Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-023-02212-w OPEN

Contextual evaluation of suicide-related posts

Mahdi Rezapour[1]✉

Suicide is a leading cause of death in the US. Online posts on social media can reveal valuable information about individuals with suicidal ideation and help prevent tragic outcomes. However, studying suicidality through online posts is challenging, as people may not be willing to share their thoughts directly due to various psychological and social barriers. Moreover, most of the previous studies focused on evaluating machine learning techniques to detect suicidal posts, rather than exploring the contextual features that are present in them. This study aimed to not only classify the posts based on sentiment analysis, but also to identify suicide-related psychiatric stressors, e.g., family problems or school stress, and examine the contextual features of the posts, especially those that are misclassified. We used two techniques of random forest and Lasso generalized linear models and found that they performed similarly. Our findings suggest that while machine learning algorithms can identify most of the potentially harmful posts, they can also introduce bias, and human intervention is needed to minimize that bias. We argue that some posts may be very difficult or impossible to tag correctly by algorithms alone, and they require human understanding and empathy.

[1] Independent researcher, Massachusetts, USA. ✉email: Rezapour2088@yahoo.com

Introduction

Social media has changed how people express their opinions and feelings. However, these platforms are also associated with increased social problems such as bullying, depression (Tsugawa et al. 2015), and suicide (Jashinsky et al. 2014). Suicide has been one of a leading cause of deaths in the United States (Wasserman, 2016), where it costs >$90 billion in 2013 alone (Shepard et al. 2016).

In the US, suicidal thoughts and behaviors are prevalent among young people, with 3.7 million of them considering suicide, more than a million making a plan for it, and more than half a million attempting suicide every year (Health and Services, 2000; Query, 2018). Death from suicide among young people has a profound impact on their loved ones, communities, and society at large (Tal et al. 2017). However, suicidal ideation and thoughts can be prevented by addressing the risk factors or removing the barriers.

Given the importance and the impacts of suicide on society, it is critical for researchers to not only tag posts as suicidal or non-suicidal, but also to better understand the emotions experienced by web users. For instance, suicidal ideations expressed in online posts are expected to be associated with emotions, and thus observing and highlighting those mental states or emotions before suicide could flag those events before they occur. They could also be used to study those emotions and find better solutions to those problems.

Moreover, studying emotions through online posts is especially important due to the existence of barriers to suicide disclosure, such as feeling embarrassed, concerned about receiving unsupportive reactions, fearful about treatment, or worried about losing one's autonomy due to increased monitoring from family members (Hom et al. 2017; Richards et al. 2019). In addition, it has been noted that people reported greater willingness to share their thoughts and feelings online than they would in face-to-face situations (Lenhart et al. 2001).

Information regarding the concerns of the general population reflected by extreme feelings due to suicidal ideation could provide clinicians and other decision-making organizations with crucial information they need for making timely decisions. For individuals experiencing suicidal thoughts and behaviors (STBs), online resources may sometimes provide the only source of social support (Mok et al. 2015). Moreover, emotions can serve as communication functions within both the brain and social groups(Johnson-Laird and Oatley, 1989). Therefore, emotions reflected by online posts can reveal important information about individuals' feelings and emotions and consequently possible suicidal behaviors. Numerous studies have been conducted to evaluate suicidal posts, and this section highlights some of them.Studies used different datasets and methods to identify suicidal thoughts. For example, using data from medical records, information such as demographic, prior self-harm episodes, mental and health diagnosis were used for risk stratification(Tran et al. 2014). In another study, a keyword-based approach for detecting at-risk contents on Twitter was used (Jashinsky et al. 2014). For instance, excluded terms included words such as "cutting myself and shaving". The study also compared at-risk users versus background users.The sample of tweets was manually labeled as "strongly concerning", "possibly concerning", or "safe to ignore"(O'dea et al. 2015). Techniques such as unigram, bag-of-words as features were used and it was found that machine learning performed as well as humans in distinguishing the categories. One of the first steps toward suicide prevention has been named as identification of suicide risk factors (Homan et al. 2014). Natural language processing (NLP) provides a great opportunity to access mental health concerns of web users for further analysis and information.

As discussed, studying emotions in the context of online posts is especially important as suicidality cannot be predicted effectively using standard practice of clinician by asking people in person about their thoughts (McHugh et al. 2019). Thus, this study was conducted to gain a better understanding of the emotions of people at risk and to examine its applicability, in addition to evaluating their performance. The dataset used in this study was obtained from Reddit. In the US, roughly 7% of adults use Reddit as a social networking site (Barthel et al. 2016). We investigated the performance of our models and then extracted some of the misclassified posts to see if they could have been classified correctly by machine learning in any way. In summary, the main objectives of this study were to use the posts from Reddit platform for the analysis to:

- First, see if the algorithm could effectively tag the posts as suicidal or non-suicidal.
- Second, highlight the posts that are misclassified by machine and discuss whether they could have been classified correctly by machine learning alone in any circumstance.
- Third, explore the plausibility of applying machine learning along with human intervention for extracting the concerns within posts to better understand and find solutions to those concerns.

One of the limitations of this study is that we analyzed stationary comments, not online reviews. In online reviews, machine algorithms could access the history of the comments and use that information to make a decision whether a comment is suicidal or not. However, in stationary comments, such as those on Reddit, there is no temporal sequence or rating system that could provide additional clues for the algorithms. Therefore, our study presents a more challenging and realistic scenario for detecting suicidal ideation online.

