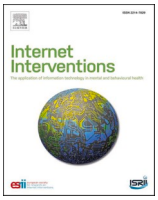
Internet Interventions 25 (2021) 100424

Contents lists available at ScienceDirect

Internet Interventions

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



Predicting acute suicidal ideation on Instagram using ensemble machine learning models

Damien Lekkas [a,c,*], Robert J. Klein [a], Nicholas C. Jacobson [a,b]

[a] Center for Technology and Behavioral Health, Geisel School of Medicine, Dartmouth College, United States of America
[b] Department of Biomedical Data Science, Geisel School of Medicine, Dartmouth College, United States of America
[c] Quantitative Biomedical Sciences Program, Dartmouth College, United States of America

ARTICLE INFO

Keywords:
Suicidal ideation
Digital phenotyping
Machine learning
Suicide prediction
Social media

ABSTRACT

Introduction: Online social networking data (SN) is a contextually and temporally rich data stream that has shown promise in the prediction of suicidal thought and behavior. Despite the clear advantages of this digital medium, predictive modeling of acute suicidal ideation (SI) currently remains underdeveloped. SN data, in conjunction with robust machine learning algorithms, may offer a promising way forward.

Methods: We applied an ensemble machine learning model on a previously published dataset of adolescents on Instagram with a prior history of lifetime SI ($N = 52$) to predict SI within the past month. Using predictors that capture language use and activity within this SN, we evaluated the performance of our out-of-sample, cross-validated model against previous efforts and leveraged a model explainer to further probe relative predictor importance and subject-level phenomenology.

Results: Linguistic and SN data predicted acute SI with an accuracy of 0.702 (sensitivity = 0.769, specificity = 0.654, AUC = 0.775). Model introspection showed a higher proportion of SN-derived predictors with substantial impact on prediction compared with linguistic predictors from structured interviews. Further analysis of subject-specific predictor importance uncovered potentially informative trends for future acute SI risk prediction.

Conclusion: Application of ensemble learning methodologies to SN data for the prediction of acute SI may mitigate the complexities and modeling challenges of SI that exist within these time scales. Future work is needed on larger, more heterogeneous populations to fine-tune digital biomarkers and more robustly test external validity.

1. Introduction

Each year, suicide claims the lives of an estimated 800,000 people worldwide (World Health Organization, 2014). Suicide is the thirteenth leading cause of death internationally and is the leading cause of death among 15- to 39-year-olds (World Health Organization, 2014). These numbers may not reflect the true prevalence of suicidal behavior as they rely on self-report surveys with limited validity as well as inconsistent criteria in registration across nations. Moreover, they fail to account for non-fatal suicide-related behavior (such as suicide attempts) which may be up to twenty times more common than fatal outcomes (World Health Organization, 2009). Given this alarming ubiquity, novel methods for identifying individuals at risk for suicide would be timely (The National Action Alliance for Suicide Prevention, R.P.T.F., 2014).

Other than past suicidal behavior, suicidal ideation (hereafter SI) is perhaps the most consistent predictor of future suicidal risk (Franklin et al., 2017; Leon et al., 1990; Mann et al., 1999; Mundt et al., 2013). In the modern era, SI is often expressed across internet platforms such as social media (Luxton et al., 2012; Marchant et al., 2017), and the relevant data has been valuable in the development of new tools aimed at predicting suicidal outcomes and related risk factors (Aladağ et al., 2018; Birjali et al., 2017; Burnap et al., 2017; Roy et al., 2020). Such online suicide risk assessment could conceivably be used to identify at-risk individuals and deliver interventions to mitigate suicidal behavior (Adrian and Lyon, 2018). The process of developing so-called "digital phenotypes" of suicide risk is in its early stages, though, and the related technology continues to grow in scope and sophistication (Braithwaite et al., 2016; Coppersmith et al., 2018; O'Dea et al., 2015). To contribute

* Corresponding author at: Center for Technology and Behavioral Health, Geisel School of Medicine, Dartmouth College, 46 Centerra Parkway, Suite 300, Lebanon, NH 03766, United States of America.
E-mail address: Damien.Lekkas.GR@dartmouth.edu (D. Lekkas).

https://doi.org/10.1016/j.invent.2021.100424
Received 12 November 2020; Received in revised form 17 June 2021; Accepted 2 July 2021
Available online 6 July 2021

to these efforts, the present investigation employed a deep ensemble machine learning approach to predict acute SI using textual and user metadata collected from the social platform, Instagram. Borrowing from the Fluid Vulnerability Theory of suicidal ideation, we conceptualized acute SI as an episodic, short-term manifestation of SI's dynamic phenomenology, namely, the current heightened expression of suicidal thoughts and feelings (Kleiman and Nock, 2018; Rudd, 2006).

Social networking (SN) data represents a promising information stream for identifying digital phenotypes of suicide risk. Public posts on social media outlets such as Reddit, Instagram, and Facebook provide an opportunity to more naturalistically observe behavior that is not altered by laboratory settings. SN data is often authored without expectation of privacy, minimizes retrospective bias, and has great utility in young cohorts given the heavy concentration of users in this age group (Nesi, 2020) - a group within which suicide is a leading cause of death. Anonymized SN data can also be obtained and analyzed in real time, allowing for rapid, in-the-moment assessment. To date, several studies have leveraged SN data to predict suicide-related outcomes. One study helped establish basic links between SN data and real-world suicidal outcomes when the authors established a connection between frequency of suicide-related tweets in a given state and actual suicide rates in that state (Jashinsky et al., 2013). In a predictive model, another study analyzed public Reddit posts related to mental health and showed that linguistic (e.g., poor coherence) and interactional (e.g., reduced social engagement) markers could be extracted and used to predict future transitions to heightened states of SI up to 9 months later (De Choudhury et al., 2016). Relatedly, a third research project used a deep learning model to examine word usage in public posts across multiple SNs (Coppersmith et al., 2018). Coppersmith and colleagues found that a well-fitted model identified individuals at risk for suicidal behavior using data from 1 to 6 months prior to the suicidal attempt with clinically meaningful precision (0.89–0.94 AUC; 70%–85% true positive rate). While the total number of attempting individuals in the dataset was small ($N = 418$), the model was built to predict suicide attempts using only text from a single post, thereby training the model on approximately 400,000 cases. Such findings are promising because they improve our capacity to identify individuals who are at risk for suicide and ultimately could enable targeted online suicide prevention strategies via emerging digital interventions (Oexle et al., 2019; Robinson et al., 2016).

Researchers interested in links between social media content and suicide have employed an array of statistical analytical approaches to the problem of suicide prediction (Grant et al., 2018; Jashinsky et al., 2013). While descriptive analyses have been foundational, suicide researchers have also been interested in prediction-based approaches that leverage past SN activity. For example, one research team developed and trained a deep learning algorithm to identify distinct markers of suicide risk among Reddit users, and used these markers in a prediction framework to pinpoint (precision = 0.79, recall = 0.83, AUC = 0.89) other users who would show SI in the future (De Choudhury et al., 2016). Risk markers were identified using both natural language processing (e.g., readability, first person singular usage) and interaction patterns (e.g., volume or length of posts, number of comments received). The present study uses a similar strategy of combining natural language processing and user interaction as risk predictors, but emphasizes the identification of acute suicidal risk. In a more recent analysis, a research team trained and validated a system of neural networks capable of reducing Twitter posts to a "0" or "1" binary classification designation on each of twelve key psychology constructs (e.g., hopelessness, stress, insomnia; Roy et al., 2020). Individual random forest models based on these key constructs identified individuals about to express SI from matched controls. Classifiers predicted SI events with an AUC of 0.86–0.90. Together, analyses of this type laid the groundwork for the present investigation. Where the aforementioned works operated on data collected across several months, this work seeks to leverage the temporally rich data structure of SN-derived information to predict and explore the dynamics of SI on a more proximal, immediate basis.

