ISSN: 0067-2904

Quantifying Suicidal Ideation on Social Media using Machine Learning: A Critical Review

Syed Tanzeel Rabani, Qamar Rayees Khan, Akib Mohi Ud Din Khanday
Department of Computer Sciences, Baba Ghulam Shah Badshah University, Rajouri, J&K

Received:7/7/2020 Accepted: 14/1/2021

Abstract

Suicidal ideation is one of the severe mental health issues and a serious social problem faced by our society. This problem has been usually dealt with through the psychological point of view, using clinical face to face settings. There are various risk factors associated with suicides, including social isolation, anxiety, depression, etc., that decrease the threshold for suicide. The COVID-19 pandemic further increases social isolation, posing a great threat to the human population. Posting suicidal thoughts on social media is gaining much attention due to the social stigma associated with the mental health. Online Social Networks (OSN) are increasingly used to express the suicidal thoughts. Recently, a top Indian actor industry took the harsh step of suicide. The last Instagram posts revealed signs of depression, which if anticipated could have saved the precious life. Recent research indicated that the public information on social media provides valuable insights on detecting the users with the suicidal ideation. The motive of this study is to provide a systematic review of the work done already in the use of social media for suicide prevention and propose a novel classification approach that classifies the suicide related tweets/ posts into three levels of distress. Moreover, our proposed classification task which was implemented through various machine learning techniques revealed high accuracy in classifying the suicidal posts. Among all algorithms, the best performing algorithm was that of the decision tree, with an F1 score ranging 0.95-0.97. After thoroughly studying the work achieved by different researchers in the area of suicide prevention, our study critically analyses those works and finds various research gaps and solves some of them. We believe that our work will motivate research community to look into other gaps that will in turn help psychiatrists, psychologists, and counsellors to protect individuals suffering from suicidal ideation.

Keywords: Twitter; Suicide; Social media; Machine Classification; Suicide Prevention

I. Introduction

The World Health Organisation (WHO) in its latest report provided detailed statistics about suicidal cases, in which it was mentioned that on an average, suicide occurs every forty seconds [1]. According to this report, more than 8 00 000 people die because of suicide every year, while almost the same number of people attempt suicide [2]. In India, student suicide rate is perturbing, as one student takes this harsh step every hour [3]. In 2016, 9474 students lost their life due to suicide as reported by the Ministry of Home Affairs. Despite the prevalence of this deadly mental health problem, the detection and intervention mechanisms are still in their infancy.

The fact about suicide is that it is not about "all or nothing" situation. An earlier work [4] provided an overview of the process of suicide, as shown in Figure 1. There are various

*Email: syedtanzeel@bgsbu.ac.in

factors that influence the suicide. These factors can be categorised as lateral (weak) and immediate (strong) risk factors [4], [5]. Distal factors do not directly influence the suicide. These include, for example, situations arising due to the lack of employment, loss of a family member, or death of a close friend. The social isolation and the cyber bullying are other factors which influence the suicide but are still considered as distal factors.

The factors which directly influence the suicidal behaviour include all those factors related to the hormonal imbalance, psychological disorder like inability to solve the problems, substance abuse, and other malfunctions of the brain like depression. When these factors come into play, they may lead to the formulation of a suicidal plan, then attempt, and finally action. The other two factors which may increase rapidly the possibility of suicide is the spread of the idea and the availability of necessary methods related to suicide. The social contagion increases the possibility of suicide by regarding it as an acceptable solution to the problems faced by individuals. There are other factors, such as protective factors, which help in preventing the suicide. These include the love for children, purpose to live, religious beliefs, and strength to face hardships.

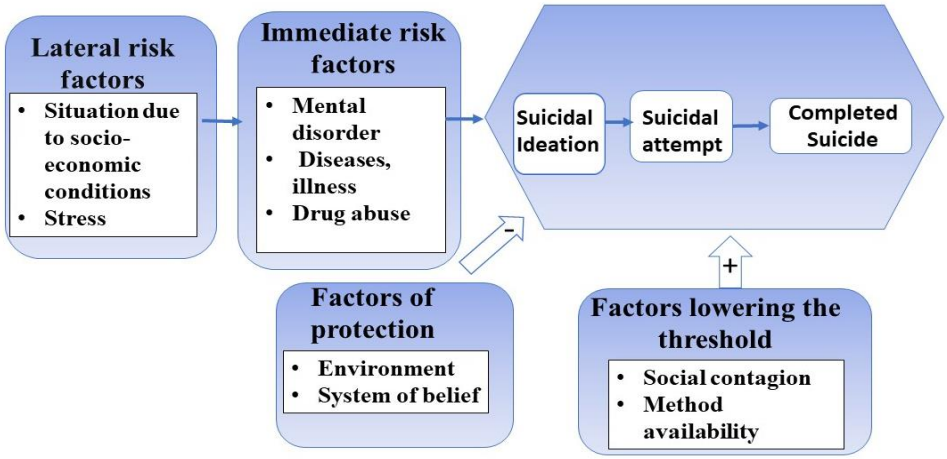


Figure1- Effect of various factors on suicidal process [6]

The Global health observatory data from the WHO indicated that low-income countries have 0.1 psychiatrists available per 100 000 people. In India, is the value is 0.75 psychiatrists per lakh population (i.e. 100 000 people) [7]. It is a well-established fact that people usually feel shy or reluctant about consulting the psychiatrist or any counsellor and do not disclose their plans before committing the act of suicide. As stigma plays a remarkable role in suicidality, clinical interventions for at-risk individuals at large scale become almost impossible. The other issue with this mental illness is the unavailability of diagnostic tests. The prediction of at-risk patients is achieved only through reporting the behaviours by mental health examiners which can include relatives, friends, siblings or psychiatrist. It is reported that 36% individuals who die due to suicide leave a note behind [8]. Research has found that many individuals talk about the suicide before completing it. Jashinsky *et al*. [9] revealed that people do post about "suicidality" on Twitter, which could help in preventing the suicide.

Internet has removed the barriers for communication and allowed to access the required information without being bound to a particular location. Web 2.0 technology further removed the barriers between the information providers and consumers. A number of scientific discoveries can be made by using large scale data analysis like [10-12]. But, for every development, there are challenges associated with it. Analysing data related to internet-based communication for suicide prevention also have many challenges. As information is

freely available, suicide influencers have the opportunity to portray the suicide as a valid and legitimate solution, leading to increase the suicidal contagion. It is also very difficult to detect the influencers due to their anonymous nature. Despite having these limitations, social media also provide some alternatives; for example, people can discuss their suicidal ideation freely, compared with discussing it only with their clinician due to the social stigma. People's attitudes, beliefs, and activities are currently found in searchable archive, thus providing the easy method for analysis. Various soft computing methods are evolving to provide a quantification mechanism and deal with the ambiguity of linguistic terms popularly being used in human statements [13].The longitudinal nature of social media, by using continuous observations, helps to analyse the changing behaviour of users, which was otherwise impossible with traditional approaches.

