Social Media + Machine Learning to Offer Clues on Suicide Ideation Concerns

Lakshmi Prayaga, Chandra Prayaga, and Amrutha Gunuru

Abstract Mental health concerns including suicide ideation are a growing concern especially among younger population. Social media also is providing a platform for people with mental health concerns to vent their frustrations and other psychological mental health concerns including suicide ideation. This research is a study on the application of topic modeling, a machine learning algorithm on big data gathered from social media to discover clues that may play a role resulting in adverse outcomes such as suicide ideation.

Keywords Social media Clues · Machine learning · Tweets · Suicide ideation · Topic modeling · Latent Dirichlet Allocation-LDA model

1 Introduction

Social media has become an integral part of our everyday lives, and its impact on our lives is extensive. For the people living in the digital world, negative criticism, mocking, or discrimination on social media tend to lead them into depression which ultimately may drive them toward drastic and unwanted outcomes including suicide ideation or suicide. Through this research, we attempt to decipher information from social media that can point to some signs associated with suicidal tendencies and can assist in detecting these early symptoms that can trigger interventions to prevent this tragic outcome. Latent Dirichlet allocation (LDA), an unsupervised machine

L. Prayaga (✉)
Department of Information Technology, University of West Florida, Pensacola, USA
e-mail: Lprayaga@uwf.edu

C. Prayaga
Physics Department, University of West Florida, Pensacola, USA

A. Gunuru
Information Technology, University of West Florida, Pensacola, USA

A. Kumar et al. (eds.), *Proceedings of the 4th International Conference on Data Science, Machine Learning and Applications*, Lecture Notes in Electrical Engineering 1038,
https://doi.org/10.1007/978-981-99-2058-7_1

learning algorithm, was used in this study on a set of tweets to study signs of suicide ideation. We present the findings in this paper.

2 Literature Review

WHO (2020) [1, 2] reports that there were 800,000 deaths worldwide due to suicide. Suicidal tendencies start during adolescence and continue to grow during early adulthood and old age [3–5]. Suicide ideation is the wish to die, plans actions to die, playing and encouraging thoughts about dying [6, 7]. The pandemic of COVID-2019 has also aggravated the situation. Harmer et al. [6] used surveys to observe the impact of Covid on suicide ideation. This group of researchers reported that during the pandemic suicide ideation was high among "respondents aged 18–24 years (25.5%), minority racial/ethnic groups (Hispanic respondents [18.6%], non-Hispanic black [black] respondents [15.1%]), self-reported unpaid caregivers for adults (30.7%), and essential workers (21.7%)".

Though suicide is rising among young adults it is often not easy to pinpoint the causes and offer interventions. However, social media is a new platform where consumers share the good and the bad in their lives in a free format. Data from social media is also accessible to researchers and academicians to harvest, process, and analyze this data.

Data collected from social media is often called big data due to its veracity, volume, and variety. However, this data is unstructured and is a core property of big data. Big data is not easy to organize and analyze due to the size and the variety of data. In this context, machine learning techniques offer the processing power required to automate the process of data collection, data preparation, data analysis, and data visualizations for big data.

Czeiser et al. [8] conducted a survey to observe the impact of Covid on adults. Regression analysis in *R* was used for this study, and they reported that 5186 people participated in the study. 33% of the population reported anxiety or depression, 29.6% reported PTSD or trauma due to COVID-19, 15.1% reported increased substance use, and 11.9% reported having symptoms of suicide ideation. These symptoms were also more prevalent in younger adults than older age groups (> 65).

After an year of the outbreak of COVID-19, [3] a study on suicidality and COVID-19 in Dec 2021, discovered an increase in the suicide ideation among patients with Covid. The report suggests that prior to the pandemic suicide was a single point of concern but the pandemic with stress due to social isolation, financial needs, depression, and limited healthcare options, made this a dual pandemic of Coronavirus and suicide since the impact of the pandemic is just not only the physical health but on mental health as well.