Most of the current knowledge on the prediction of suicidal behavior is derived from epidemiologic studies that assess SI and suicide attempts across the lifespan with time intervals that are no less than twelve months in duration (Glenn and Nock, 2014). Accordingly, there is a dearth of studies focusing specifically on acute suicidal behavior and therefore limited information regarding short-term risk factors. Fortunately, modern trends in communication have provided a promising new avenue (Allen et al., 2019), with the dynamic nature of the SN environment conducive to capturing SI that is newly developing or transient. Brown et al. (2019a) and Brown et al. (2019b) is among the few recent researchers that have actively sought to apply data derived from a SN platform to explore acute SI predictive efficacy. In their work, they investigated the association between acute SI and language use as well as user activity and engagement on Instagram. Logistic regression analyses were used to model the predictive capabilities of various linguistic and activity metrics on the presence/absence of acute SI (past month) in a cohort of German adolescents ($N = 52$) with a previous lifetime history of SI. Despite the strengths of their data collection procedure and the associated novelty of the information they leverage, the researchers did not examine the out-of-sample predictive capacity of their model, effectively hindering the opportunity to investigate model generalization and to evaluate the broader applicability of their findings. Their model yielded the capacity to correctly assign 69% of participants to expression of acute SI versus non-acute SI. Such promise warrants further consideration and extension through additional modeling techniques.

To build upon the efforts of Brown et al. (2019a) and Brown et al. (2019b) and contribute a quantitatively rigorous modeling pipeline in the prediction of acute SI online, the current study employed a stacked ensemble machine learning methodology to this publicly available Instagram data set. Additional statistical and visualization techniques that aid in understanding "how the model learned" the data will also be used to provide a more complete view of the dynamics underlying the predictors and guide inference into their relative predictive utility. Specifically, this model explanation will provide the ability to parse predictors by relative importance and allow for the ability to potentially uncover patterns in predictive utility across ecologically valid (i.e., SN constructs such as number of followers and number of likes) and semantically-derived (i.e., linguistic features from interviews) elements. Accordingly, this study is driven by the following hypotheses:

(1) Leveraging a consensus ensemble machine learning model to predict acute SI on Instagram using textual and user engagement metadata will lead to comparable or higher prediction accuracy ($\geq$0.69) in out-of-fold performance compared with the in-sample logistic regression modeling implemented previously.

(2) Given the less contrived setting of SN use and interaction, Instagram user metadata-derived features will have higher average feature importance values, reflecting more significant contributions to model prediction of acute SI across individuals, compared with those features derived from language use in interviews.

2. Materials & methods

2.1. Study population & data set

This work utilized the data collected from a study investigating the associations among acute suicidality, language use, and Instagram activity (Brown et al., 2019b). In this research endeavor, a subset of German adolescents was selected from a larger study that investigated the occurrence of non-suicidal self-injury on Instagram (Brown and Plener, 2017). Public Instagram user data and post content was collected from subjects for four weeks prior to personal interview via Instagram messenger. A 4-week SN collection window was ideal because (i)

gathering data from a single day or single week may not allow our ML models sufficient data to find a signal in the noise, and (ii) there was little risk of 4 weeks being too far in the past given that SI, while tending to fluctuate from day-to-day, has also shown long-term predictive validity as a suicidal risk factor (Leon et al., 1990; Mann et al., 1999; Mundt et al., 2013). The resulting dataset analyzed in the current work therefore consists of a randomized subset of $N = 52$ study participants [mean age = 16.6 years, median age = 16 years, $N = 41$ (78.8%) female, $N = 40$ (76.9%) attending high school, $N = 7$ (13.5%) attending university or professional school, $N = 2$ (3.8%) unemployed] for which an interview via Instagram messenger was conducted and who reported a lifetime history of SI (Brown et al., 2019a). Participants were made aware of the Instagram data collection procedure after the fact, but prior to the interview. Informed consent was obtained regarding subsequent use and publication of both the Instagram and interview data.

2.2. Baseline features

2.2.1. Natural language text analysis

Quantitative interview data and image captions on Instagram were analyzed using the Linguistic Inquiry and Word Count (LIWC) software (Pennebaker et al., 2015) with a validated German dictionary loaded (Wolf et al., 2008). This text analysis tool leverages a dictionary of nearly 6400 expert curated words that have been evaluated for their psychometric properties. Each word is accompanied by a variety of categorical definitions including basic linguistic dimensions as well as a diverse set of psychological processes including affect, social, cognitive, perceptual, and biological. Using these LIWC categories, the following predictors were derived from interviews and used in the current analysis:

(1) total word count
(2) percent of words that were first-person pronouns
(3) percent of words that indicate expression of emotion (affect)
(4) percent of words that are associated with negative emotions (e.g., hatred, sadness)
(5) percent of words that are represented as cognitive processes (e.g., think, perhaps, and else)

The Flesch-Reading-Ease Index (FRE) (Flesch, 1949) was also utilized to calculate the readability and comprehensibility of adolescent responses in the interview. The FRE is a metric that is normalized between 0 and 100, with higher scores indicative of higher ease of readability. A written passage with a score of 25, for example, represents text that is much more difficult to understand than one with a score of 90, the latter of which would be defined as easily comprehensible by an average eleven-year-old student. FRE indices were determined from the average sentence length and the average number of syllables per word as originally defined (Flesch, 1949). All considered, this analysis approach yielded a total of 6 linguistic text analysis features.

2.2.2. Instagram features

Additional Instagram user-specific metadata was also collected and calculated across five features:

(1) number of total followers
(2) number following
(3) number of pictures posted within the last month
(4) average number of comments per picture within the last month
(5) average number of likes per picture within the past month

2.2.3. Feature engineering

This current study implemented additional feature engineering from the data provided in Section 2.2.2 above to more holistically capture user activity on Instagram. The four features created include:

(1) follow ratio ($\frac{total\ followers}{total\ following}$) (Longobardi et al., 2020; Woodruff et al., 2018)
(2) engagement ($\frac{number\ of\ posted\ pictures*(average\ likes+average\ comments)}{total\ followers}$)
(3) sum of average comments and average likes per follower ($\frac{average\ comments+average\ likes}{total\ followers}$) (De Vries, 2019)
(4) average comments-to-average-likes ratio ($\frac{average\ comments}{average\ likes}$)

The engagement metric was derived from a common formula of engagement rate by reach (ERR) implemented in modern social media analytics (Sehl, 2019), but is slightly modified due to the fact that only averages, and not total counts, of likes and comments were reported.