Method

This section will briefly discuss the implemented techniques of GLMNET and random forest, and then discuss the preprocessing of the dataset.

GLMNET. GLMNET is a function in R that fits a generalized linear model via penalized maximum likelihood. It is known that before doing any training, the data should only include useful features and exclude noisy and multicollinear features. Those possible noisy features can distract the model from training itself by overfitting, so we make sure all those unimportant features are removed.

The following description is based on the work in the literature review (Friedman et al. 2017). Few important steps are taken in the GLMNET model, which make it different from simple generalized linear model (GLM) of the multinominal logit, including feature selection, and removal of multicollinear variables. To be consistent, the same steps were also used before implementing the random forest technique.

Another important aspect is cross-validation to prevent overfitting of this model by using different training sets and changing the test dataset. This way, we make sure the model can predict future instances as well. We had initially >100 features in the dataset. As the algorithm uses the whole dataset given, it is important to make sure that instead of using all dataset observations, it only keeps the important ones and discards others.

In the implemented technique of Least Absolute Shrinkage and Selection Operator (Lasso), if a predictor does not have much

Table 1 Examples of irrelevant noisy texts and extra considered stop words.

Removed keywords	Terms removed from Stop words
"because", "much", "go", "back", "get", "get","m","thing","s", "even","i_was", "get", "things","thing","don't","person","also","just","like"	"myself", "no", "not", "should"

predictive power, the penalty associated with the coefficient will force the coefficient to be zero, which is the same as eliminating it from the model.

Unlike ridge regression, Lasso is more of a variable selection technique. The objective function of Lasso is to minimize the sum of squared errors (SSE) of the model. It is known that the ridge penalty shrinks the coefficients of correlated predictors towards each other, while Lasso tends to pick one of them and discard the others. The GLMNET algorithm employs cyclical coordinate descent by successively optimizing the objective function over each parameter while keeping others fixed, and cycles repeatedly until convergence. When lambda becomes large, it pushes coefficients toward zero to minimize the below equation. Also, in case of a large lambda, coefficient has to be zero, which can be seen from the below equation

$$SSE = \sum\left(\widehat{y}_i - y_i\right)^2 + \lambda \sum_{i=1}^{P} \left|\beta_i\right| \qquad (1)$$

In summary, Lasso model minimizes the residual sum of squares while considering the sum of the absolute value of coefficients. The problem with Ridge method is that while it shrinks some coefficients, it does not make any of them an absolute zero.

It should be noted that before feeding all attributes to the model, feature selection was conducted by an expert to manually remove those unigram features that are not valuable, see Table 1 on the right, or those that lack generality, see Table 1 on the left. To minimize subjective feature extraction, we mainly focused on higher n-grams.

RF. The second implemented technique is random forest, which this section briefly elaborates on. Random forest (RF) combines the outputs of several decision trees to reach a final decision based on voting, which can be used for both classification and regression. As RF is based on many decision trees, it is more intuitive to briefly describe that algorithm.

Decision tree, in simple words, is based on answering questions and making decisions on nodes and splitting the trees. The decision to make a split or a branch can be based on the significance of features on the splits. Although a single decision tree is prone to bias or overfitting, those issues can be addressed in random forest technique due to inclusion of many decision trees.

Each decision tree is comprised of a data sample drawn from training data with replacement or bootstrap sample. Bootstrapped dataset is created from the training dataset by randomly selecting from that dataset, and by selecting a sample more than once. Now we use a subset of features or columns for each bootstrapped dataset. At the node of each tree, the most important feature will be used, and we go further down the tree by considering the remaining features.

The recursion stops when no significant feature is left for splitting. The final decision is based on voting of all trees. It should be noted that some datasets observations are repeated a few times in each bootstrapped and some are not included at all. Those that are not included in the bootstrapped samples are called out-of-bag samples. Out-of-bag samples, then, pass through trees and make a vote to label them.

Data pre-processing. The study was based on social media postings from Suicide Watch, using data from Reddit. To keep only the frequent terms, or terms that appear more frequently, we trimmed our term matrix by including only those terms that occur more than $n = 2{,}000$ times. That value was based on trial and error to come up with meaningful terms and a reasonable number of terms to be used in machine learning algorithm.

As the vocabulary of emotions and other general terms contain words from all categories of noun, verb, adjective, adverb, and pronoun, we broke them down into the roots/stems of those terms and used only those. For instance, consider "loved" and the corresponding adverb of "lovely", which turn into "love". To account for the heterogeneity of the emotions, it is important not to just focus on a particular term by means of unigram but terms in their context.

We did other preprocessing steps for removal of noises in the contents of our posts. Those include, for instance, conversion to lower case, removal of punctuations, numbers, symbols, and hyphens. Before doing more than a unigram, and after stop words removal, we checked the identified items and if they were very general and not related to suicide ideation, we excluded them from the analysis.

Some of those terms include terms highlighted in Table 1. That was especially important to account for the generality of the trained model. In other words, each emotional signal is associated with a particular physiological pattern, where patterns have their own neurochemical basis (Johnson-Laird and Oatley, 1989). At the same time, emotion was defined as eliciting conditions, cognitive evaluation and an action (Frijda, 1986).

It might be said that these seemingly unrelated terms highlight terms that are likelier to be used by users, so they should not have been discarded. In response to that concern, we did not use those terms mainly to account for the heterogeneity of possible emotional terms that could be used to express emotions.