Suicide ideation and self-harm (SH) among combat veterans is also a concern. However, SI and SH are a result of a complex mix of variables. A recent study [9] used 738 surveys collected from combat veterans. These surveys contained 192 variables. Data collected included variables that were multifaceted and not just related

to mental health or isolation which are usually flags for tendencies leading to SI or SH. In this context, the authors observed that machine learning was able to take ten of the variables that were not related to identifying SH or SI and yet detect the presence of SH or SI with a 75.3% accuracy. This study suggests that machine learning can be a good instrument to analyze large datasets and find predictors that can be used in SI SH risk assessment of patients.

Other researchers [10] observed that there was a strong correlation and confirmation on the association between depressive symptoms and social isolation with suicide ideation. Machine learning algorithms, namely random forest, K nearest neighbor, and neural networks, were used to study this relationship. The results also suggest that women had relatively higher suicidal tendencies 33.3%. It was determined that social influence has both direct and indirect relation to depression.

Another characteristic feature of suicide ideation is that suicidal thoughts may not necessarily be constant. In a study, researchers [6] observed that the characteristics of the suicide ideation fluctuate dramatically. Suicide ideation can be either active or passive. Active ideation denotes experiencing suicidal tendencies now and consciously plan in a specific way to inflict self-harm. However, passive ideation indicates the wish to die and without a specific plan to harm themselves. This study reported that 31% of the people who committed suicide are either an inpatient or outpatient who were treated within a year and about 57% had contact with the mental care professionals. They noticed the rate of clinically depressed is low when compared to the suicides related to new-onset of depression.

Research on suicide ideation using Reddit's users [11] has revealed that the suicide posts had a higher score than non-suicidal posts. Posts on authenticity, anxiety, mentality, and depression had a higher scale and perception, and attention and mind-thinking were on the lower scale. They evaluated the process by using machine learning and deep hybrid learning with an accuracy of 95%. From the results, they derive that the users committing suicide exhibit physiological or agitation.

Our contribution to this literature review is that we study data posted on social media, specifically Twitter to observe if a. findings from posts made on this platform are in line with findings from other forms of data collection such as surveys, electronic health records and b. if posts have any additions to commonly known flags for suicide ideation. The methodology used and results from our study are presented below.

3 Data and Methodology

To study the posts made on social media related to the topic of suicide, 89,519 tweets were gathered and analyzed using topic modeling. Tweets were gathered just by using the search phrase suicide. No other criteria were used for this preliminary study.

4 Data Preparation and Cleaning

The tweets collected were converted to a text document and cleaned for data processing. Data cleaning procedures used included removing retweets, stop words, punctuation marks, hyperlinks, leading blanks, and spaces.

5 Data Exploration with Word clouds

Word clouds are a good way to explore what is contained in a block of text. They are a form of visual representation that show words in various sizes based on the frequency of their occurrence in a block of text. The more the frequency of the word the larger is its size. Figure 1 is a word cloud from the 89,519 tweets that were collected. From this visualization, some of the words that stand out are end, everyone, student, count, iPhone, Android, and lost among many other words. These words can further be attributed to commonly associated sentiments related to suicide. Word clouds thus also provide the sentiments and thoughts of the audience related to a specific topic.