2.3. Outcome metric

The presence/absence of acute SI was the primary outcome of interest. It was assessed during an Instagram messenger interview with the question, "Are you currently thinking about, or planning to, end your life?" Accordingly, it was defined as positive if the response to this question was "yes" and defined as negative if the response to this question was "no".

2.4. Data preprocessing

All 15 features were individually standardized such that data had a mean of 0 and standard deviation of 1 for use in subsequent models. Subjects for which Instagram user data was not available ($n = 5$) were removed prior to analysis.

2.5. Baseline model testing

In their paper, Brown et al. (2019a) and Brown et al. (2019b) ran an in-sample, stepwise logistic regression with interview language use variables and Instagram activity and found that only negative emotion in interviews was significantly associated with acute SI outcome. They then ran a second univariate logistic regression model with negative emotion as the sole predictor to arrive at an "optimal" model that predicts odds of acute SI with an accuracy of 0.69 (sensitivity = 0.84, specificity = 0.57) at a cut-off of 0.7. This approach, while common, has some notable limitations. Using one model to effectively select notable predictors for a subsequent model on the same dataset leads to bias and an overly optimistic assessment of predictive performance (Reunanen, 2003; Varma and Simon, 2006). As is, the pipeline left much for additional exploration in terms of the relative predictive utility of its various linguistic and Instagram activity features. To serve as a more statistically rigorous initial point of comparison with the current analysis, an out-of-sample, repeated, ten-fold cross-validated stepwise logistic regression model was constructed using the same variables to interrogate association with acute SI. The R packages *caret* and *glmStepAIC* were used to construct and run the model developed by Brown et al. (2019a) and Brown et al. (2019b) within this repeated, cross-validated framework. Variable importance using ROC curve analysis on each predictor, model performance assessed through both accuracy and AUC, as well as final model variable coefficients and significance values are reported.

2.6. Machine learning model building and implementation

The machine learning pipeline was built and run in R (v3.6.1) using the *caret* package. Prediction of individual presence/absence of acute suicidal thoughts was treated as a binary classification task and leveraged all 11 baseline (Sections 2.2.1 and 2.2.2) as well as the four additional derived features (Section 2.2.3). As a first step, seven individual models were run in a repeated, ten-fold cross-validated framework. These include (i) an Extreme Gradient Boosted Tree (xgboost) (Chen and Guestrin, 2016), (ii) a boosted logistic decision tree

(logitboost) (Dettling and Bühlmann, 2003), (iii) a generalized linear model via penalized maximum likelihood (glmnet) (Friedman et al., 2010), (iv) k-nearest neighbors (knn) (Hechenbichler, 2004), (v) a three-layer (i.e., with 1 hidden layer) feed-forward neural network (nnet) (Venables and Ripley, 2002), (vi) aggregated and averaged random seed neural nets (avnnet) (Venables and Ripley, 2002), and (vii) a naive Bayes classifier (naiveBayes) (Majka, 2019). To mitigate data leakage (and an overestimation of model performance) as a result of hyper-parameter tuning at this level, no hyperparameter tuning was performed on the seven lower-level models. Models were run at default package hyperparameter values with the exception of glmnet which followed hyperparameter recommendations outlined in a separate meta-analytical study (Probst et al., 2019). Similar to established procedures outlined in a previous pipeline for predicting psychiatric illness from electronic health records (Nemesure et al., 2020), the prediction probabilities from each model were then used as features for acute suicidal thought prediction in the following five ensemble learning models: (i) xgboost, (ii) logitboost, (iii) knn, (iv) nnet, and (v) avnnet. In this metalayer of stacked models, each model was run within a ten-fold repeated, cross-validated framework with grid search hyperparameter tuning for maximum accuracy using the automatic grid search feature in *caret*. Lastly, predictions across these five stacked ensemble models were averaged to arrive at a final consensus prediction for the acute SI binary classification task. Model accuracy, Kappa score, AUC (sensitivity vs 1 - specificity), and F1 score are reported for each of the five ensemble models as well as for the final consensus model.

2.7. Model introspection and feature importance

Machine learning approaches have traditionally suffered from a lack of transparency which includes an inability to understand how a model "learns the data" and subsequently arrives at a decision. The moniker of "black-box model" reflects this palpable loss of interpretability of machine learning models when compared with traditional statistical models. However, recent methodological advancements have begun to address this limitation and provide a means by which machine learning model predictions can be explained across features at both the global (entire dataset) and local (each data point) level. SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2017) is one such method based upon the Shapley values of game theory (Shapley, 1953). Where Shapley values are conceptualized as relative payouts to players in a cooperative game based on their relative contribution, SHAP equates players as feature values in a prediction task game. As such, SHAP aims to explain the prediction outcome of each sample in the dataset through calculation of each feature's (e.g., Instagram likes) contribution to that prediction. The resulting values are thus interpreted as the relative magnitudes by which features influence prediction outcomes. The SHAP framework is particularly attractive for this analytical pipeline because it is model agnostic and thus applicable across all model types (e.g., linear, tree-based). The *iBreakdown* package in R was used to predict SHAP values for all features across each individual in the dataset. SHAP values were visualized for the consensus ensemble machine learning model using the data structures provided in the *SHAPforxgboost* R package.

For added clarity, Figs. 1 and 2 outline the analytical pipeline described above.

3. Results

3.1. Baseline model

The out-of-sample, ten-fold repeated, cross-validated step-down logistic regression implementation resulted in a final model with both negative emotion in interviews (β = 0.82, OR = 2.28) and number of followers on Instagram (β = 0.96, OR = 2.61) as the only regressors. Similar to the results obtained by Brown et al. (2019a) and Brown et al. (2019b), only negative emotion in interviews was found to be statistically significant ($p = 0.026$) in the prediction of acute SI. Moreover, the resampling results of the model reported an accuracy of 55.6%, a Kappa = 0.087, and an AUC = 0.560 (sensitivity/recall = 0.524, specificity = 0.692, F = 0.550) indicating inferior predictive power than previously reported for the in-sample analysis. Variable importance analysis carried out using the *varImp* function in *caret* further indicated negative emotion in interviews as the most important variable (scaled importance value (SIV) = 100.000). This was followed by percent of affect words in interviews (SIV = 71.831) and FRE (SIV = 47.183) by large margins of difference. Percent of cognitive mechanism words in interviews (SIV = 13.380), number following on Instagram (SIV = 2.113), and the mean number of comments on Instagram (SIV = 0.000) were found to be the least important.