A good amount of research has been published about detecting the suicidality on the Social Networking Sites (SNS) [14], [15]. Therefore, the main motive of this systematic review was to confirm the feasibility of using the social media for the detection and prevention of potential suicidal actions. Our research work contributes to the already existing literature in the following ways:

(1) A critical review was constructed by highlighting the limitations of various research articles published in the domain of suicide prevention.
(2) The limitations and research gaps found are provided below briefly in a Tabular form for a ready reference to the researchers working in this domain.
(3) A novel dataset of more than 12534 suicidal posts was extracted from various social networking sites like Twitter and Reddit and further annotated into various levels of suicide concerns with the aid of an annotation scheme developed in consultation with psychiatrists and psychologists.
(4) A multi-classification scheme is proposed that distinguishes between various levels of suicidal ideation.
(5) A hybrid feature engineering mechanism is developed that helps in extracting the relevant features for the multi classification of distress.

The paper consists of various sections. Section II discusses the methods and domains of Suicidal Ideation Detection. Section III discusses the data collection techniques used in the existing work. Section IV describes the extensive literature survey. Section V discusses the findings and limitations of some recent papers. Section VI describes the conclusions and section VII discusses the future work in the area.

II. Methods and Domains of Suicidal Ideation Detection

Due to the enormous increase in suicidal cases, researchers are studying this mental health concern from different perspectives. The traditional approach is through clinical methods by having clinical patient interaction [16]. Another approach is by analysing the suicidal text on social media using machine learning techniques [9-11]. Studying social media for suicide prevention has gained much attention as people express their feelings freely due to its anonymous nature. There has been a boom in social network sites. The different events that occur on social media include the location and time of that event, which can be used to extract the necessary information related to that event. In this section, we introduce our systematic review about the use of OSN for suicide and related mental health detection and prevention approaches. The different techniques and domains of suicide detection were elaborated by Shaoxing [17], which are summarised and described in Figure 2.

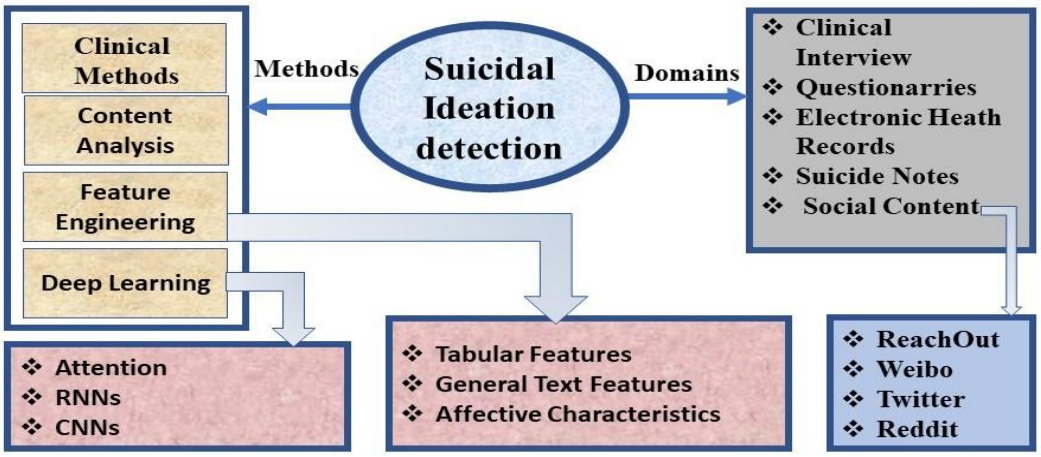


Figure 2-Categorization of suicidal ideation detection methods and domains [17].

III. Data Collection Techniques

Colombo *et.al.* [18] collected data related to suicide from ten websites. Two thousand posts were collected and annotated. After analysing the posts, some unimportant terms were discarded using term frequency-inverse document frequency (TFIDF) technique. The TFIDF generated 62 terms that helped the annotators to categorize the posts in seven categories.

Choudhury *et al*. [19] used crowdsourcing to collect the data on social media related to the Major Depressive Disorder (MDD). Crowd workers participated enthusiastically and were asked to share their Twitter account, guaranteeing that their privacy will be maintained and data will be mined anonymously using computer software. The authors used scales like CES-D which consists of 20 questions to measure the depression level of population. They used the diagnostic and statistical manual defined by the American Psychiatric Association (APA) to measure the depression symptoms.

Burnap *et al.* [20] manually created a vocabulary of suicidal terms generated from various websites and collected the relevant posts by using those terms. The posts were annotated and classified into different classes based upon their context.

Birjali *et al*. [21] also constructed a manual vocabulary consisting of various themes related to suicide for data collected from Twitter using Twitter API. The authors used Twitter4J application programming interface (API) to extract the required data. The Figure 3. describes the flowchart of the approach through which the authors collected the tweets from Twitter.

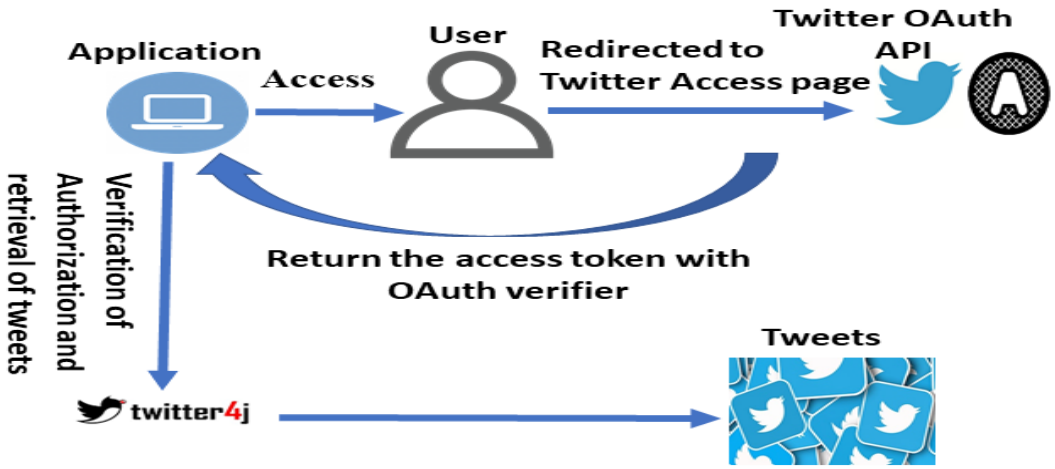


Figure 3. Architecture used for collection of tweets for training the suicidal prediction model [21].

IV. Literature review

Discovering people with suicidal ideation via SNS is a difficult and very tedious task that needs a remarkable amount of research. In this article, a critical review was constructed and different research gaps were identified. Research papers were selected based on the answers to the following three questions:

What are the different ways in which suicidal thoughts are communicated?

What is the power of social media to reach and identify the at-risk people for suicide prevention?

What are the experiences of all those people who used the social media platforms for suicide related purposes?

The various keywords that form the basis of the research were (suicide OR depression OR anxiety OR major depressive disorder OR suicidal ideation OR self-injury) AND (online OR Twitter OR Facebook OR Reddit OR Myspace OR social networking sites OR Orkut). After the thorough search, only the articles from the previous 10 years that had a preventive focus were retained. Then, these articles were divided based on similarity of the problem into various clusters. Various articles in each cluster were studied year wise and their limitations were noted down. Although we studied all the contents of each article, but more focus was given to the future scope of the article, to reach an idea of the questions that remained unanswered.