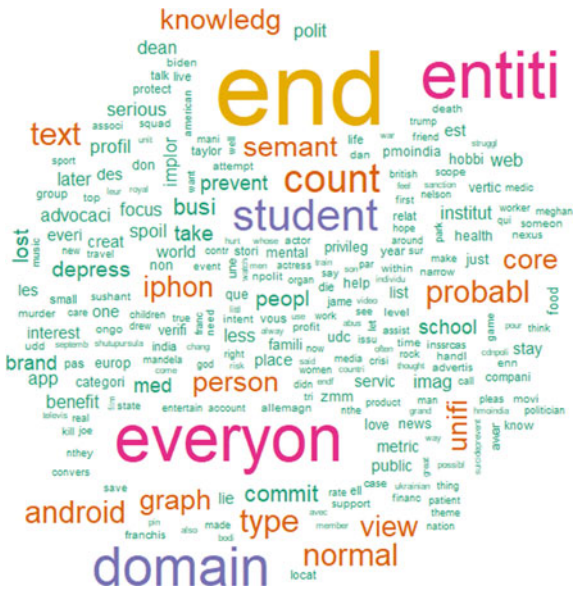

Fig. 1 Word cloud

6 ML—Topic Modeling

To further analyze the sentiments of the audience and words related to a specific topic such as suicide ideation, topic modeling was used to identify major topics from the conversations of tweets. Topic modeling is an unsupervised machine learning algorithm that is used to group sets of words that make up topics from documents. Through topic modeling, we can detect the words and patterns which have similar expressions or meanings. Filtering the top 20 frequently used terms, we can derive other parameters related to the suicide ideation. Figure 2 represents the top 20 words from the body of text, and Fig. 3 shows the words that were used more than 6000 times in the document.

Figure 3 is a graphical representation of the words used more than six thousand times in the tweets collected related to suicide.

```
> freq[head(ord, n = 20)]
   suicid       end descript   entiti  everyon   domain
    69380     48704     41410     29739     29658     23995
taxonomi   student     count      text      type   normal
    23208     18570     16644     12874     12645     12559
 probabl   android     iphon      view     graph knowledg
    12484     11855     11193     10938     10665     10638
     core    semant
    10620     10607
```

Fig. 2 Top 20 words

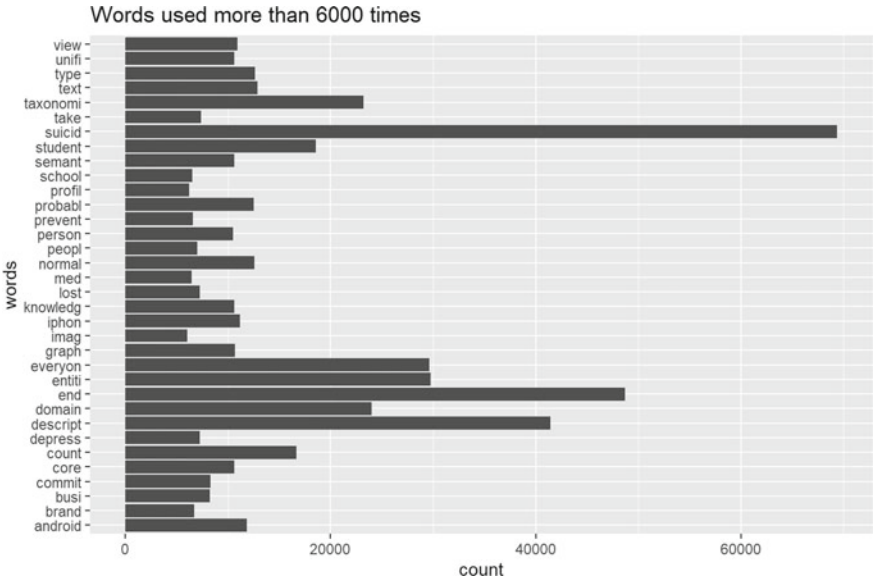

Fig. 3 Words used > 6000 instances

After ignoring some words that have a more semantic relevance than an emotional context, a set of words emerge from Fig. 3 that are of interest and candidates for further analysis. These words include student, end, depression, … At the preliminary level, these words also agree with other studies which identify "depression", "young adults/ students", thinking that the "end", is near as some common words related to suicide ideation.

7 Application of the Latent Dirichlet Allocation (LDA) Model

The LDA in topic modeling builds words per topic and topic per document. The LDA model was applied to the tweets dataset. The model identifies major topics and words that are present in each topic. Figure 4 displays the high frequency words in each topic. The number of topics (3 in this case) were chosen as a starting point number.

To determine the optimal number of topics for a given dataset, the following packages and functions were used. The Package is idatunuing, and the four functions to determine the number of topics are Griffiths2004, CaoJuan2009, Arun2010, and Deveaud2014. Results from applying these methods are presented in Figs. 5 and 6.