3.2. Consensus ensemble model performance

A machine learning approach was used to predict acute suicidality on Instagram using both linguistic and user activity-based predictors. The modeling pipeline comprised five ensemble models that used the output from the seven lower-level models as predictors. As summarized in Table 1, accuracy across these five ensemble models ranged from 0.523–0.697 (sensitivity/recall = 0.190–0.762, specificity = 0.692–0.962, F1 = 0.308–0.727) with AUCs in the range of 0.510–0.720. The neural net-based models performed the best, and the k-nearest neighbors algorithm performed the worst. When the output predictions for each of these models was averaged to form the final consensus predictions, performance reflected a superior accuracy of 0.702 (sensitivity/recall = 0.769, specificity = 0.654, F1 = 0.741) and an AUC of 0.755. This ensemble machine learning approach, wrapped within 10-fold repeated cross-validation, achieved an overall predictive performance that was superior to baseline (AUC = 0.755 compared to AUC = 0.560). The baseline model attempted to replicate the previous efforts of Brown et al. (2019a) and Brown et al. (2019b) in an out-of-

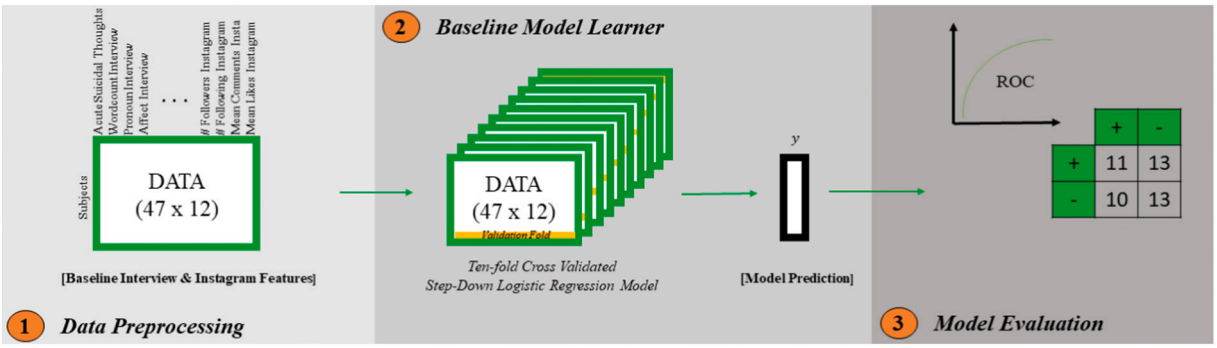


Fig. 1. Analytical pipeline of baseline comparison model.

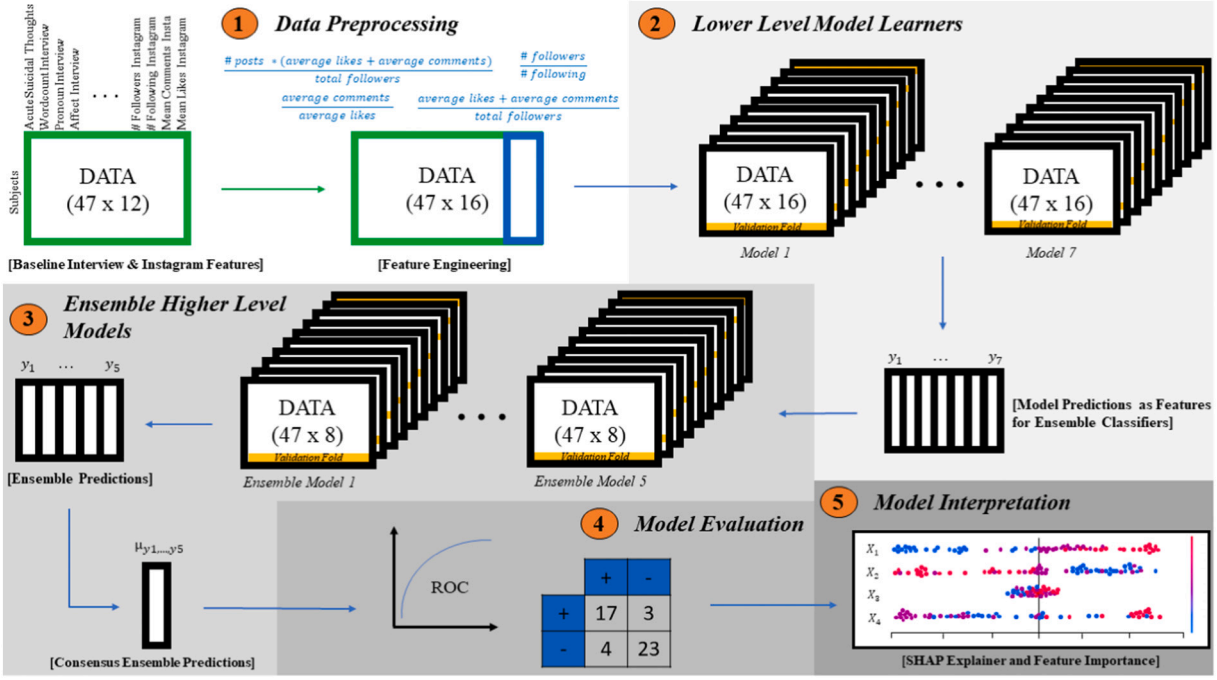


Fig. 2. Analytical pipeline of consensus ensemble model.

Table 1
Ensemble and consensus ensemble model results.

Ensemble models	Accuracy	Kappa	AUROC	Specificity	Sensitivity/recall	F1 score
xgboost	0.633	0.241	0.650	0.692	0.619	0.619
logitboost	0.605	0.197	0.660	0.731	0.571	0.600
knn	0.523	0.025	0.510	0.962	0.190	0.308
nnet	0.697	0.391	0.730	0.731	0.762	0.727
avnnet	0.680	0.353	0.720	0.692	0.762	0.711
Consensus	0.702	0.392	0.755	0.654	0.769	0.741

sample framework for a more direct comparison of model performance. Moreover, the ensemble framework implemented for this study, despite evaluation out-of-sample, matched the accuracy of previous efforts (70.2% compared to the previous 69%) where the metric was interrogated and calculated with an in-sample paradigm.

3.3. Feature contributions in ensemble model prediction using Shapley scores

SHAP values implicated (i) number of accounts the subject followed, (ii) engagement, (iii) negative emotion in interviews, (iv) number of accounts following the subject, and (v) mean likes as the top five most influential features on model prediction. Notably, four of the five predictors were derived from Instagram activity and not from the linguistic characteristics of structured interview content. However, the presence of negative emotion in interviews within these top five features echoes the findings of Brown et al. (2019a) and Brown et al. (2019b) as well as the results of the baseline model where this sole linguistic characteristic was found to be significant in acute SI prediction.

Fig. 4 illustrates the subject-level differences and trends in feature importance. A few notable patterns emerged. First, those with a higher number of accounts followed by the subject tended to have an increasingly positive impact on model prediction, indicating a propensity toward classification as acutely suicidal. The same was true for larger numbers of followers and with overall engagement scores. Individuals with higher values of these features tended to influence the model toward positive predictions of acute SI. Second, individuals with lower numbers of mean likes positively influenced the model toward predictions favoring an acutely suicidal state. Third, the majority of subjects with lower or average follow ratios (i.e., a relatively higher number of accounts the subject follows compared with the number of accounts following the subject) as well as lower or average comment-like ratios had a negative or neutral impact on model prediction with positive impacts to model prediction observed only with higher ratios.

4. Discussion

This study utilized a dataset consisting of LIWC predictors derived from both interview and Instagram post content, as well as those describing Instagram activity over the past month, of German adolescents with a prior lifetime history of SI. The primary result of the current model pipeline indicated that linguistic and SN activity variables are capable of predicting acute SI with an accuracy of 70.2% (specificity = 0.654, sensitivity/recall = 0.769, F1 = 0.741, AUC = 0.755).

