Bayesian information criterion value, highest stationary R2 value (the variance attributable to the fitted time-series model), and, a not significant Ljung-Box Q statistic (indicating whether residuals could be assumed white noise, with stated df). The Q statistic tests whether the autocorrelations of the model residuals at any or some lags are different from zero. Violation of this assumption indicates a relative loss of efficiency of parameter estimation.

We visually inspected the time-series plots for trends in the data, and the plots of the autocorrelation function (ACF) and the partial autocorrelation function (PACF) for important autocorrelative patterns, in addition to the Q statistic. Due to many outliers and technical anomalies, model building progressed manually for the Lifeline data, first performing a differencing operation (d = 1) to remove a clearly visible trend in the time-series plot and then adding autoregressive parameters and random shocks one at a time (starting with lag 1 in both cases) until additional parameter did not reach significance anymore. Outliers were subsequently detected with the SPSS Expert Modeler function and incorporated into the final model. First differencing (d = 1) was also used for the independent variables in the analysis of Lifeline call data.

For model selection, we deliberately aimed at obtaining the simplest and most parsimonious models that were compatible with the data (following the principle of Occam's razor), i.e., models with the smallest possible orders of the parameters p, d, q, P, D, and Q. This approach conforms to recommendations in the literature, which highlight that there may be more than one “correct” model and that the most parsimonious model should be selected (e.g., Schaffer et al., 2021).

SARIMA modelling (and time-series modelling in general) does not allow to directly establish evidence of causal effects. For instance, outcomes still could have been caused by variables which were not part of the investigated model. However, in the course of “natural experiments” and in the absence of the possibility of conducting randomized controlled trials, time-series modelling rigorously provides a more robust and valid assessment of the associations of independent variables with outcomes than other approaches (see also Schaffer et al., 2021). Therefore, our analysis provides evidence of associations, adjusting for temporal effects.

References

Schaffer, A. L., Dobbins, T. A., & Pearson, S.-A. (2021). Interrupted time series analysis using autoregressive integrated moving average (ARIMA) models: A guide for evaluating large-scale health interventions. BMC Medical Research Methodology, 21, Article 58. https://doi.org/10.1186/s12874-021-01235-8

Niederkrotenthaler T, Tran U, Gould M, Sinyor M, Sumner S, Strauss MJ, Voracek M, Till B, Murphy S, Gonzalez F, Spittal MJ, Draper J (2021). Association of Logic's Hip Hop Song 1-800-273-8255 with Lifeline Calls and Suicides in the United States: A Time-Series Analysis. BMJ 375:e067726.

Niederkrotenthaler T, Stack S, Till B, Sinyor M, Pirkis J, Garcia D, Rockett IRH, Tran US (2019). Association of Increased Youth Suicides in the United States With the Release of 13 Reasons Why. JAMA Psychiatry 76(9):933-940.

Sinyor M, Williams M, Tran US, et al. Suicides in young people in Ontario following the release of “13 Reasons Why”. Can J Psychiatry 2019; 64: 798–804.