Figure 5 is a result of using the four methods in the ldatuning package to discover the optimal number of topics for this dataset. Figure 6 shows the two best applicable methods: CaoJuan2009 and Deveaud2014 methods to determine the optimal number of topics. These two methods were chosen as the most suitable methods are due to the fact that the two methods CaoJuan2009 and Deveaud2014 show a peak and a low at the number sixteen. The other two methods were not very informative for this dataset. Figure 7 is the results of using sixteen as the number of topics and extracting the top five terms per topic.

Figure 8 is a visual representation of the top ten terms from each topic. These words were chosen by the algorithm since they have a high beta value.

Fig. 4 Top words from three topics

```
> top10terms_3
      Topic 1    Topic 2   Topic 3
 [1,] "descript" "end"     "suicid"
 [2,] "entiti"   "count"   "student"
 [3,] "domain"   "text"    "everyon"
 [4,] "taxonomi" "probabl" "iphon"
 [5,] "view"     "everyon" "take"
 [6,] "graph"    "type"    "depress"
 [7,] "knowledg" "normal"  "lost"
 [8,] "core"     "commit"  "school"
 [9,] "semant"   "prevent" "med"
[10,] "unifi"    "profil"  "institut"
```

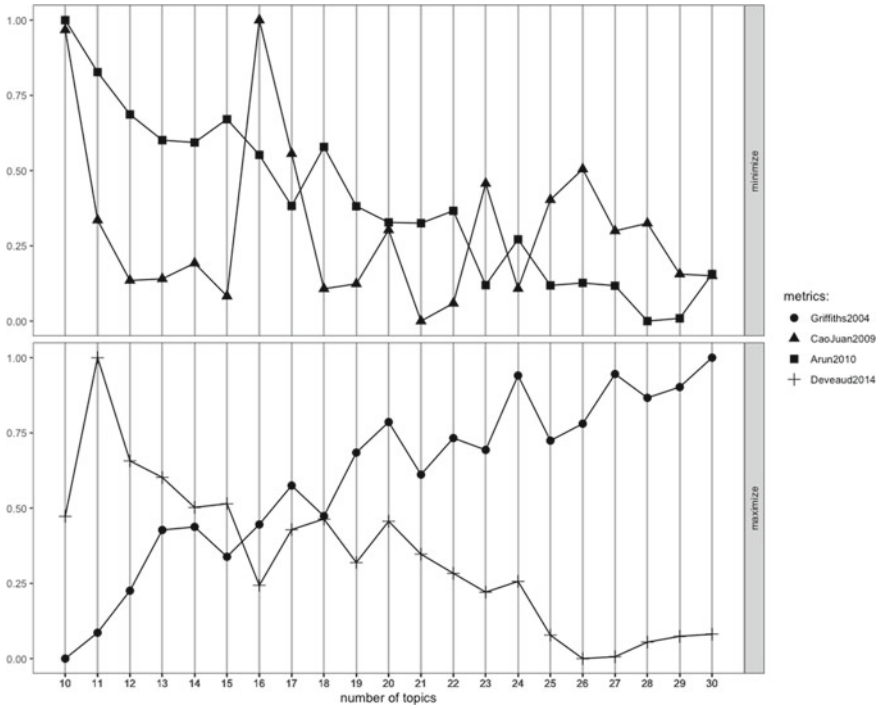

Fig. 5 Four methods to discover optimal topics

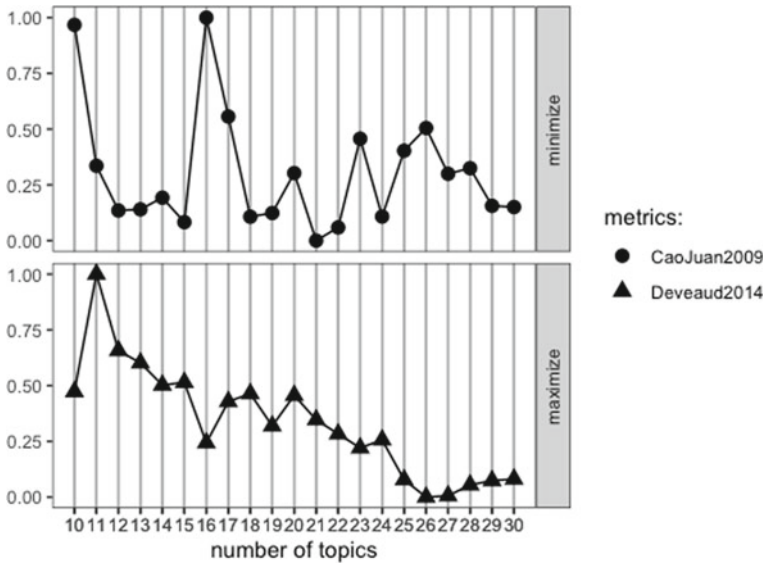

Fig. 6 CaoJuan2009 and Deveaud2014 methods to determine the optimal number of topics

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
[1,]	"take"	"suicid"	"suicid"	"entiti"	"polit"	"stay"	"end"	"suicid"
[2,]	"depress"	"pmoindia"	"commit"	"descript"	"entiti"	"lie"	"everyon"	"que"
[3,]	"account"	"india"	"one"	"busi"	"world"	"suicid"	"iphon"	"someon"
[4,]	"school"	"nexus"	"die"	"domain"	"peopl"	"privileg"	"suicid"	"know"
[5,]	"suicid"	"death"	"don"	"interest"	"descript"	"zmm"	"love"	"une"

	Topic 9	Topic 10	Topic 11	Topic 12	Topic 13	Topic 14	Topic 15	Topic 16
[1,]	"end"	"les"	"student"	"app"	"count"	"android"	"suicid"	"descript"