The various gaps were highlighted so that they may be addressed in future studies. The limitations of recent papers are listed in Table 1.

K. M. Harris *et al*. [22] focussed on the vulnerable individuals. They investigated uncertain and indistinct connections between suicidality and explicit web use. Data was collected through online survey from the users experiencing suicidal ideation and also from non suicidal users. Various behaviours related to suicide, symptoms of depression, help seeking, building relationship online, and online behaviour were determined. It was found that users like to communicate with people without discovering their identity and do not wish to seek help from the family members and mental health professionals. The researchers reported that the majority of persons feel less estranged online and less suicidal afterwards. It was less likely that the at-risk suicidal users seek help from others offline for their problems. It was also found that users typically prefer solving their problems by including the method of avoiding persons. The method of survey used here indicates that in the traumatic situations, intervention by researchers in not possible.

Wen-Cheng *et al.* [23] focussed on the intervention of high-risk individuals so that they can be detected early to prevent them from taking the harsh step of suicide. Brief symptom rating scale-5 known as "Feeling Thermometer" was introduced on facebook as a survey. The survey results were scored in four levels as normal, mild emotional distress, moderate distress, and severe mood disturbance. This investigation exhibits that Facebook can help or assist counsellors or suicide prevention gatekeepers to execute their tasks. The application gives a straightforward mental test and gives some valuable remarks with the outcomes so as to achieve early discovery and early treatment. The self-destructive inclination perception framework could be coordinated with other physical information; for example, clinical records, wellbeing records, diet records, family ancestry, and restoration frameworks, into a total personal health record (PHR) framework.

Moreno *et al.* [24] evaluated Facebook status of college students who were diagnosed of having depression or major depressive episode according to the Diagnostic and Statistical Manual (DSM) criteria. The researchers evaluated two hundred profiles. Statistical calculations were performed by SATA version 9.0. Categorical comparison was conducted using persons x^2 tests. Normalization was achieved to the distribution of friends. On

examining the collected data, it was found that data consists of a female population (43.55%) with an average age of twenty years. A percentage of 25% of the analysed data consists of those profiles having the indications of misery, while 2.5% met the rules for major depressive episode (MDE). The proprietors of profile were bound to have wretchedness if there was more than one online reaction from their companions to an announcement unveiling despondency manifestations or if Facebook was utilized all the more and as often as possible. Only college students were focussed on in this study; hence, it will be unwarranted to generalise the results to other young adult populations. Validity of the displayed status is unclear. Formal determination cannot be made without clinical setting, including term, seriousness and recurrence of the showed manifestations. Relationship between the showed sorrow manifestations and self-announced despondency side effects ought to have been assessed utilizing a clinical scale.

Another study [25] examined the picture of a person who had cut his wrist and posted it on Weibo, a Chinese SNS. It was found that, besides the large number of positive responses on that picture, many reposts revealed that social networks could be used to detect the at-risk posts and help intervene in suicidal attempts. A case study [26] also provided similar findings. It discussed the case of a person (client) who posted a suicidal content on Facebook indicating his suicidal ideation. The clinician, after discovering the content of his client, took necessary actions leading to his hospital admission. The authors in this case study also discussed the ethical challenges related to the intervention, which include issues related to informed consent, judgement of clinician, confidentiality, and privacy. Despite having these challenges, the study reflected how these public posts on social media about suicidality would lead help in intervention.

Jashinsky *et al*. [9] identified the risk factors related to suicide on Twitter by matching geographical rates of suicide with the statistical data. The extracted tweet set was refined by removing the sarcastic tweets and geolocated tweets were analysed. After the data filtering, tweets were grouped in terms of state and rates of at-risk tweets were calculated by finding the baseline values and departure. Suicide related tweets were compared against the rates of actual suicide information, as provided by national data. This study could not be generalized as the number of tweets was not sufficient and the study could not differentiate between levels of suicidal intent. While there has been a breakthrough in collecting data from depression related illnesses, there is a need to build some real time apps for data collection, which will help in assessing the suicide risk of the user and could alert the well-wishers of the concerned user.

Kailasam *et al*. [27] analysed the case history of a person who had a Major depressive disorder (MDD) and non-adherence, with no history of suicide attempt or impatient treatment. The patient was counselled properly after determining the suicidal ideation online. Clinicians emphasized that the social media can be used to prevent potential suicidal individuals by analysing their content on social media using machine learning techniques. Besides having such power, researchers find it very difficult to identify or analyse emotions of people using these posts, as the context of emotions may be also related to various specific events or situations.

Cash *et al.* [28] investigated MySpace SNS to locate the various manners by which young people use to remark on their self-destructive contemplations and goals. The study analysed various comments that were collected from the adolescent profiles, having age between 13 and 24 years. Various analytical techniques were used at different phases of the model. The results showed that the relationship subtheme represents most of the suicidal comments. This study was conducted only on MySpace despite the fact that many popular SNSs, like Twitter and Facebook exist. Seriousness of comments was determined by only two coders. No objective measure was used. There was a lack of demographic data; hence the collected

sample did not allow the generalisability of results. Associations between demographic variables and suicide related comments were not studied.

O'Dea *et al.* [29] examined the risk of tweets using machine learning approach**.** The target was to design the classifier that could help in automating and replicating the accuracy of human coders. Machine learning algorithms were implemented using Scikit-Learn toolkit. Feature representation was performed with different variants, including unigrams, TFIDF, and Filter. Support vector machine and logistic regression were tested with different variants of the feature space discussed above. The results from the combined data were analysed, 14% of all these tweets were detected to be strongly concerning, 57% were possibly concerning, and 29 % were treated as safe to ignore. Agreement rate value among coders for the combined dataset was found to be 74. SVMs with no filter TFIDF was found to be the best performing algorithm. The research can further be explored by increasing the phrases related to suicide to guarantee more articulations of suicidality. Offline measures need to be taken in consideration for validating the risk. Various approaches could be used by consulting the well-being wishes, investigated through questionnaires and through clinical consultations. The users who have died of suicide need to be analysed, which can provide the authentic risk of suicide. There is a need to understand the context of the tweets by including various forms. This may be within the post of the tweet by including images, hashtags, emoticons, and retweets, or it may include the previous tweets and replies of followers within the account of the Twitter user. Replies help to understand the risk of the user in a better way as it can be assumed that whoever replies would know the user personally. External context can include the offline social analysis and the emotional state of user.

Coppersmith *et al*. [30] examined the data of young twitter users having age between 15 to 29 years, who previously attempted suicide. Analysis of patterns in languages and emotions was provided in detail. Human annotation analysis was performed to examine the tweets of suicide attempters, their demographic information, etc. Linguistic differences in dataset were visualized using Venn clouds, wherein the language was examined at the level of tokens**.** The methods that provide scores at a per tweet level were considered. Character n-gram language models were used, which was then followed by logistic regression via scikit-learn tool for classification purposes. The results showed that the number of anger and sadness tweets from the persons attempting suicide was remarkably higher before the suicide attempt than afterwards. It was found that there was an increase in the sadness level in tweets before the suicide attempt, with a rise in levels of rage and sorrow in tweets after a week of a suicidal attempt. Fear and disgust tweets were in line with neurotypicals before the attempt, but they decrease below the levels of neurotypicals afterwards. Loneliness tweets were consistently low before the suicide attempt, which were further decreased afterwards. It was reported that people who attempt suicide generate a large number of tweets before the attempt, indicating a call for help. Various established metrics were used to validate the classification the classification results with the existing research on suicide. The data explored here is primarily from women aged between 15-29. The behaviours, motivations, and stressors of users selected are likely significantly different from those of other at-risk groups (transgendered individuals or middle-aged males).