Another reason that we did not use those terms is related to the objective of this study, which was not to solely predict the current observations but to reach a better understanding of the model performance in a general sense, so those terms were discarded to prevent possible bias. In this study, we also did not consider precision and recall but a simple confusion matrix due to non-skewness of the dataset.

Those terms that are removed from stop words and thus kept in the analysis were highlighted in Table 1 on the right hand side. "no" or "not", Table 1 on the right hand side, were kept due to negation of sentence and improving the accuracy of the model. On the other hand, some words, Table 1 on the left hand side, were excluded by an operator due to lack of generality of the trained model.

Again, we considered bi- and trigrams for words as we expected that unigrams would be less robust, and thus many of them were initially discarded from the process by an operator. Feature selection based on GLMNET was considered for only including those features that seemed to be important based on the algorithm and after being evaluated by an operator.

Data. The dataset that we used was collected using Pushshift API, which is a service that provides access to Reddit data. The dataset contains posts from the "SuicideWatch" and "depression" subreddits of Reddit, labeled as suicide and depression respectively. The dataset also contains non-suicide posts from the "teenagers" subreddit.

The original data source of the dataset is Reddit, which is a social media platform where users can post and comment on various topics. Reddit consists of many subreddits, which are communities dedicated to specific topics or interests. Users can join and participate in subreddits that match their preferences. Reddit data can be accessed through its official API or through third-party services such as Pushshift. The dataset covers the period from Dec 16, 2008 to Jan 2, 2021 for “SuicideWatch” posts, from Jan 1, 2009 to Jan 2, 2021 for “depression” posts, and from Jan 1, 2019 to Jan 2, 2021 for “teenagers” posts. The dataset has 232, 150 observations, which is publicly available and can be found on (Kaggle, 2021). The posts were labeled by the authors of the dataset. He collected the posts from different reddits using Pushshift API and assigned them labels based on the category they belong to (Kaggle, 2021)

Results

Machine learning. Initially, 101 attributes were considered to be used in the machine learning technique, considering uni-, bi- and trigrams. Generalized linear model based on Lasso was used for keeping only important features in the model. 20% of observations were used for the test dataset, while 89,989 or 80% were used for training. Based on the feature elimination of GLMNET, the only excluded features were “suicid_thought”, “one_day”, and “relationship”.

Two machine learning techniques of the popular random forest and GLMNET technique were used. As can be seen from Table 2, while RF outperformed based on the suicide category, GLMNET outperformed the RF model for the non-suicide category. However, considering both categories, RF outperformed GLMNET technique.

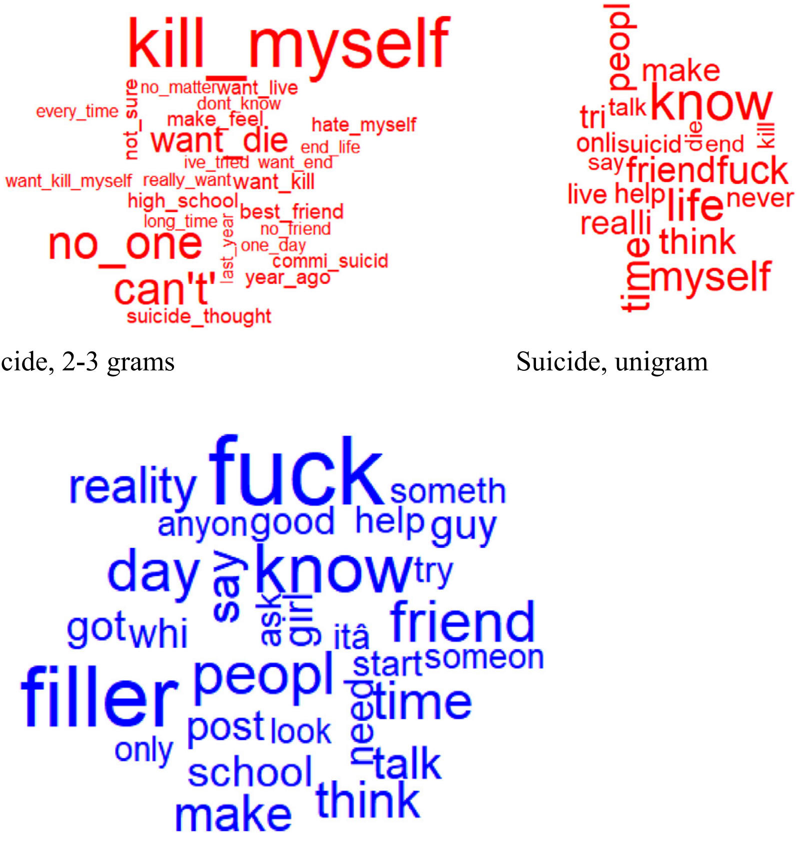
Also, word-clouds of some of the popular terms are included in Fig. 1. It should be noted that to have a better vision, the word-clouds of only single and bi-grams are depicted. Also, as the unigrams were not informative for the non-suicide category, expressing some unrelated terms, that part was excluded from the figure. Although that is still the case for the suicide category, that part is depicted as an example.

As can be seen from Fig. 1, suicide unigram terms are not very informative due to lack of generality, so the majority of those terms were excluded from the analysis. As important features across each singlecategory were considered separately, similar terms could be observed.

Evaluation of misclassified posts. American foundation for suicide prevention (AFSP) has identified characteristics and three risk factors that increase an individual’s risk including (1) health factor (e.g., mental health or chronic pain), (2) environmental

Table 2 Confusion matrix of the test dataset.