As mentioned, this work is an extension of the research conducted by Brown et al. (2019a) and Brown et al. (2019b). In this research, Brown and colleagues investigated a link between acute suicidal ideation, Instagram activity, and language use based on a traditional in-sample analysis of the data. In an initial analysis, the current work began by recapitulating the Brown et al. (2019a) and Brown et al. (2019b) logistic regression model using a ten-fold repeated cross-validation framework. The predictive performance achieved by our re-analysis was much lower than the reported 69% accuracy of the original method. Without employing an out-of-sample, cross-fold validation approach, the data

used to build the original regression model was the same as the data used to quantify its performance, thus our results suggest that the 69% benchmark accuracy was likely an overestimate of the model's capacity to predict. Correcting this, the approximately 54% predictive accuracy achieved with a near-zero Kappa in the current re-analysis suggests that the original logistic regression model may not have been informative on, nor generalizable to, data that it had not previously seen. Despite this performance discrepancy, both the previous efforts of Brown et al. (2019a) and Brown et al. (2019b) and the current model highlighted negative emotion words in interview as the most important variable driving the logistic regression.

The impetus of the current research stemmed in part from the above results. The central question became whether machine learning classification modeling could be leveraged to achieve higher predictive performance of acute suicidal ideation in an out-of-sample paradigm, thus further highlighting the potential utility of Instagram activity and language use information. The current analysis utilized seven model types spanning decision tree-based, supervised clustering, neural network, linear, and probabilistic classifiers for an inclusive and agnostic approach to analysis. The performance results of these models are shown in Table 1 and reflect superior final accuracy of 70.2%, a value that is comparable to that previously reported in the in-sample analysis (Accuracy = 69.0%) and approximately 18% higher than the out-of-sample logistic regression approach (Accuracy = 52.0%). The AUC curves in Fig. 3 depict the trade-off in correctly identifying those who are acutely suicidal and correctly identifying those who are not. Specifically, the attained AUC of 0.755 (sensitivity = 0.769, specificity = 0.654) corresponds with the ability to correctly classify acutely suicidal individuals as being acutely suicidal 76.9% of the time. This statistically significant ($p < 0.05$) improvement in model predictive ability is dramatic compared with the baseline out-of-sample logistic regression (Fig. 3, right).

These findings suggest that the present machine learning approach is quite promising as an avenue for future development in the prediction of acute suicidal thought and behavior online. The present results indicated a balanced improvement in this binary classification task of a complex outcome using data that, by most SN "big data" standards, is characterized by a small sample size and a sparse set of informative predictors. It is not unreasonable to assume that this ensemble approach may prove effective with larger, more feature-rich datasets across different SN platforms since the variables that inform the models are generalizable to other online settings such as Twitter, Facebook, and Reddit. Collection of linguistic and user metadata in the short-term for acute SI prediction may be especially pertinent on platforms primarily operating on more dynamic, interactive timescales such as Twitter. Despite this potential for feature space generalizability, it is important to recognize that training entailed the specific detection of acute SI among individuals with a lifetime history of SI. The resulting discriminatory model, while trained for an especially difficult task given the baseline risk of individuals in this cohort, would not be suitable for deployment in a general population with never-SI controls. This does not necessarily detract from its performance as the inherent nuance of the classification renders the obtained accuracy especially noteworthy.

One traditional shortcoming of machine learning models has been their lack of transparency in how data is used to arrive at a prediction. Recognizing this limitation, the present work was interested in decomposing the resulting ensemble consensus machine learning model in such a way that the relative influence of each variable used to predict status of acute SI could be investigated. The relatively novel SHAP framework served this purpose, and the results of the SHAP analysis offered some notable insights into suicide risk prediction. Fig. 4 illustrates, in ranked order of importance (from top to bottom), the features that were most important in predicting acute SI. Interestingly, four of the top five most influential predictors were associated with social media use behavior rather than with interview-related linguistic content. This speaks to the possible benefits of leveraging discrete behaviors such as "liking" and "following" in addition to natural text processing strategies. Moreover, the free-form, passive collection of behavioral information on social media use may offer a non-intrusive alternative to discern uniquely subtle, abnormal patterns of activity that can be linked to fluctuating suicidal states and serve as informative variables for predictive models. It is important to also mention that negative emotion words in interview were found to be the third most informative predictor and the only linguistic-based feature within the top five most influential features. Recapitulating the results from Brown et al. (2019a) and the corrected baseline model in the current study, its recurrent prominence in the ensemble pipeline further highlights this linguistic feature for predictive application in future research.

In short, the SHAP results suggest that passively collected SN interaction data, especially follower, following, engagement, and like-based metrics, may be particularly useful predictors of future acute SI. This finding carries implications for the identification of suicide risk and

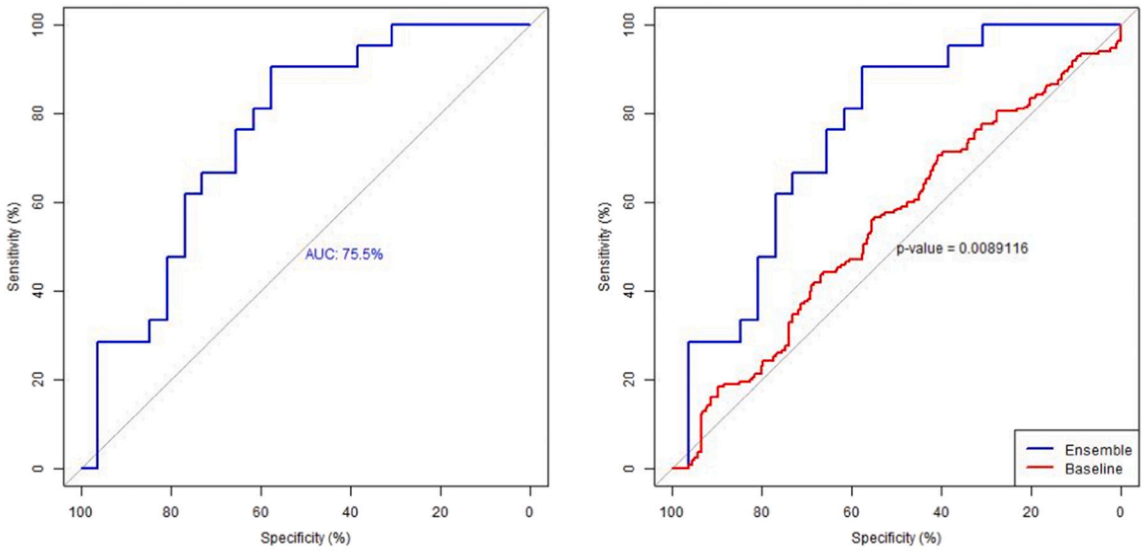


Fig. 3. ROC curves of consensus ensemble model performance.
Note. (A) ROC curve reflects an AUROC of 0.755 (sensitivity = 0.769, specificity = 0.654). (B) Ensemble consensus model ROC (blue) compared with baseline model (red) indicates statistically significant ($p < 0.05$) improvement in predictive performance.