Colombo *et al*. [18] investigated the connectivity and correspondence attributes of those clients that mean to do self-destruction. ocial graphs were analysed to obtain the first hand information about the networks between users. The graphs were obtained using the identification of friends, mutual friends, followers, and sharing of tweets by the suicidal users. The results showed that suicidal users help and communicated with each other by having the follower or following relationships. Moreover, there were 75% of mutual / common links between suicidal users as shown by the social graphs. Other findings by researchers showed up to 42% of reciprocity. On investigating the communication, it was found that the average

value of sharing the suicidal content was higher than that found in other earlier studies. This gives an indication of higher contagion effect. The research could be further extended by working on the limitations which arise due to the smaller dataset used in this research. Further examination might be done on one stage away neighbours, for example, by breaking down the companions of companions and retweeters of retweeters. Examination could be reached out to more than one-bounce away neighbours (companions of companions, retweeters of retweeters). Deep understanding can be done by analysing the age and gender of these type of suicidal users and their social networks.

Moulahi *et al.* [31] leveraged the Conditional Random Fields in tracking the suicidal ideation on social media. Twitter API was used to collect the data by using the keywords that were recognised as risk factors by the American Psychological Association Psychological and Emotional Features, which were also used to understand the context of suicidal ideation. Contextual features were used to observe the posts during the previous session. The authors analysed the emotional changes over time. A framework "DARE TO CARE" was proposed and evaluated with different Machine Learning Algorithms.

Joseph *et al.* [32] created a prediction model for higher risk users by studying various risk factors, such as anxiety, hopelessness, and depression, using data mining techniques. Risk factors were calculated using different metrics of psychology. The various measures which were applied included Beck hopelessness scale, depression scale, hospital anxiety scale, and suicidal ideation subscale. The six data mining algorithms for classification were compared so as to predict the suicidal behaviour. Among the six algorithms, regression was found to have the highest prediction accuracy.

Vioules *et al.* [33] introduced a new approach that measures the warning signs of suicide, detects the posts that contain the content related to suicide, and automatically identifies the sudden changes in the user behaviour. By using research from the field of psychology, behavioural features were designed and developed that measure the level of risk of a person concerned with online on Twitter behaviour. Two classes of behavioural features, namely user-driven and post-driven, were established. The text of the post is addressed by using text score, which holds the crucial information related to an individual's current mood and mental health. Classification of the text is done through the two approaches of NLP and Distress classifier. The results indicated that NLP text-scoring approach effectively isolates the tweets that have trouble related material and acts as a ground-breaking contribution to the suicidal ideation framework. The full methodology of change of emotion detection was performed only on two Twitter users. Many users' timelines need to be included for further testing. The testing would be better if different age groups of users having mental distress are examined. Different levels of distress classes could be refined by having fine grained classes of emotion like anger, fear, sadness, etc.

Noureen *et al*. [34] discussed the various classifiers in classifying psychotic behaviour of individuals. Data collection was done using Facebook API. The posts were pre-processed using various techniques like tokenization, removal of stop words, stemming the features, etc. The results showed that no single method can be mentioned as a benchmark for user psychotic behaviour classification. The various other machine learning techniques can be used and implemented with proper validation for better accuracy and results.

Shahreen *et al*. [35] focused on Twitter by analysing the text, using machine learning and neural networks. For neural networks, three types of weight optimizers were used. The authors were interested in obtaining the Tweet Id and Tweet Text. Feature extraction was done by following the two approaches of Count Vectorizer and Tf-Idf. Support Vector Machine (SVM) attained an accuracy of 95.2% while Neural Networks attained 97.6%. In future multilingual and multiple social network sites need to be analysed. Hybrid approach

can be also analysed and applications should be developed to analyse the users of suicidal intent in real time.

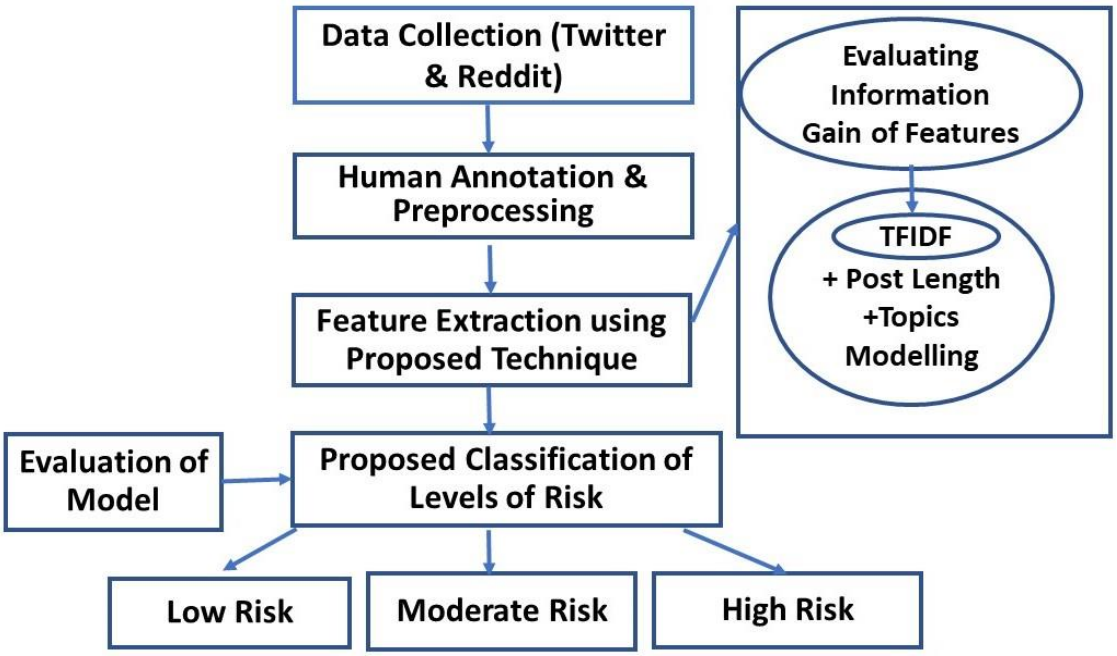
Chiroma *et al*. [36] determined the performance of machine learning algorithms in identifying the suicide related content on Twitter. Annotated data by P. Burnap [37] was used. The researchers collected a total of 2000 tweets and categorized them manually into seven classes concerned with suicide in any way. The first class, named as "Suicide", included all those tweets that reflect suicidal ideation in a strict sense. The other categories included the Campaign, Flippant, Support, Memorial, Reports, and others. All the tweets that had an agreement rate of lower than 75% were removed. The text was transformed by using standard pre-processing techniques. Four machine classification techniques were used. The experiment showed that the Random Forest technique results has the best precision and F measure values for the flippant class, while Naïve Bayes has the best recall values for the suicide class. For the multi class dataset of the non-suicidal classes, the SVM has the best precision score, Naive Bayes has the best recall score, and decision tree has the best F measure. The best value for the flippant class is obtained by decision tree F measure of 0.446. F measure and accuracy were used to compare the classifiers, by comparing the Binary and Multi Class datasets. It was found that the decision tree has a best F measure and accuracy for the multi class data