Techniques	Context	Predicted, non-suicide	Predicted, suicide
RF	Non-suicide	9972	1293
	Suicide	1688	9545
GLMNET	Non-suicide	10,547	718
	Suicide	2466	8767



Suicide, 2-3 grams

Suicide, unigram

Non-suicide, unigram

Fig. 1 Word cloud of suicide on the top, and non-suicide at the bottom. Various unigrams and bigrams of suicide and non-suicide posts.

factors (e.g., harassment, stressful life events), (3) historical factors (e.g., previous suicide attempts or family history) (Organization, 2020).

In this paper, we did not only focus on the application of machine learning techniques but also on evaluating the contextual features of the posts. To further understand the performance of our model and to highlight some items that are tagged incorrectly, Table 3 is provided. Although it has been discussed that the emotional components of words' meanings do not go outside their meanings (Johnson-Laird and Oatley, 1989), sometimes due to the context of the sentence, terms highlight a meaning that machine learning or even human could not anticipate, ID = 5.

Table 3 for instance, highlights those features related to particular factors such as school, which could elaborate on the cause of suicide. For instance, ID = 14 "a really peaceful way to die…". Which was tagged as non-suicide wrongly, false negative. Also, it should be noted that we expect that the word "die", for instance, was excluded by ML due to the frequent use of the term for both categories. This is one of those posts that are found to be challenging to be tagged by both human and machine.

When digging deeper into other problems, we found that there are cases where users use terms in an ironic or unexpected context like poems. These cases are also very challenging to predict for any machine, like ID = 5, ID = 4 or ID = 7. For instance "this topic is so boring and I am going to kill myself" being called as edge cases ID = 2. These instances are very unlikely to be tagged correctly by any algorithm unless the posts during a time period are considered. Some of texts also incorporate complex emotions which are unlikely to be deciphered by ordinary non-machines, e.g., ID = 8.

Discussion

Although emotions have been named as partly heterogeneous set of events that happen in a person experiencing that event (Johnson-Laird and Oatley, 1989), based on Mandler's theory, those are meaningful external events that are used for labeling arousal (Brown et al. 1962). As a result, by incorporating seemingly meaningful features, it is expected to help the algorithm to identify important features in future incidents.

Suicide posts evaluation could be used as clues for timely intervention of the web user's intent to suicide. Analyzing the online posts is especially significant, as based on our discussion, sometimes human beings might evade expressing their real emotions in real in-person communications.

There are various ways of communicating emotions such as physical movement or social posts. The reasons behind a lack of understanding for both machines and even human being in analyzing some social posts features is partly due to lack of facial/body expression, history of the past posts, which challenge the process of identification of real meanings of posts. That could be observed from the context of the majority of above contextual texts.

Also, it has been observed that the use of different emotions for various feelings is a subjective task, and thus emotions are largely heterogenous and their classification might be impossible (Johnson-Laird and Oatley, 1989). On the same note, analyzing online posts is not just about flagging suicide texts or their classification but also, in a bigger picture, what those posts means and how can they be used to help the population at risk.

To address those two points, we partially incorporated human being knowledge to exclude unimportant features due to lack of generality, such as unigram, and also, we discussed why standard algorithms are unable of tagging some posts due to their nature. However, in this study, we also considered some meaningful unigrams such as "kill, and" suicide". After keeping those unigrams by an operator, we kept those features in our model, if they are kept in the model by means of feature elimination.

We used two machine learning techniques of random forest and GLMNET, where both achieve almost similar accuracies. We found it more meaningful to look at more n-grams than a single gram, especially for suicidal thoughts.

Although the majority of the suicidal posts were associated with regex such as "kill/suicide/death/want_kill", the same holds true for false positives where those terms were used, for instance, in a joke or lyric contexts. Especially for those false positives, there is no language clue for algorithm to tag them correctly unless history of past posts are available.

We acknowledge the limitations of this study. The shortcoming of our dataset was its balanced nature, which is impractical in real-life problems. In other words, in real life problem, it is expected that the majority of posts to be non-suicidal and suicidal posts account for a significant minority of posts. However, we assume that the data was preprocessed and balanced by the Reddit websites.

The texts in the dataset were labeled as 'suicide' or 'non-suicide' based on the source reddit they were collected from. This labeling method was done by the author of the dataset, who used Pushshift API to collect the posts. We acknowledge that this method may not be accurate or reliable, as some posts may not reflect the true mental state of the posters, and some posts may belong to multiple categories. Therefore, we suggest that a mental health professional should proofread the dataset and verify the labels. This is one of the limitations of our study, and we hope to address it in future work.

It has been argued that no single coherent outcome is likely to be a production of words analysis, being referred to emotions (James et al. 1890). That was observed from presented table in the results section, incorporating false positive or false negative. Although we acknowledged the lack of strength of machine learning in tagging those posts, we did feature evaluation of the contextual features of the posts to minimize the error for the current dataset.

That is especially important due to the heterogeneity of the posts. Our analysis is limited to the time of data collection and dynamic evaluations of posts might be needed for real-life problems. For instance, it is unlikely that the same terms or expressions would be used consistently over time or across different platforms or communities.