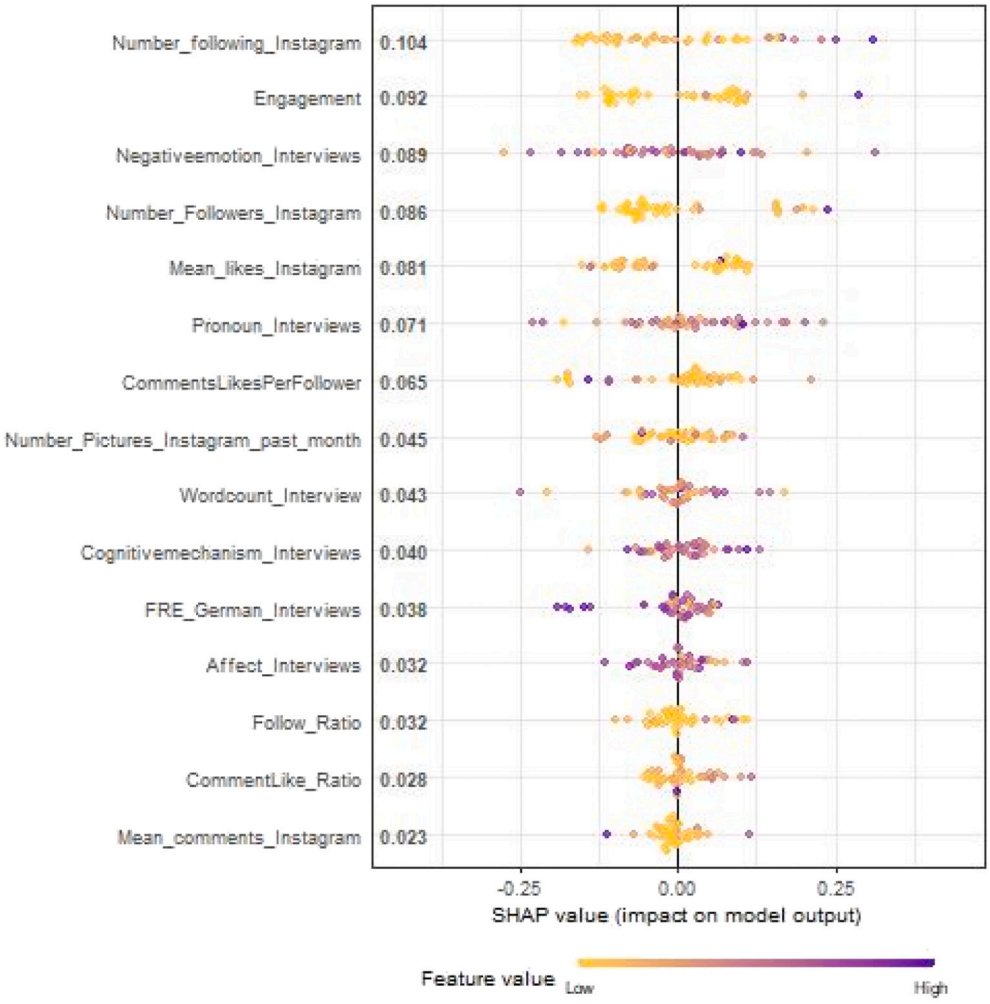


Fig. 4. 2D Plot of SHAP values.
Note. The graph illustrates the relative importance of all features utilized for the prediction of acute suicidal thought in the consensus ensemble machine learning model. The average (across $n = 47$ subjects) SHAP value for each feature is listed next to its respective name and ordered from highest to lowest impact on model prediction.

suicide prevention within the broader medium of online social interaction. Looking at subject-specific values within the SHAP framework may also prove fruitful. For example, in Fig. 4, lower averages of mean likes per post tended to have a positive impact on model prediction. The implication makes intuitive sense from the perspective of SI. Subjects who exhibit acute SI may be more likely to post socially taboo and/or less publicly favorable content that will not garner as many likes relative to more neutral or positively received content posted by those without acute SI. Drawing from established psychological theory, the social implications of acute SI expression within this medium may also be related to excessive reassurance behaviors initially proposed by Coyne in his Interpersonal Theory of Depression (Coyne, 1976). Coyne posits that constant reassurance seeking by depressed individuals may serve to alleviate doubts related to notions of self-worth and in the process leads to the frustration and irritation of friends and family. The observed association between decreased receptivity, here quantified as mean number of likes, and acute SI prediction may highlight an underlying aversion response warrants further investigation. Moreover, this operationalization contributes to a growing consensus regarding the stigmatizing nature of STB (Emul et al., 2011; Kalish, 1966; Lester, 1992; Lester and Walker, 2006; Scocco et al., 2012).

This study is innovative in its use of ensemble machine learning to predict acute suicidal ideation as well as in its application of SHAP values to explain the model and derive insight into potential risk factors. While the current work was capable of detecting some meaningful signal, the small size of the cohort limited the power and interpretation of SHAP since the dynamics of various predictor values have small case representation. As such, further analyses with larger cohorts and more comprehensive SN-based metrics are necessary to reconcile observed patterns with established psychological theory and maximize the descriptive affordances of the SHAP approach. The lack of fine-grained temporal dynamics in the predictors also restricted the confidence in making specific and definitive comments on associations; however, the overall trends in the model architecture serve as promising pilot results that draw attention to the potential significance of SN-derived data in the formulation of acute SI predictive algorithms.

Model performance results were also promising despite the limitations in predictor availability and small sample size. The exhaustive approach to model selection and overall architecture of the pipeline is a notable strength of this work. Specifically, the use of repeated cross-validation in training results in comparatively minimal bias to model performance (Jacobucci et al., 2020), especially when benchmarked against the in-sample approach of the original study. Despite these strengths, there are additional important limitations to consider.

First and foremost, the small sample size of the data restricted the ability to utilize a true train-test split, thus hyperparameter tuning was performed alongside repeated cross-validation. While this limits generalizability and may lead to more optimistic performance results, an exploration of performance across all candidate ensemble models within the hyperparameter grid search space revealed consistency and comparable performance to the selected optimum even among the majority of the lesser performing model iterations (see Supplemental Tables S1–S5 for summaries and hyperparameter values). This suggests that the choice to perform hyperparameter tuning within cross-

validation in a dataset that was unfortunately not large enough to statistically support a more robust approach to validation, while a limitation of the study, has limited impact on the findings.

Next, the summative nature of the variables did not allow for the construction of models that could capture short-term trajectories of acute suicidal ideation risk factors and behaviors. While the current predictions were based on summative snapshots of recent language choice and online behavior, changes that took place throughout this period of time could not be considered. By extension, this meant that we did not have an ability to probe the predictive utility of these features within even more proximate time scales (e.g., one week prior to interview). Future research efforts will benefit from collecting data across critical, shortened periods of interest and modeling the resulting predictors within a temporal framework, possibly through application of time-series-based methods such as spectral analysis or differential time-varying effect models (Jacobson et al., 2019). Intuitively, the known dynamic properties of acute SI would benefit from a trajectory-based transformation of factors as these are likely to be more informative compared with a temporal representations that loosely define a period of interest.