Table 1-Limitations found in some recent papers related to suicidality

Author	Title of Paper	Year of Publication	Classifier used	Limitations
Moreno *et al.* [24]	Feeling bad on facebook: depression disclosures by college students on a social networking site.	2011	No classifier used	(1) Web profiles from only one SNS websites (Facebook) were analysed (2) Validity of displayed status is unclear. Formal diagnosis can't be made without clinical context, including duration, severity and frequency of the displayed symptoms.
Won *et al*.[38]	Predicting national suicide numbers with social media data.	2013	No classifier used	(1) All variables that were having association with suicide were not included in this model
O'Dea *et al*.[29]	Detecting suicidality on twitter	2015	(1) Support vector machine (SVM) (2) Logistic regression (LGR)	(1) Less suicide-related key words were used for extraction (2) Offline measures were ignored, for example by consulting family and friends, using questionnaires or by clinical consultations. (3) Context of tweet was not properly analysed as retweets; emoticons were not analysed
Colombo *et al*. [18]	Analysing the connectivity and communication of suicidal users on twitter.	2016	No classifier used	(1) Analysis could not be extended in this research to more than one-hop away neighbours (friends of friends,

				retweeters of retweeters). (2) Demographic characteristics were not analysed
Coppersmith *et al.* [30]	Exploratory analysis of social media prior to a suicide attempt.	2016	Logistic regression	(1) The cases haven't been verified, though validated with existing research on suicide (2) The data explored here is primarily from women aged between 15-29. (3) All people investigated here survived their suicide attempts, so there may be a systematic difference between those in our dataset and those who die by suicide. Thus, this study can affect the generalizability of findings.
Haung *et al.* [39]	Exploring timelines of confirmed suicide incidents through social media.	2017	No Classifier used	(1) When comparing the patterns of users grouped by their reason for committing suicide, more variation in social media behaviour was found. (2) Comments, reposts and likes were not explored. Those interactions could be examined to see their influence on the trajectory of user's suicidal timeline.
Vioules *et.al.* [33]	Detection of suicide-related posts in twitter data streams.	2018	(1) Random forest (2) Sequential minimal 0ptimization (SMO) (3) J48 tree (4) Simple logistic	(1) The full methodology of emotion detection was performed only on two Twitter users. Many user timelines need to be included for further testing. The testing would be better if different age groups of users having mental distress are examined.
Chadha *et.al.* [40]	A survey on prediction of suicidal ideation using machine and ensemble learning.	2019	(1) Multinomial naïve bayes (2) Bernoulli naïve bayes (3) Decision tree (4) Logistic regression (5) Support vector machine (6) Voting ensemble (7) Adaboost ensemble (8) Random forest	(1) Precision value was not more than 50% and recall value was less than 50%, indicating false positive and false negative alarms. (2) Only Binary classification is performed which does not provide the real stress information. (3) Data used in training the machine learning model is very few. (4) Less number of

				features are used to train the machine learning model.
Roy *et al*. [41]	A machine learning approach predicts future risk to suicidal ideation from social media data.	2020	Random Forest model using neural network outputs	(1) Data from only Twitter was included in the study. (2) Suicidal behaviour and suicidal thoughts were not studied, which are the strongest indicators of suicidal ideation.
Zheng *et al*. [42]	Development of an early-warning system for high-risk patients of suicide attempt using deep learning and electronic health records.	2020	Deep neural network having input layer of 117 dimensions, 3 hidden layers, and a scaler output layer.	(1) The study didn't include uncoded suicide attempt data. (2) Mortality associated with subsequent suicide attempts were not included in the study.



V. I. Proposed Methodology

The methodology adopted for detecting the suicidal tweets is based upon the gaps found in the literature and consultation with mental health experts. It was found that there is a need of a machine learning model that could classify the suicidal posts into various levels of distress, such that people falling in higher levels of distress could be identified on priority and lives could be saved. The overall picture of the methodology is shown in Figure 4.

Figure 4. Proposed methodology for identification and classification of levels of suicidal risk.

The adopted methodology consists of five steps; (1) Data collection, (2) Human annotation and pre-processing, (3) Feature extraction using the proposed technique, (4) Proposed

classification of levels of risk, (5) Evaluation and validation. Data was collected using Twitter API. A total of 40000 tweets and Reddit posts were extracted in a span of one year. Out of the whole data, 12534 tweets and Reddit posts were retained after discarding the irrelevant posts. The posts were annotated using a novel classification scheme, as discussed in Figure 4, developed in consultation with mental health experts. It was found that people with high-risk use the words like "wish," "want" in a post as "I wish to kill myself right now". People with low risk use the words like "feel" in their post as "I feel I am drowning". The tweets and posts were refined using a standard pre-processing technique [43]. A novel feature engineering mechanism was developed that extracted the relevant attributes from the dataset which were fed to the machine learning model for the classification of suicidal risk. Motivated by an earlier work [44], we extracted the features using term frequency inverse document frequency (TFIDF), Post length, and Topic Modelling. We then improved the feature extraction mechanism by the use of the information gain algorithm, along with the ranker search method to find the most significant features. The threshold for discarding the less significant attributes is set to 0.0020. The maximum limit of the feature set is set to 424. TFIDF is a technique that assigns the score to the particular feature, based upon its presence in the particular document and whole corpus. In the context of our problem, TFIDF was described by the following equation:

$$tfidf(f,p,D) = tf(f,p) \times idf(p,D) \quad (1)$$

$$idf(f,D) = \log \frac{|D|}{1|\{p \epsilon D: f \epsilon p\}|} \quad (2)$$

where f denotes the word as a feature, p denotes the post from which features are extracted, and D represents the document space (set of all posts and tweets)

The post length is used as a feature, as the length of suicidal and non-suicidal text significantly varies. Topic features are also used to extract the 10 different topics from the suicidal posts, using Latent Dirichlet Allocation (LDA) [45]. The various machine learning algorithms that were used and fine-tuned include the decision tree, logistic regression, multinomial naive bayes, and support vector machine.

V. Results and Discussion

Twitter API was used to extract the suicidal tweets using various keywords and hashtags that represented potential suicide thoughts. These keywords were collected from various research articles, websites, and forums [29], [46]. Reddit posts about suicide were adopted from a previous study [44]. After pre-processing and extracting the relevant features using the proposed feature engineering technique shown in Figure 4, machine learning algorithms were applied to classify the tweets and posts into the three levels of risk. The results generated by applying various pre-processing steps on the dataset used for experimentation are shown in Figure 5. Also, Figure 6. shows the various statistical measures of various features that were used in our dataset.