Instead of binary classification, a finer assessment of people might be needed to distinguish between more concerning, compared with low, or moderate comments. So, for future studies, it is recommended to consider that point. Another advantage of more precise classifications, such as strongly concerning, could be due to immediate actions, which could be made in a timely manner

Another shortcoming of this study was the lack of demographic, e.g., age and gender, and other characteristics for the users posts in machine learning technique. That information is especially important to identify the population at higher need of support, in addition to helping the algorithm to better tag them.

The technique covers limited suicidal expressions, so they suffer from a lack of strength for detection of implicit mentions of suicide such as "i think it is better to take the exist now!" or "i want to take the pain away now". Those possible instances are numerous. So, again, the intervention of non-machine individuals, although seem not very practical, is always needed to prevent dire consequences.

Another limitation is the lack of information regarding the actual outcomes of the attempts. That is also impractical in most cases and web centers should find a way to identify those users that commit suicide in real life and evaluate their posts even after that incidents to prevent future cases.

Table 3 Examples and reasons of misclassification errors on suicide detection.

ID	Boundary	Reason for misclassification	Comment
1	Non-suicide, predicted as suicide	Negation	Im confident in my math but *physic* makes me wat to die i want
2	Non-suicide, predicted as suicide	Entity in wrong context	Gonna kill myself, cant take bullying no longer
3	Non-suicide, predicted as suicide	Entity in wrong context	So I just lost all of my friends because my mind was telling me that they want me to die I dont know what to do like I really want to kill myself because of it and now there all just sorta laughing at me and tbh I see no way out so does anyone have a way out
4	Non suicide, predicted as suicide	Entity in wrong context	"I want to kill myself I want to kill and someone shaine again, Iâ€™m not safe in not a good place but I donâ€™t want to talk my mom she will freak. Hit me"
5	Non suicide, predicted as suicide	Entity in wrong context, joking	"slugsoul says edgy shit on reddit heyyyyyhey hey hello i want to kill myself!!!!!!!!!!!!! haha so funnnuuuuyy!!!!??!!!!!!! so quirky!!!!!!!!!!!!!!!!!!!!!! Âi!!!!!!Âi!!!!Âi!!!!! very original!!Âi"
6	Non suicide, predicted as suicide	Entity in wrong context, joking	"White skinny straight cisgender people be like: life gets better !!!! yea probably for you lol but im fucking tired of being miserable and suffering every fucking day and really fucking want to kill myself"
7	Non suicide, predicted as suicide	Entity in wrong context, poem	"How time flies. I wrote this short poem in January, and 9 months later, I feel better than what i felt writing the poem. I know you know that I stink \nthan whatever you can ever think. \nRight now, my mind, heart, and soul,\nall is sinking down below. \nNow that I'm in this state of mind, \nI think I'll end this life of mine\nnothing's gonna stop from above and below\nCuz now I'm jumping off the window.\n\n\n\n \nJust hold on. What you're going through will get better even though it feels like it will last forever. :)"
8	Non suicide, predict suicide	Entity in wrong context, joking	"Can someone talkkkk???? Feeling lonely af and depressed. If u can then pls donâ€™t be an asshole and say Iâ€™m suicide for stupid reason or smth. Thx in advance"
9	Non suicide, predict suicide	Entity in wrong context, talking about others	"A Question Iâ€™m sorry if this is insensitive to anyone, but why is everyone depressed and suicidal right now? I just donâ€™t understand..."
10	Non suicide, predict suicide	Entity in wrong context, talking about others	"Can someone talkkkk???? Feeling lonely af and depressed. If u can then pls donâ€™t be an asshole and say Iâ€™m suicide for stupid reason or smth. Thx in advance"
11	Non suicide, predict suicide	Entity in wrong context, talking about others	"A Question Iâ€™m sorry if this is insensitive to anyone, but why is everyone depressed and suicidal right now? I just donâ€™t understand..."
12	Suicide, predicted as not suicide	comment does not contain any explicit suicidal ideation or intent from the author, but rather from the author's girlfriend. The machine may have also been confused by the use of slang terms such as "fuckers" and "vent", which may not be associated with suicide in its vocabulary. The machine may have also missed the contextual clues such as the long distance relationship, the lack of support from the parents, and the fear of intervention.	"my girlfriend is suicidalshe posted \"i'll see you fuckers in hell\" to her fb nearly an hour ago. i panicked. she's just reblogged something to her vent (read: depression black and white) tumblr. i don't know whether shes still going to go through with it or not. she also lives 6000 miles away from me in mexico; assuming shes decided not to kill herself, should i still try and contact the police through my countries embassy? if they were to burst in and she decided to go to bed, would i make her situation worse? i should mention her parents arent winning any awards and she definitely doesnt want them finding out. thank you"
13	Suicide, predicted as not suicide	The machine may have failed to recognize the expression "I wanna die" as a sign of suicidal ideation, as it may have been used frequently in both categories. The machine may have also missed the emotional tone of the comment, such as the use of "broken", "love", and "tears". The machine may have also been confused by the spelling errors and informal language, such as "tucked up" and "abd".	"Girlfriend and I are doneIâ€™m broken now.. I love her with all my heart but I tucked up and now i have tears as a man abd I wanna die thank you."