Another limitation is one of generalizability. As mentioned in the original analysis of the data, information on socio-demographic data could not be validated given the anonymous nature of the online interviews. Moreover, the data was derived from a German cohort. The degree to which language and culture impacted the variables that were collected and implemented in the modeling pipeline is not clear, nor can a statement be made regarding which linguistic features may be more or less predictive of acute suicidal ideation in one language or culture relative to another. The homogeneity, small size, and uniquely high-risk profile of the sample population, coupled with a dearth of currently available research in the acute SI domain, limits the direct application of this study's findings. Consequently, the results are primarily useful as a new modeling baseline and benchmark for discriminating those with acute SI prediction from those without active SI, but a history of SI. In addition, this work illustrated how aspects of predictive architecture can be applied to infer relationships among risk factors within a cohort.

Progress toward better prediction of acute suicidal behaviors, including ideation, will require more densely collected data from more heterogeneous samples. Luckily, SN platforms lend themselves well to providing such data, thus it is the responsibility of researchers to sample appropriately and take full advantage of the rich contextual information available. Of course, designing predictive models capable of detecting patterns from complex variable structure is just as important. With data in hand, the model performance illustrated by this work suggests that an ensemble machine learning approach is a promising place to start.

CRediT authorship contribution statement

Damien Lekkas: Conceptualization, Methodology, Software, Formal analysis, Writing - Original draft preparation, Writing - Review & editing, Visualization. **Robert J. Klein:** Writing - Original draft preparation, Writing - Review & editing. **Nicholas C. Jacobson:** Methodology, Formal analysis, Writing - Review & editing.

Funding

This work was funded by an institutional grant from the National Institute on Drug Abuse (NIDA-5P30DA02992610).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.invent.2021.100424.

References

Adrian, M., Lyon, A.R., 2018. In: Moreno, M.A., Radovic, A. (Eds.), Social Media Data for Online Adolescent Suicide Risk Identification: Considerations for Integration Within Platforms, Clinics, and Schools, Technology and Adolescent Mental Health. Springer International Publishing, pp. 155–170. https://doi.org/10.1007/978-3-319-69638-6_12.

Aladağ, A.E., Muderrisoglu, S., Akbas, N.B., Zahmacioglu, O., Bingol, H.O., 2018. Detecting suicidal ideation on forums: proof-of-concept study. J. Med. Internet Res. 20 (6), e215. https://doi.org/10.2196/jmir.9840.

Allen, N.B., Nelson, B.W., Brent, D., Auerbach, R.P., 2019. Short-term prediction of suicidal thoughts and behaviors in adolescents: can recent developments in technology and computational science provide a breakthrough? J. Affect. Disord. 250, 163–169. https://doi.org/10.1016/j.jad.2019.03.044.

Birjali, M., Beni-Hssane, A., Erritali, M., 2017. Machine learning and semantic sentiment analysis based algorithms for suicide sentiment prediction in social networks. Procedia Comput. Sci. 113, 65–72. https://doi.org/10.1016/j.procs.2017.08.290.

Braithwaite, S.R., Giraud-Carrier, C., West, J., Barnes, M.D., Hanson, C.L., 2016. Validating machine learning algorithms for twitter data against established measures of suicidality. JMIR Ment. Health 3 (2), e21. https://doi.org/10.2196/mental.4822.

Brown, R.C., Plener, P.L., 2017. Non-suicidal self-injury in adolescence. Curr. Psychiatry Rep. 19 (3) https://doi.org/10.1007/s11920-017-0767-9.

Brown, R., Bendig, E., Fischer, T., Goldwich, D., Baumeister, H., Plener, P.L., 2019a. Brown_Bendig_Instagram_Suicidality_Dataset.xlsx. https://doi.org/10.6084/m9.figshare.7763333.v1.

Brown, R.C., Bendig, E., Fischer, T., Goldwich, A.D., Baumeister, H., Plener, P.L., 2019b. Can acute suicidality be predicted by Instagram data? Results from qualitative and quantitative language analyses. PLoS ONE 14 (9), e0220623. https://doi.org/10.1371/journal.pone.0220623.

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

Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794. https://doi.org/10.1145/2939672.2939785.

Coppersmith, G., Leary, R., Crutchley, P., Fine, A., 2018. Natural language processing of social media as screening for suicide risk. Biomed. Inform. Insights 10. https://doi.org/10.1177/1178222618792860, 1178222618792860.

Coyne, J.C., 1976. Toward an interactional description of depression. Psychiatry 39 (1), 28–40. https://doi.org/10.1080/00332747.1976.11023874.

De Choudhury, M., Kiciman, E., Dredze, M., Coppersmith, G., Kumar, M., 2016. Discovering Shifts to Suicidal Ideation From Mental Health Content in Social Media. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pp. 2098–2110. https://doi.org/10.1145/2858036.2858207.

De Vries, E.L.E., 2019. When more likes is not better: the consequences of high and low likes-to-followers ratios for perceived account credibility and social media marketing effectiveness. Mark. Lett. 30 (3), 275–291. https://doi.org/10.1007/s11002-019-09496-6.

Dettling, M., Bühlmann, P., 2003. Boosting for tumor classification with gene expression data. Bioinformatics (Oxf.) 19 (9), 1061–1069. https://doi.org/10.1093/bioinformatics/btf867.

Emul, M., Uzunoglu, Z., Sevinç, H., Güzel, Ç., Yılmaz, Ç., Erkut, D., Arıkan, K., 2011. The attitudes of preclinical and clinical Turkish medical students toward suicide attempters. Crisis 32 (3), 128–133. https://doi.org/10.1027/0227-5910/a000065.

Flesch, R., 1949. The art of readable writing. Harper & Brothers.

Franklin, J.C., Ribeiro, J.D., Fox, K.R., Bentley, K.H., Kleiman, E.M., Huang, X., Musacchio, K.M., Jaroszewski, A.C., Chang, B.P., Nock, M.K., 2017. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. Psychol. Bull. 143 (2), 187–232. https://doi.org/10.1037/bul0000084.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33 (1), 1–22.

Glenn, C.R., Nock, M.K., 2014. Improving the short-term prediction of suicidal behavior. Am. J. Prev. Med. 47 (3, Supplement 2), S176–S180. https://doi.org/10.1016/j.amepre.2014.06.004.

Grant, R.N., Kucher, D., León, A.M., Gemmell, J.F., Raicu, D.S., Fodeh, S.J., 2018. Automatic extraction of informal topics from online suicidal ideation. BMC Bioinf. 19 (8), 211 https://doi.org/10.1186/s12859-018-2197-z.

Hechenbichler, S., 2004. Weighted k-Nearest-Neighbor Techniques and Ordinal Classification, Discussion Paper 399, SFB 386. Ludwig-Maximilians University Munich, pp. 1–16.

Jacobson, N.C., Chow, S.-M., Newman, M.G., 2019. The differential time-varying effect model (DTVEM): a tool for diagnosing and modeling time lags in intensive longitudinal data. Behav. Res. Methods 51 (1), 295–315. https://doi.org/10.3758/s13428-018-1101-0.

Jacobucci, R., Littlefield, A.K., Millner, A.J., Kleiman, E., Steinley, D., 2020. Pairing Machine Learning and Clinical Psychology: How You Evaluate Predictive Performance Matters. PsyArXiv. https://doi.org/10.31234/osf.io/2yber.