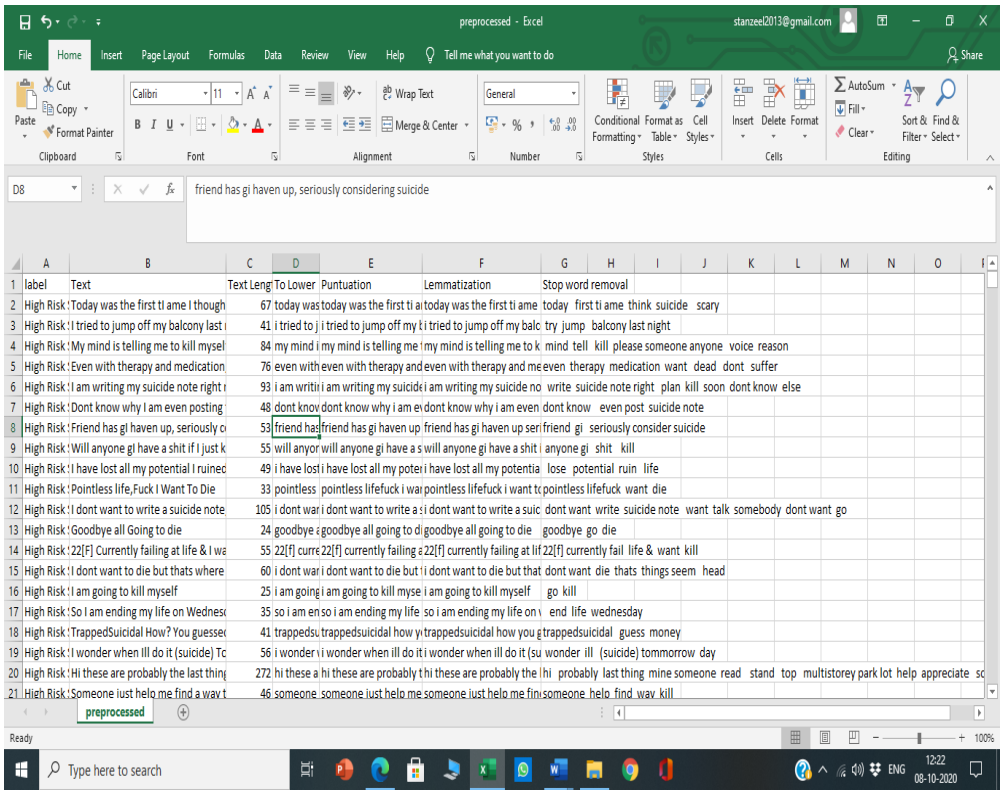


Figure 5-Various pre-processing steps used on the dataset for cleaning the data.

	A	B	C	D	E
1	Features	Standard [	Mean	Minimum	Maximum
2	depress	0.027682	0.114521	0	1
3	post	0.007519	0.070353	0	1
4	worse	0.010087	0.07157	0	0.827672
5	thoughts	0.010605	0.071239	0	0.862159
6	want kill	0.01039	0.073172	0	1
7	later	0.003909	0.044442	0	0.839043
8	live	0.004201	0.045698	0	0.769324
9	go kill	0.005358	0.054931	0	1
10	want end	0.007904	0.064822	0	1
11	ive	0.004743	0.053434	0	0.749548
12	worthless	0.008715	0.070228	0	1
13	end	0.013878	0.082239	0	0.874562
14	suicidal	0.005206	0.057532	0	1
15	rt	0.004574	0.052105	0	0.87948
16	bar	0.004868	0.047648	0	0.675821
17	need help	0.008624	0.07243	0	0.857616
18	pain	0.0511	0.12417	0	1
19	feel	0.02192	0.096676	0	0.68594
20	die sleep	0.008714	0.065791	0	0.752952
21	anxiety	0.008668	0.067653	0	0.760968

Figure 6- Statistical measures of some of the features

We selected Python as a language to implement the various machine learning algorithms. Different packages that were used are NLTK, Pandas, Sckitlearn, and others. The results generated are summed up in Table 2.

Table 2-Classifiers and their performance based upon F-measure

Metric used	Classifier	Suicidal Ideation Risk classification		
		Level 0 (No Risk)	**Level_1 (Moderate Risk)**	**Level_ 3 (High Risk)**
F- Measure	Decision Tree	0.96	0.95	0.97
	Multinomial Naïve Bayes	0.89	0.85	0.86
	Support Vector Machine	0.89	0.86	0.85
	Logistic Regression	0.89	0.92	0.86

The comparison of various implemented machine learning algorithms is also shown in Figur 7. In evaluating the results, it was found that F1 had a very impressive score that ranged from 0.85-0.97. Among all machine learning algorithms applied on the dataset, decision tree outperformed the other three algorithms, with f-score values of 0.96, 0.95, and 0.97 for low risk, medium risk, and high risk, respectively.

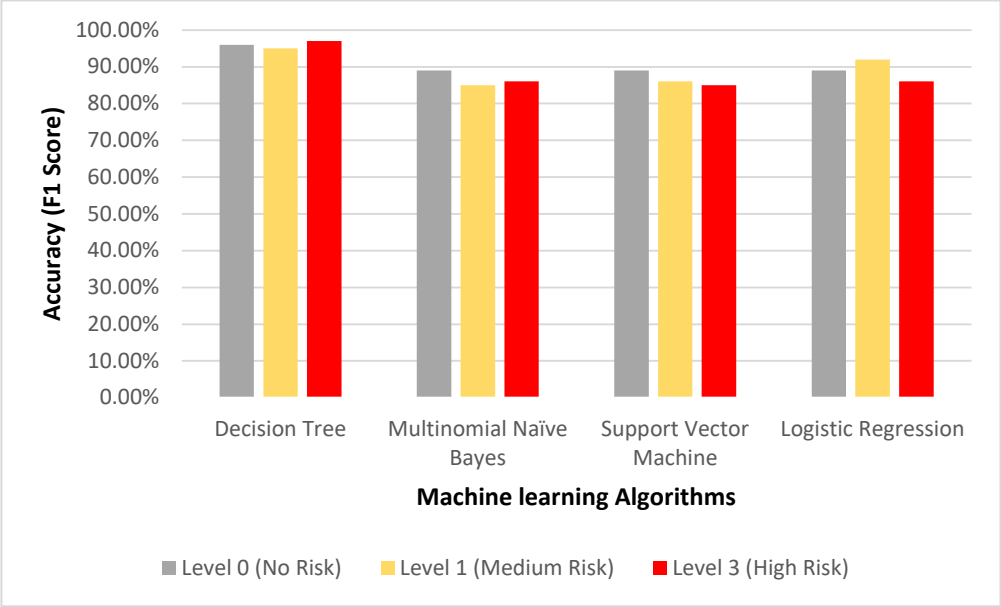


Figure 7-Comparison of various machine learning algorithms based upon F1-score in classifying the tweets into three levels of suicidal ideation.