Recommendation. It has been noted that willingness for seeking help from both formal (e.g., health care professional), or informal (e.g., partners, friends) tend to decrease as the risk of suicide increases (Seward and Harris, 2016). That is likely to be related to so many confounding factors that are associated with the thought of suicide. In an ideal condition, the trend of reaching from depression, for instance, to the idea of suicide might be investigated to better understand and help people in need. Understanding the emotions of posters is critical in the field of psychology and for understanding the suicidal behaviors due to its unprecedented consequences,

It is important to understand the underlying cause of emotions for better finding appropriate remedies. Especially some emotions are resultants of self-evaluation in relation to others. For instance,

Table 3 (continued)

ID	Boundary	Reason for misclassification	Comment
14	Suicide, predicted as not suicide	The machine may have failed to recognize the seriousness of the comment, as it may have interpreted the phrase "Why can't I be born again" as a rhetorical question or a wishful thinking, rather than a suicidal thought. The machine may have also missed the contextual factors that contributed to the comment, such as the political situation in Poland, the desire to move to France, and the purchase of a helium canister. The machine may have also been confused by the use of sarcasm, such as "lovely shitty Poland" and "really peaceful way to die".	"Why canâ€™t I be born again in a better countryMy lovely shitty Poland is vetoing the European COVID budget, probably will get kicked out of the eu, all I want to do is to finish school, and move to France for university, I just want to be with my girlfriend, study to become a doctor and help people, meanwhile my country might get kicked from the eu, so I likely wonâ€™t be able to immigrate, and even if I will, itâ€™ll still cost me so much to go university.\nI bough a helium canister, I guess if we get booted out before I finish school, Iâ€™ll just asphyxiate myself, apparently itâ€™s a really peaceful way to die…"
15	Suicide, predicted as not suicide	The machine may have failed to recognize the connection between loneliness and suicide, as it may have assumed that wanting a girlfriend is a common and normal desire for a teenager. The machine may have also missed the emotional intensity of the comment, such as the use of "desperate", "loved", and "die sometimes". The machine may have also been confused by the lack of punctuation and capitalization, which may indicate a low level of effort or attention.	"Being lonely makes me wanna die sometimesI'm (M14) so desperate for a girlfriend. I just want to be loved. I've never been in a relationship before and have been rejected so many times…"
16	Suicide but predict no suicide, high school	The machine may have failed to recognize the bullying and harassment that the comment experienced, as it may have not understood the meaning of the nicknames that were used to mock him. The machine may have also missed the contrast between his outward appearance and his inner feelings, such as being a "relatively happy guy" who smiles, but feeling like garbage and being on the brink. The machine may have also been confused by the use of slang terms, such as "guy", "jokes", and "playing around".	"I Can't Take It Anymore!I'm currently attending high school as a junior. When I was growing up, I knew I was always different from the others. Now, they make fun of me, and treat me like garbage everywhere I go. I'm on the brink, and I need help. They always call me Gayrex, Kylorex, faggotrex, arex, and nigrex just because I have posture like a dinosaur and a larger than average nose. I feel like life isn't worth living anymore. "
17	Suicide but predict no suicide, high school	The machine may have failed to recognize the family problems that the comment faced, as it may have not understood the negative implications of words such as "never listens", "makes fun", "unhappy", and "lectures". The machine may have also missed the discrepancy between his school life and his home life, such as being a freshman in high school who is expected to be happy and successful, but feeling hated by his family. The machine may have also been confused by the lack of punctuation and capitalization, which may indicate a low level of effort or attention.	"I hate my familyI'm a freshman in High School. Reletivly happy guy and tend to smile, from 7:20 to 4:30 everyday. My father is a guy who never listens and always makes jokes about me frowning and makes it seem like I'm just playing around. My sister has a tendency to do whatever she can to make me unhappy whenever. My grandfather is a judging person who always lectures me for how I should be. I don't know what to do anymore…"
18	Suicide, but predict no suicide high school	The machine may have failed to recognize the loss of motivation and hope that the comment expressed, as it may have not understood the meaning of phrases such as "lost all motivation", "treated like trash", and "can't handle". The machine may have also missed the stress and pressure that the comment experienced, such as being in the home stretch of his senior year in high school and having to deal with academic and social expectations. The machine may have also been confused by the use of punctuation and capitalization, which may indicate a high level of effort or attention.	"I canâ€™t take it anymore.Iâ€™m in the home stretch of my senior year in high school and I have lost all motivation to continue living. Iâ€™m tired of being treated like trash by everyone around me, and I donâ€™t think I can handle the stress anymore."

we found that much of the suicide posts were related to school topics, which might be expected as positive relationship with peers have been named important for psychological well-beings (Bishop and Inderbitzen, 1995), whereas peer rejection has been linked to serious problems such as drug abuse, and depression (Laursen et al. 1996). Thus, after preprocessing the posts, and tagging the posts to the terms of interest, e.g., "school", those posts could be extracted to highlight the significance and their concerns. Observations are especially important to be stratified based on various geographic areas, to account for the heterogeneity of the group. After taking various policy making steps, the behavioral shifts should be also closely watched.