[2,]	"text"	"est"	"suicid"	"web"	"end"	"suicid"	"prevent"	"taxonomi"
[3,]	"type"	"europ"	"med"	"end"	"profil"	"everyon"	"health"	"domain"
[4,]	"normal"	"des"	"school"	"everyon"	"imag"	"support"	"mental"	"entiti"
[5,]	"probabl"	"pas"	"lost"	"suicid"	"public"	"friend"	"help"	"view"

Fig. 7 Top five words from each topic

Beta values for each word are calculated to show the importance of each word in a topic. The higher the beta value the more important is the word in that topic. Figure 8 is a matrix with a sample of the beta value for words in each topic. The three columns, namely topic, term, and beta list the topic number to which the term belongs and its corresponding beta value. From this sample data, it can be noted that row 3 contains the words friend with a beta value of 1.40e-2 which thus is the most important word in topic number three (Fig. 9).

8 Gamma

Gama values represent the contribution or importance of a topic to the document. The higher the gama value, the more important is that topic in that document. Figure 10 is a sample of ten topics and their relative importance to the document.

As a final step of topic modeling on tweets, the top three terms from each topic were concatenated into string values to offer some descriptions for those topics. Figure 11 corresponds to these concatenated string values from each topic.

9 Discussion

The results from topic modeling provide the following observations.

1. Our research confirms prior work and observes that depression, anxiety, and loneliness are some of the main concerns in young adults that could be factors that are conducive toward suicidal ideation. Topics 1, 3 reflect these themes.
2. We also note other observations such as political and educational contexts are also playing a role in suicide ideation, possibly due to stress caused by these contexts. These would be good topics for further research. Topics 5, 11 suggest these inferences.