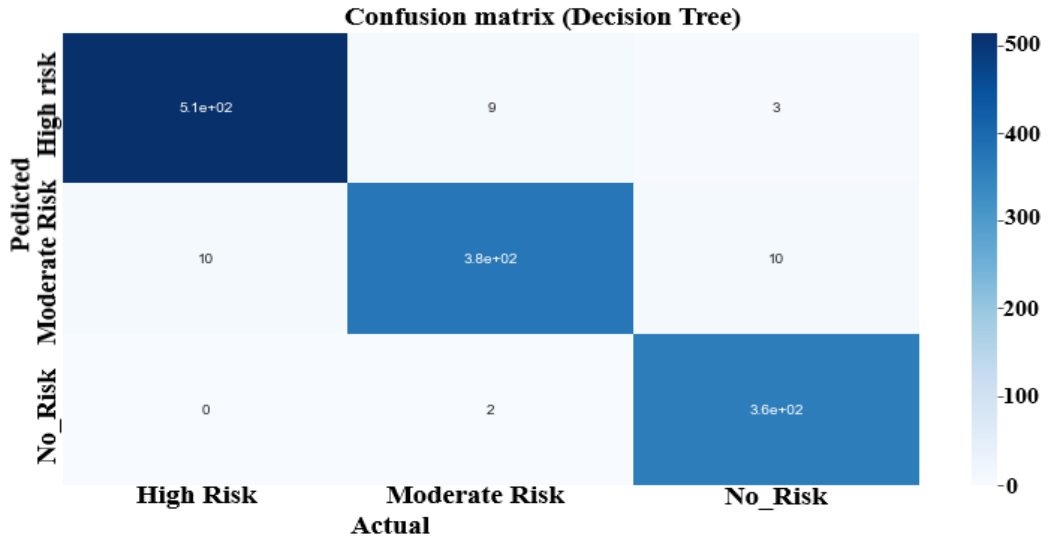


Figure 8-Confusion matrix of decision tree (best performing algorithm) on the dataset used.

Conclusions

In this paper, a thorough review of previous research on mining and analysis of social media data related with suicidal ideation was constructed. Some of the limitations of the recent work were identified. Different machine learning classifiers and evaluation metrics were also studied. Some of the limitations and gaps in previous research were worked upon by proposing a hybrid feature engineering mechanism that was applied on our dataset. Based on that, a novel classification technique was proposed by fuzzifying the levels of distress. Among the four well know machine learning algorithms, decision tree outperformed all algorithms in classifying the tweets/posts, with F-score values of 0.96, 0.95, and 0.97 for low risk, moderate risk, and high-risk classes, respectively. From the study, it was found that no classifier can be treated as a benchmark for classification. The findings suggest that people feel very much comfortable to post their feelings on social media, which could be a reliable tool to reach a large population suffering from suicidal ideation. This could help the intervention by counsellors as it allows the at-risk persons to receive support without worrying about social stigma. The social media could thus help in understanding the suicide in a better manner and, in turn, benefits a large population, which was not otherwise possible using traditional approaches.

VI. Future direction

The current survey confirms that social media is a rich source to analyse people's behaviour, even though by short text. However, there are some limitations which need to be focussed upon, such as finding the context of tweets or posts externally. The unachieved level of expectation of the learning curve needs to be addressed so that reliability and validity of the model could be improved. The at-risk tweets/ posts need to be fuzzified into various levels of distress and based upon emotions, such that persons falling into the lower levels could also be intervened before they reach the higher levels. The real time suicide risk system could be improved and optimized by identifying the various locations that have the higher rates of strongly concerned tweets or posts and then combining it with various population measures to provide the automated overview of higher risk. Deep learning techniques need to be incorporated to increase the efficiency of automatic classification of tweets in many levels of distress.

References

[1] "Suicide." https://www.who.int/news-room/fact-sheets/detail/suicide (accessed Apr. 08, 2020).

[2] W. H. Organization, "Preventing suicide: A Global Imperative," *World Heal. Organ.*, 2014, doi: 10.1093/qjmed/hch106.

[3] Chethan Kumar, "student suicides: One student kills self every hour in India | India News - Times of India," 2018. .

[4] B. Desmet, "Automatic text classi cation for suicide prevention," pp. 1–205, 2014.

[5] E. Seong *et al.*, "Relationship of Social and Behavioral Characteristics to Suicidality in Community Adolescents With Self-Harm: Considering Contagion and Connection on Social Media," *Front. Psychol.*, vol. 12, p. 2102, Jul. 2021, doi: 10.3389/FPSYG. 2021.691438 /BIBTEX.

[6] B. Desmet, "Automatic text classification for suicide prevention," 2014.

[7] "WHO | Psychiatrists and nurses (per 100 000 population)," *WHO*, 2019, Accessed: Jun. 29, 2019. [Online]. Available: https://www.who.int/gho/mental_ health/human_resources / psychiatrists_nurses/en/.

[8] M. K. M. T. Shioiri, A. Nishimura, K. Akazawa, R. Abe, H. Nushida, Y. Ueno, "Incidence of note-leaving remains constant despite increasing suicide rates," *Psychiatry Clin. Neurosci.*, pp. 226–228, 2005.

[9] J. Jashinsky *et al.*, "Tracking suicide risk factors through Twitter in the US," *Crisis*, vol. 35, no. 1, pp. 51–59, 2014, doi: 10.1027/0227-5910/a000234.

[10] S. Sofi, A. Adil, H. Amin Kar, R. Jangir, and S. A. Sofi, "Analysis of multi-diseases using big data for improvement in healthcare," *IEEE UP Sect. Conf. Electr. Comput. Electron.*, 2015, doi: 10.1109/UPCON.2015.7456696.

[11] A. M. U. D. Khanday, Q. R. Khan, and S. T. Rabani, "SVMBPI: Support Vector Machine-Based Propaganda Identification," pp. 445–455, 2021, doi: 10.1007/978-981-16-1056-1_35.

[12] S. Islam, S. Jamwal, and M. H. Mir, "Leveraging Fog Computing for SmartInternet of ThingsCrop Monitoring Farming in Covid-19 Era," *Ann. R.S.C.B*, vol. 25, no. 6, pp. 10410–10420, 2021.

[13] N. J. Khan, G. Ahamad, M. Naseem, and Q. R. Khan, "Fuzzy Discrete Event System (FDES): A Survey," *Lect. Notes Electr. Eng.*, vol. 723 LNEE, pp. 531–544, 2021, doi: 10.1007/978-981-33-4080-0_51.

[14] S. T. Rabani, Q. R. Khan, and A. M. U. D. Khanday, "Detection of Suicidal Ideation on Twitter using Machine Learning & Ensemble Approaches," *Baghdad Sci. J.*, vol. 17, no. 4, pp. 1328–1339, 2020, doi: http://dx.doi.org/10.21123/bsj.2020.17.4.1328.

[15] S. T. Rabani, Q. R. Khan, and A. Khanday, "A nove approach to predict the level of suicidal ideation on social networks Using machine and ensemble learning," *ICTACT J. SOFT Comput.*, vol. 11, no. 2, pp. 2288–2293, 2021, doi: 10.21917/ijsc.2021.0327.

[16] V. Venek, S. Scherer, L. P. Morency, A. S. Rizzo, and J. Pestian, "Adolescent Suicidal Risk Assessment in Clinician-Patient Interaction," *IEEE Trans. Affect. Comput.*, vol. 8, no. 2, pp. 204–215, 2017, doi: 10.1109/TAFFC.2016.2518665.

[17] S. Ji, S. Pan, X. Li, E. Cambria, G. Long, and Z. Huang, "Suicidal Ideation Detection: A Review of Machine Learning Methods and Applications," pp. 1–13, 2019, [Online]. Available: http://arxiv.org/abs/1910.12611.