Table 3 (continued)

ID	Boundary	Reason for misclassification	Comment
19	Suicide, but predict no suicide high school	The machine may have failed to recognize the frustration and disappointment that the comment felt, as it may have not understood the meaning of words such as "left", "felt like shit", and "nothing to live for". The machine may have also missed the impact of his injury and his inability to skateboard, which may have been a source of joy and identity for him. The machine may have also been confused by the use of slang terms, such as "skatepark", "skateboarding", and "skate video".	"That's itToday I went to the skatepark and everyone was skateboarding and I haven't skateboarded since my last injury and swore off from skateboarding. I only went because I got it in my mind to prove to myself that I can put together a cool skate video and learn some things at the skatepark and I felt like shit and left. Everyone was friends or knew each other and it was like a small community. Last time I felt like shit I was in high school. I can't take it anymore I have nothing to live for now.\n"
20	suicide predicted as no suicide	The machine may have failed to recognize the low self-esteem and self-hatred that the comment expressed, as it may have not understood the meaning of words such as "useless", "walked over", and "hate". The machine may have also missed the emotional attachment and dependency that the comment had on the girl he loved, which may have made him feel hopeless and helpless. The machine may have also been confused by the spelling errors and informal language, such as "peice" and "dont".	Im a useless peice of shitI let myself get walked over by the girl i love because im afraid if i dont let her i will never get to see her. She says im her best friend but i just dont know. I hate everything about myself and i hate myself for thinking that way. Plenty of people have it so much worse then me but im just stuck here feeling sorry for myself
21	suicide predicted as no suicide	The machine may have failed to recognize the suicidal intention and plan that the comment implied, as it may have not understood the meaning of phrases such as "say goodbye" and "not really burden anyone". The machine may have also missed the lack of hope and purpose that the comment felt, which may have made him feel like giving up. The machine may have also been confused by the use of punctuation and capitalization, which may indicate a high level of effort or attention.	Want to say goodbye, but not really burden anyone one last timeSaying it here is a bit of a compromise. Was hoping something came along that made it worth living, hate myself for just clinging on to an existence that gets worse every year"
22	suicide predict as no suicide	The machine may have failed to recognize the self-harm and substance abuse that the comment engaged in, as it may have not understood the meaning of words such as "snorted", "razor", and "thighs". The machine may have also missed the grief and stress that the comment experienced, which may have made him feel overwhelmed and depressed. The machine may have also been confused by the use of slang terms, such as "ambien", "klonopin", and "booze".	Snorted two ambien, some klonopin and drank some booze, took a razor to my thighsI'm a 32 year old married male. My mom is dying from cancer and I cannot deal with it. There is infighting between my wife and my sister due to the added stress. I'm stuck in the middle, incapacitated, incapable of coping with whatever is going on. I can't stand seeing my mother and father cry. The world is just so sad.

Although the stressors could vary from individual to individual, a better understanding of people's concerns, and consequently providing better training and techniques such as counseling could help individuals with a better support they need. For instance, individual support through meeting with parents and students counseling and sustained supports, in case that school bullying is a major stressor, were identified as intensive individual intervention (Swearer et al. 2009).

Identifying the concerns could be discussed by school so better policies and intervention could be taken by those entities. Also, it has been noted that individuals tend to seek support from resources such as social media (Homan et al. 2014), however they should also be trained to seek assistant to take assistant from professional like clinicians

In case of availability, the history of the posts could be used to exclude many tagged posts that are inconsistent or contradictory with previous or subsequent posts. For example, if a user posted a joke about suicide before or after posting a serious comment about suicide, the machine may have difficulty distinguishing between them. The history of the posts could also provide more context and insight into the user's situation and emotions over time.

That was confirmed by the content of snapshots of some of the misclassified posts in this study. Large tech companies that are active in sharing the posts with users might be obliged to allocate a portion of their revenue for better investigating and understanding what exposes people at higher risk of suicide or self-harm. That information is especially important to be shared with the public, such as schools, so appropriate countermeasures for younger people could take place. Evaluation of the posts is not only about the model prediction power but also a better understanding of the population's representative concerns. Those are especially important for policies to better understand what triggers self-injury and suicide ideation and to prevent those incidents. It also highlighted that some misclassified texts are very unlikely to be predicted by any algorithm correctly, so expert interventions, along with machines, are needed.

In addition to these practical recommendations, we also suggest some methodological and technical recommendations for future research and practice:

- More advanced natural language processing techniques that can capture the nuances and subtleties of online communication such as sarcasm, irony, humor, and metaphors.

These techniques may include sentiment analysis, emotion detection, word embeddings, or deep learning models that can learn from large amounts of data and context.

- More diverse and representative datasets that reflect the real-life distribution and variation of suicidal and non-suicidal posts. These datasets may include different sources such as social media platforms, forums, blogs, or chat rooms; different languages or dialects; different demographics or populations; and different time periods or seasons.
- More comprehensive and multidimensional features that can account for the complexity and heterogeneity of emotions and suicide ideation. These features may include not only textual but also visual or auditory cues; not only lexical but also syntactic or semantic aspects; not only individual but also social or environmental factors; and not only static but also dynamic or temporal changes.
- More collaborative and interdisciplinary approaches that can combine machine learning with human expertize and intervention. These approaches may include involving mental health professionals, counselors, or peers in the design, evaluation, or implementation of the machine learning models; providing feedback, guidance, or support to the users who are at risk or in need of help; and developing ethical, legal, and social implications of using machine learning for suicide prevention.

Another direction for future research is to explore the use of more advanced natural language processing techniques that can capture the nuances and subtleties of online communication, such as long short term memory (LSTM) or Bidirectional Encoder Representations from Transformers (BERT). These are types of deep learning models that can learn from large amounts of data and context and generate more expressive and contextualized word embeddings. Compared to random forest and GLMNET, which are based on bag-of-words or n-gram features, LSTM or BERT may be able to better handle the variability and complexity of online texts, such as sarcasm, irony, humor, and metaphors. However, these techniques also pose some challenges, such as requiring more computational resources, more labeled data, more fine-tuning, and more interpretability. Therefore, it is important to evaluate their performance and feasibility for suicide detection in different settings and scenarios.