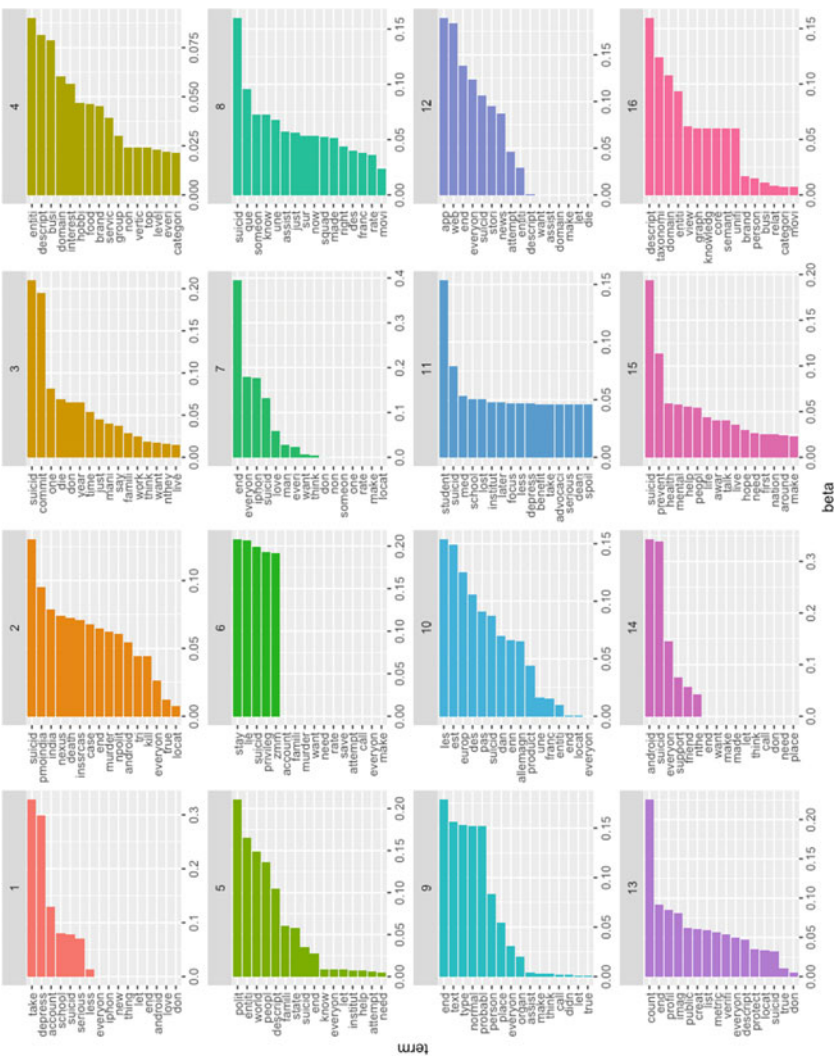

Fig. 8 Graphical representation of words from each topic

Fig. 9 Sample beta values for terms from different topics

```
> beta_topics
# A tibble: 324 × 3
   topic term       beta
   <int> <chr>     <dbl>
 1     1 friend 7.15e-14
 2     2 friend 2.89e-15
 3     3 friend 1.40e- 2
 4     4 friend 3.18e-23
 5     1 peopl  3.63e- 9
 6     2 peopl  6.72e-10
 7     3 peopl  4.90e- 2
 8     4 peopl  1.77e-35
 9     1 suicid 1.61e- 1
10     2 suicid 3.80e- 1
# … with 314 more rows
# ℹ Use `print(n = ...)` to see more rows
```

Fig. 10 Gama values per topic

```
> gamma_document
# A tibble: 232,792 × 3
   document topic   gamma
   <chr>    <int>   <dbl>
 1 2            1 0.0318
 2 3            1 0.0761
 3 4            1 0.0170
 4 5            1 0.00535
 5 6            1 0.0201
 6 7            1 0.0201
 7 8            1 0.342
 8 9            1 0.00535
 9 10           1 0.00535
10 11           1 0.0449
# … with 232,782 more rows
# ℹ Use `print(n = ...)` to see more rows
```

Fig. 11 Concatenation of top three terms from each topic

```
"topic 1 NewTopicName: take depress account"
"topic 2 NewTopicName: suicid pmoindia india"
"topic 3 NewTopicName: suicid commit one"
"topic 4 NewTopicName: entiti descript busi"
"topic 5 NewTopicName: polit entiti world"
"topic 6 NewTopicName: stay lie suicid"
"topic 7 NewTopicName: end everyon iphon"
"topic 8 NewTopicName: suicid que someon"
"topic 9 NewTopicName: end text type"
"topic 10 NewTopicName: les est europ"
"topic 11 NewTopicName: student suicid med"
"topic 12 NewTopicName: app web end"
"topic 13 NewTopicName: count end profil"
"topic 14 NewTopicName: android suicid everyon"
"topic 15 NewTopicName: suicid prevent health"
"topic 16 NewTopicName: descript taxonomi domain"
```