[18] G. B. Colombo, P. Burnap, A. Hodorog, and J. Scourfield, "Analysing the connectivity and communication of suicidal users on twitter," *Comput. Commun.*, vol. 73, pp. 291–300, 2016, doi: 10.1016/j.comcom.2015.07.018.

[19] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression through social media," in *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, 2013, pp. 128–137.

[20] P. Burnap, G. Colombo, R. Amery, A. Hodorog, and J. Scourfield, "Multi-class machine classification of suicide-related communication on Twitter," *Online Soc. Networks Media*, vol. 2, pp. 32–44, 2017.

[21] A. M. Birjali Marouane Beni-Hssane, "Prediction of Suicidal Ideation in Twitter Data using Machine Learning algorithms," *Int. Arab Conf. Inf. Technol.*, pp. 2–6, 2016.

[22] K. M. Harris, J. P. McLean, and J. Sheffield, "Examining suicide-risk individuals who go online

for suicide-related purposes," *Arch. Suicide Res.*, vol. 13, no. 3, pp. 264–276, 2009, doi: 10.1080/13811110903044419.

[23] Wen-Cheng Chiang, Po-Husn Cheng, Mei-Ju Su, Heng-Shuen Chen, Ssu-Wei Wu, and Jia-Kuan Lin, "Socio-health with personal mental health records: Suicidal-tendency observation system on Facebook for Taiwanese adolescents and young adults," in *2011 IEEE 13th International Conference on e-Health Networking, Applications and Services*, Jun. 2011, pp. 46–51, doi: 10.1109/HEALTH.2011.6026784.

[24] M. A. Moreno *et al.*, "Feeling bad on facebook: Depression disclosures by college students on a social networking site," *Depress. Anxiety*, vol. 28, no. 6, pp. 447–455, 2011, doi: 10.1002 /da. 20805.

[25] K. Fu, Q. Cheng, P. W. C. Wong, and P. S. F. Yip, "Responses to a Self-Presented Suicide Attempt in Social Media," *Crisis*, vol. 34, no. 6, pp. 406–412, Nov. 2013, doi: 10.1027/0227-5910/a000221.

[26] K. Lehavot, D. Ben-Zeev, and R. E. Neville, "Ethical considerations and social media: A case of suicidal postings on facebook," *J. Dual Diagn.*, vol. 8, no. 4, pp. 341–346, 2012, doi: 10.1080 /15504263.2012.718928.

[27] V. K. Kailasam and E. Samuels, "Can social media help mental health practitioners prevent suicides?," *Curr. Psychiatr.*, vol. 14, no. 2, pp. 37–39, 51, 2015.

[28] S. J. Cash, M. Thelwall, S. N. Peck, J. Z. Ferrell, and J. A. Bridge, "Adolescent Suicide Statements on MySpace," *Cyberpsychology, Behav. Soc. Netw.*, vol. 16, no. 3, pp. 166–174, 2013, doi: 10.1089/cyber.2012.0098.

[29] B. O'Dea, S. Wan, P. J. Batterham, A. L. Calear, C. Paris, and H. Christensen, "Detecting suicidality on twitter," *Internet Interv.*, vol. 2, no. 2, pp. 183–188, 2015, doi: 10.1016/ j. invent.2015.03.005.

[30] G. Coppersmith, K. Ngo, R. Leary, and A. Wood, "Exploratory Analysis of Social Media Prior to a Suicide Attempt," *Proc. Third Work. Comput. Lingusitics Clin. Psychol.*, pp. 106–117, 2016, doi: 10.18653/v1/W16-0311.

[31] B. Moulahi, J. Azé, and S. Bringay, "DARE to care: A context-aware framework to track suicidal ideation on social media," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10570 LNCS, pp. 346–353, 2017, doi: 10.1007/978-3-319-68786-5_28.

[32] A. Joseph and B. Ramamurthy, "Suicidal behavior prediction using data mining techniques," *Int. J. Mech. Eng. Technol.*, vol. 9, no. 4, pp. 293–301, 2018.

[33] M. J. Vioules, B. Moulahi, J. Aze, and S. Bringay, "Detection of suicide-related posts in Twitter data streams," *IBM J. Res. Dev.*, vol. 62, no. 1, pp. 7:1-7:12, 2018, doi: 10.1147/JRD.2017 .2768678.

[34] A. Noureen, U. Qamar, and M. Ali, "Semantic analysis of social media and associated psychotic behavior," *ICNC-FSKD 2017 - 13th Int. Conf. Nat. Comput. Fuzzy Syst. Knowl. Discov.*, pp. 1621–1630, 2018, doi: 10.1109/FSKD.2017.8393009.

[35] N. Shahreen, "Suicidal Trend Analysis of Twitter using Machine Learning and Neural Network," *2018 Int. Conf. Bangla Speech Lang. Process.*, pp. 1–5, 2020.

[36] F. Chiroma, H. A. N. Liu, and M. Cocea, "Text Classification For Suicide Related Tweets," *2018 Int. Conf. Mach. Learn. Cybern.*, vol. 2, pp. 587–592.

[37] P. Burnap, G. Colombo, and J. Scourfield, "Machine Classification and Analysis of Suicide-Related Communication on Twitter," 2015, doi: 10.1145/2700171.2791023.

[38] H. H. Won *et al.*, "Predicting National Suicide Numbers with Social Media Data," *PLoS One*, vol. 8, no. 4, pp. 1–6, 2013, doi: 10.1371/journal.pone.0061809.

[39] X. Huang, L. Xing, J. R. Brubaker, and M. J. Paul, "Exploring Timelines of Confirmed Suicide Incidents Through Social Media," *Proc. - 2017 IEEE Int. Conf. Healthc. Informatics, ICHI 2017*, pp. 470–477, 2017, doi: 10.1109/ICHI.2017.47.

[40] A. Chadha and B. Kaushik, "A Survey on Prediction of Suicidal Ideation Using Machine and Ensemble Learning," *Comput. J.*, vol. 00, no. 00, pp. 1–17, 2019, doi: 10.1093/comjnl/bxz120.

[41] A. Roy, K. Nikolitch, R. McGinn, S. Jinah, W. Klement, and Z. A. Kaminsky, "A machine learning approach predicts future risk to suicidal ideation from social media data," *npj Digit. Med.*, vol. 3, no. 1, pp. 1–12, 2020, doi: 10.1038/s41746-020-0287-6.

[42] L. Zheng *et al.*, "Development of an early-warning system for high-risk patients for suicide attempt using deep learning and electronic health records," *Transl. Psychiatry*, vol. 10, no. 1, pp. 1–10, 2020, doi: 10.1038/s41398-020-0684-2.

[43] A. M. U. D. Khanday, S. T. Rabani, Q. R. Khan, N. Rouf, and M. Mohi Ud Din, "Machine learning based approaches for detecting COVID-19 using clinical text data," *Int. J. Inf. Technol.*, 2020, doi: 10.1007/s41870-020-00495-9.

[44] S. Ji, C. P. Yu, S.-F. Fung, S. Pan, and G. Long, "Supervised Learning for Suicidal Ideation Detection in Online User Content," *Hindawi Complex.*, vol. 2018, pp. 1–10, 2018, doi: 10.1155/2018/6157249.

[45] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1033, 2013, doi: 10.1016/B978-0-12-411519-4.00006-9.

[46] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Association, 2013.