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. Drug Abuse 10.

Kalish, R.A., 1966. Social distance and the dying. Community Ment. Health J. 2 (2), 152–155. https://doi.org/10.1007/BF01420690.

Kleiman, E.M., Nock, M.K., 2018. Real-time assessment of suicidal thoughts and behaviors. Curr. Opin. Psychol. 22, 33–37. https://doi.org/10.1016/j.copsyc.2017.07.026.

Leon, A.C., Friedman, R.A., Sweeney, J.A., Brown, R.P., Mann, J.J., 1990. Statistical issues in the identification of risk factors for suicidal behavior: the application of survival analysis. Psychiatry Res. 31 (1), 99–108. https://doi.org/10.1016/0165-1781(90)90112-i.

Lester, D., 1992. The stigma against dying and suicidal patients. OMEGA 26, 71–75.

Lester, D., Walker, R.L., 2006. The stigma for attempting suicide and the loss to suicide prevention efforts. Crisis 27 (3), 147–148. https://doi.org/10.1027/0227-5910.27.3.147.

Longobardi, C., Settanni, M., Fabris, M.A., Marengo, D., 2020. Follow or be followed: exploring the links between Instagram popularity, social media addiction, cyber victimization, and subjective happiness in Italian adolescents. Child Youth Serv. Rev. 113, 104955 https://doi.org/10.1016/j.childyouth.2020.104955.

Lundberg, S., Lee, S.-I., 2017. A Unified Approach to Interpreting Model Predictions. ArXiv:1705.07874 [Cs, Stat]. http://arxiv.org/abs/1705.07874.

Luxton, D.D., June, J.D., Fairall, J.M., 2012. Social media and suicide: a public health perspective. Am. J. Public Health 102 (S2), S195–S200. https://doi.org/10.2105/AJPH.2011.300608.

Majka, M., 2019. *naivebayes: High Performance Implementation of the Naive Bayes Algorithm in R* (0.9.7) [Computer software]. https://CRAN.R-project.org/package=naivebayes.

Mann, J.J., Waternaux, C., Haas, G.L., Malone, K.M., 1999. Toward a clinical model of suicidal behavior in psychiatric patients. Am. J. Psychiatry 156 (2), 181–189. https://doi.org/10.1176/ajp.156.2.181.

Marchant, A., Hawton, K., Stewart, A., Montgomery, P., Singaravelu, V., Lloyd, K., Purdy, N., Daine, K., John, A., 2017. A systematic review of the relationship between internet use, self-harm and suicidal behaviour in young people: the good, the bad and the unknown. PLoS ONE 12 (8), e0181722. https://doi.org/10.1371/journal.pone.0181722.

Mundt, J.C., Greist, J.H., Jefferson, J.W., Federico, M., Mann, J.J., Posner, K., 2013. Prediction of suicidal behavior in clinical research by lifetime suicidal ideation and behavior ascertained by the electronic Columbia-suicide severity rating scale. J. Clin. Psychiatry 74 (09), 887–893. https://doi.org/10.4088/JCP.13m08398.

Nemesure, M.D., Heinz, M., Huang, R., Jacobson, N.C., 2020. Predictive Modeling of Psychiatric Illness Using Electronic Health Records and a Novel Machine Learning Approach With Artificial Intelligence [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/fhdr5.

Nesi, J., 2020. The impact of social media on youth mental health: challenges and opportunities. N. C. Med. J. 81 (2), 116–121. https://doi.org/10.18043/ncm.81.2.116.

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

Oexle, N., Niederkrotenthaler, T., DeLeo, D., 2019. Emerging trends in suicide prevention research. Curr. Opin. Psychiatry 32 (4), 336–341. https://doi.org/10.1097/YCO.0000000000000507.

Pennebaker, J.W., Boyd, R.L., Jordan, K., Blackburn, K., 2015. The Development and Psychometric Properties of LIWC2015. https://repositories.lib.utexas.edu/handle/2152/31333.

Probst, P., Boulesteix, A.-L., Bischl, B., 2019. Tunability: importance of hyperparameters of machine learning algorithms. J. Mach. Learn. Res. 20, 32.

Reunanen, J., 2003. Overfitting in making comparisons between variable selection methods. J. Mach. Learn. Res. 3 (Mar), 1371–1382.

Robinson, J., Cox, G., Bailey, E., Hetrick, S., Rodrigues, M., Fisher, S., Herrman, H., 2016. Social media and suicide prevention: a systematic review. Early Interv. Psychiatry 10 (2), 103–121. https://doi.org/10.1111/eip.12229.

Roy, A., Nikolitch, K., McGinn, R., Jinah, S., Klement, W., Kaminsky, Z.A., 2020. A machine learning approach predicts future risk to suicidal ideation from social media data. NPJ Digit. Med. 3 (1), 1–12. https://doi.org/10.1038/s41746-020-0287-6.

Rudd, M.D., 2006. Fluid vulnerability theory: a cognitive approach to understanding the process of acute and chronic suicide risk. In: Cognition and Suicide: Theory, Research, and Therapy. American Psychological Association, pp. 355–368. https://doi.org/10.1037/11377-016.

Scocco, P., Castriotta, C., Toffol, E., Preti, A., 2012. Stigma of Suicide Attempt (STOSA) scale and Stigma of Suicide and Suicide Survivor (STOSASS) scale: two new assessment tools. Psychiatry Res. 200 (2), 872–878. https://doi.org/10.1016/j.psychres.2012.06.033.

Sehl, K., 2019. 6 Ways to Calculate Engagement Rate on Social Media. Social Media Marketing & Management Dashboard. April 10. https://blog.hootsuite.com/calculate-engagement-rate/.

Shapley, L.S., 1953. A value for n-person games. In: Contributions to the Theory of Games, 2(28), pp. 307–317.

The National Action Alliance for Suicide Prevention, R.P.T.F, 2014. Suicide Research Prioritization Plan of Action. https://theactionalliance.org/resource/prioritized-research-agenda-suicide-prevention-action-plan-savelives.

Varma, S., Simon, R., 2006. Bias in error estimation when using cross-validation for model selection. BMC Bioinf. 7 (1), 91. https://doi.org/10.1186/1471-2105-7-91.

Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics With S, 4th ed. Springer-Verlag. https://doi.org/10.1007/978-0-387-21706-2.

Wolf, M., Horn, A.B., Mehl, M.R., Haug, S., Pennebaker, J.W., Kordy, H., 2008. Computer-aided quantitative textanalysis: equivalence and reliability of the German adaptation of the linguistic inquiry and word count. Diagnostica 54 (2), 85–98. https://doi.org/10.1026/0012-1924.54.2.85.

Woodruff, S., Santarossa, S., Lacasse, J., 2018. Posting #selfie on Instagram: what are people talking about? J. Soc. Media Soc. 7 (1), 4–14.

World Health Organization, 2009. World suicide prevention day media release: suicide prevention. http://www.who.int/mental_health/prevention/suicide/suicideprevent/en.

World Health Organization, 2014. Preventing suicide: a global imperative. http://www.who.int/mental health/suicide-prevention/world report 2014/en/.