3. Mobile devices such as iPhone and Android phones as noted in topics 7 and 14 are also being used a lot to suggest that people use these devices to post their feelings.
4. There is also discussion on help required for suicide ideation as shown in topic 15.

From this analysis it is can be argued that social media such as Twitter or Facebook offer a wealth of information on difficult topics such as suicide ideation and other mental health illnesses that academic researchers can glean information from. These platforms allow for free communication and some sort of anonymity that allows people to share their thoughts in an unguarded format that enriches the depth of information on the topics under discussion. It is hoped that advances in machine learning can analyze such natural language conversations making up a voluminous chunk of big data and provide informative takeaways to address challenges such as curbing suicide ideation and other mental health concerns.

10 Future Work

Limitations to our research are the number of tweets considered, since we only analyzed 89,519 tweets, the derivations or the conclusions may vary when applied to larger datasets. A future project is to compare the LDA models with other machine learning models and compare the accuracies of different models. A social research problem would also be to consider what factors in political and educational contexts are causing adverse mental health issues.

References

1. WHO. https://www.who.int/health-topics/suicide#tab=tab_1
2. WHO. https://www.who.int/news-room/fact-sheets/detail/suicide
3. Hawton K, Saunders KEA, O'Connor RC (2012) Self-harm and suicide in adolescents. Lancet 379:2373–2382. https://doi.org/10.1016/S0140-6736(12)60322-5
4. Klonsky ED, May AM, Saffer BY (2016) Suicide, suicide attempts, and suicidal ideation. Annu Rev Clin Psychol 12:307–330. https://doi.org/10.1146/annurev-clinpsy-021815-093204
5. Dendup T, Zhao Y, Dorji T, Phuntsho S (2020) Risk factors associated with suicidal ideation and suicide attempts in Bhutan: an analysis of the 2014 bhutan STEPS survey data. PLoS ONE 15:e0225888. https://doi.org/10.1371/journal.pone.0225888
6. Harmer B, Lee S, Duong TVH, Saadabadi A (2022) Suicidal ideation. In: StatPearls [Internet]. StatPearls Publishing, Treasure Island. PMID: 33351435
7. Morese R, Longobardi C (2020) Suicidal ideation in adolescence: a perspective view on the role of the ventromedial prefrontal cortex. Front Psychol 11:713. https://doi.org/10.3389/fpsyg.2020.00713
8. Czeisler MÉ, Lane RI, Wiley JF, Czeisler CA, Howard ME, Rajaratnam SMW (2021) Follow-up survey of US adult reports of mental health, substance use, and suicidal ideation during the COVID-19 pandemic, September 2020. JAMA Netw Open 4(2):e2037665. https://doi.org/10.1001/jamanetworkopen.2020.37665. https://jamanetwork.com/journals/jamanetworkopen/article-abstract/2776559

9. Colic S, He JC, Richardson JD, St. Cyr K, Reilly JP, Hasey GM (2022) A machine learning approach to identification of self-harm and suicidal ideation among military and police Veterans. J Mil Veteran Fam Health. 8:56–67. https://doi.org/10.3138/jmvfh-2021-0035
10. Kim S, Lee K (2022) The effectiveness of predicting suicidal ideation through depressive symptoms and social isolation using machine learning techniques. J Pers Med 12:516. https://doi.org/10.3390/jpm12040516
11. Yeskuatov E, Chua SL, Foo LK (2022) Leveraging reddit for suicidal ideation detection: a review of machine learning and natural language processing techniques. Int J Environ Res Public Health 19(16):10347. https://doi.org/10.3390/ijerph191610347.PMID:36011981; PMCID:PMC9407719