Data availability

The data is publicly available at https://www.kaggle.com/datasets/nikhileswarkomati/suicide-watch/data.

Received: 19 November 2022; Accepted: 28 September 2023;
Published online: 01 December 2023

References

Barthel, M, Stocking, G, Holcomb, J, & Mitchell, A (2016) Seven-in-ten Reddit users get news on the site. Pew Research Centre, Washington

Bishop JA, Inderbitzen HM (1995) Peer acceptance and friendship: an investigation of their relation to self-esteem. J Early Adoles 15(4):476–489

Brown, R, Galanter, E, Hess, EH, & Mandler, G (1962) New directions in psychology. American Psychological Association, Washington

Friedman J, Hastie T, Simon N, Tibshirani R, Hastie MT, Matrix D (2017) Package 'glmnet'. J Statist Softw 33(1):1–22

Frijda, NH (1986) The emotions. Cambridge University Press

Health, U. D. O., & Services, H. (2000) Substance abuse and mental health services administration. Treatment Episode Data Set (TEDS)

Hom MA, Stanley IH, Podlogar MC, Joiner Jr TE (2017) Are you having thoughts of suicide? Examining experiences with disclosing and denying suicidal ideation. J Clin Psychol 73(10):1382–1392

Homan, C, Johar, R, Liu, T, Lytle, M, Silenzio, V, & Alm, CO (2014) Toward macro-insights for suicide prevention: Analyzing fine-grained distress at scale. Paper presented at the proceedings of the workshop on computational linguistics and clinical psychology from linguistic signal to clinical reality, Association of Computational Linguistics, 107–117 January 2014

James W, Burkhardt F, Bowers F, Skrupskelis IK (1890) The Principles of Psychology. Macmillan, London

Jashinsky J, Burton SH, Hanson CL, West J, Giraud-Carrier C, Barnes MD, Argyle T (2014) Tracking suicide risk factors through Twitter in the US. Cris J Cris Interv Suicide Prev 35(1):51

Johnson-Laird PN, Oatley K (1989) The language of emotions: an analysis of a semantic field. Cogn Emotion 3(2):81–123

Kaggle. (2021) Suicide and depression detection. https://www.kaggle.com/datasets/nikhileswarkomati/suicide-watch

Laursen, B, Hartup, WW, & Koplas, AL (1996) Towards understanding peer conflict. Merrill-palmer quarterly. Wayne University Press, 76–102 January 1996

Lenhart, A, Lewis, O, & Rainie, L (2001) Teenage life online. https://www.pewresearch.org/internet/2001/06/21/teenage-life-online/

McHugh CM, Corderoy A, Ryan CJ, Hickie IB, Large MM (2019) Association between suicidal ideation and suicide: meta-analyses of odds ratios, sensitivity, specificity and positive predictive value. BJ Psych Open 5:2

Mok K, Jorm AF, Pirkis J (2015) Suicide-related Internet use: a review. Aust New Zealand J Psychiat 49(8):697–705

O'dea B, Wan S, Batterham PJ, Calear AL, Paris C, Christensen H (2015) Detecting suicidality on Twitter. Intern Interven 2(2):183–188

Organization, W. H. (2020) Mental health action plan 2013–2020. World Health Organization

Query, CW-BIS (2018) Reporting System (WISQARS)[online]. 2006. National center for injury prevention and control, centers for disease control and prevention (producer). Fildes, Brian and Langford, Jim March 2002

Richards JE, Whiteside U, Ludman EJ, Pabiniak C, Kirlin B, Hidalgo R, Simon G (2019) Understanding why patients may not report suicidal ideation at a health care visit prior to a suicide attempt: a qualitative study. Psychiat Serv 70(1):40–45

Seward AL, Harris KM (2016) Offline versus online suicide-related help seeking: changing domains, changing paradigms. J Clin Psychol 72(6):606–620

Shepard DS, Gurewich D, Lwin AK, Reed Jr GA, Silverman MM (2016) Suicide and suicidal attempts in the United States: costs and policy implications. Suicide Life Threat Behav 46(3):352–362

Swearer, SM, Espelage, DL, & Napolitano, SA (2009) Bullying prevention and intervention: realistic strategies for schools. Guilford Press, Washington

Tal I, Mauro C, Reynolds III CF, Shear MK, Simon N, Lebowitz B, Iglewicz A (2017) Complicated grief after suicide bereavement and other causes of death. Death Stud 41(5):267–275

Tran T, Luo W, Phung D, Harvey R, Berk M, Kennedy RL, Venkatesh S (2014) Risk stratification using data from electronic medical records better predicts suicide risks than clinician assessments. BMC Psychiat 14(1):1–9

Tsugawa, S, Kikuchi, Y, Kishino, F, Nakajima, K, Itoh, Y & Ohsaki, H (2015) Recognizing depression from twitter activity. Paper presented at the Proceedings of the 33rd annual ACM conference on human factors in computing systems, ACM Conference, 3187–3196 April 2015

Wasserman, D (2016) Suicide: an unnecessary death. Oxford University Press, MD

Author contributions

MR is the sole contributor of this study.

Ethical approval

The data was obtained from online resource and from posted comments, so all comments are publicly available.

Informed consent

The data was obtained from online resource and from posted comments, so all comments are publicly available.

Competing interests

The author declares no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Mahdi Rezapour.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.