FACULTY OF MEDICINE
AND HEALTH SCIENCES

CAN MACHINE LEARNING IDENTIFY SUICIDAL RISK ON SOCIAL MEDIA PLATFORMS?

A Systematic Review

Remi Moerkerke
Student number: 01900565

Supervisor(s): Prof. Dr. Chris Baeken, Paula Horczak

A dissertation submitted to Ghent University in partial fulfilment of the requirements for the degree of Master of Medicine in Medicine

Academic year: 2022– 2023

GHENT
UNIVERSITY

Copyright Declaration

11/13/2023

Remi Moerkerke

Prof. Dr. Chris Baeken

i.o.

co-promotor

Maatschappelijke Outreach

Zelfmoord vormt een probleem dat niet gemakkelijk aan te pakken is in onze samenleving. Het is moeilijk om te weten te komen welke mensen met donkere gedachtes zitten en het risico lopen om zelfmoord te plegen. Dit komt vooral doordat mensen deze gevoelens niet vaak uiten aan de buitenwereld. Met de digitalisering van het dagelijkse leven zijn sociale media platformen een belangrijk onderdeel geworden van onze communicatie en interacties met anderen. Deze platformen worden vandaar ook soms gebruikt om bepaalde emoties te uiten die anders opgekropt blijven. Sociale media biedt dus een rijke bron aan data en gegevens om onderzoek te doen naar mentale gezondheid. Machine learning is een vorm van technologie die goed om kan met grote hoeveelheden van data om te analyseren. Deze technologie lijkt een veelbelovend hulpmiddel om potentiële indicaties van zelfmoordgedachtes of zelfmoordrisico's te identificeren op sociale media. Deze review heeft gekeken naar het huidige onderzoek naar het gebruik van machine learning modellen om suïcidale ideatie of mensen die het risico lopen om zelfmoord te plegen te detecteren op sociale media. Er zijn verschillende studies gevonden die elk een eigen methode ontwikkeld hebben om te proberen een machine learning model te ontwikkelen die in staat is om suïcidale ideatie te herkennen of mensen die het risico lopen om zelfmoord te plegen te detecteren. Dit werd gedaan op basis van verschillende factoren zoals taalgebruik in social media posts maar ook aantal volgers, gemiddelde uur van een social media post online plaatsen, enz. Deze studies tonen veelbelovende resultaten maar door de verscheidenheid aan aanpakken was het moeilijk om de studies onderling te vergelijken en te bekijken welke methode het beste was. Vandaar dat er meer onderzoek nodig is die de verschillende methodes verder exploreert op en gestandaardiseerde manier. Verder is het ook belangrijk dat er rekening gehouden wordt met ethische en privacy gerelateerde kwesties, vooraleer machine learning modellen gebruikt kunnen worden voor het detecteren van mensen op sociale media die het risico lopen om zelfmoord te plegen

Table of Contents

Copyright Declaration ii

Maatschappelijke Outreach iii

Abstract EN 1

Abstract NL 2

Introduction 4

- Suicide, Suicidal Ideation, and Suicide Prevention 4
- Social Media 6
- Machine Learning 10

Methods 19

- Eligibility Criteria 19
- Information Sources 19
- Search Strategy 20
- Selection Process 20
- Data Collection and Data Extraction 21

Results 22

- Study Selection 22
- Study Characteristics 23
- Results of Individual Studies 26

Discussion 29

- Data 29
- Methodological Considerations 31
- Study Results 35
- Recommendations for Future Research 36
- Implications and Ethical Considerations 37
- Limitation of This Review 37

Conclusion 38

References 39

Addenda i

Abstract EN

Introduction

The detection of suicidality and prevention of suicide remain critical public health challenges, often hindered by the difficulties of identifying individuals at risk. With the digitalization of daily life, social media platforms have become an important part of our communication and interactions. This extensive digital landscape provides a rich source of data, offering a lot of potential for mental health research. Machine learning, through its analytical capabilities, emerges as a promising tool to navigate this vast digital landscape and identify potential indicators of suicidal ideation or suicide risk, opening doors to more effective prevention strategies.

Objective

The objective of this review was to provide an overview of existing research on the use of machine learning models to detect suicidal ideation or at-risk individuals on social media platforms, comparing different approaches along with recommendations for future research.

Method

PubMed, IEEE Xplore, Embase and Web of Science were searched for eligible studies that employed machine learning models to detect suicidal ideation or identify users at-risk of suicide. Studies were then screened and selected based on the eligibility criteria. The included studies were analyzed, and important information was extracted and presented in a table. Due to the diversity in methods among the studies, a categorization model was proposed to classify the studies into four groups based on their methods to facilitate comparison.

Results

A total of 32 studies were included in this review, each employing a variety of machine learning models, showing promising yet diverse outcomes in classifying suicidal ideation or identifying at-risk users. Due to the lack of homogenous methods and performance metrics used to report outcomes, a direct comparison was difficult.

Conclusion

While current research offers valuable insights into online behaviors linked to suicidality, further exploration and refinement are essential. Addressing ethical, privacy, and representational concerns is critical before implementing machine learning models for screening and preventative projects. The review highlights the need for standardized methodologies and diverse, clinically informed approaches to comprehensively address mental health complexities within social media contexts.

Abstract NL

Introductie

De detectie en preventie van zelfmoord blijven belangrijke uitdagingen voor de volksgezondheid, die vaak gehinderd worden door de moeilijkheden om individuen die risico lopen te identificeren. Met de digitalisering van het dagelijks leven zijn sociale media platformen een belangrijk onderdeel geworden van onze communicatie en interacties. Dit uitgebreide digitaal landschap vormt een rijke bron van data en biedt verder mogelijkheden voor onderzoek naar geestelijke gezondheid. Machine learning, een technologie met analytische vermogens, lijkt een veelbelovend hulpmiddel om door deze uitgestrekte zee aan digitale data te navigeren en potentiële indicatoren van suïcidale ideatie of zelfmoordrisico's te identificeren, wat verder gebruikt kan worden als fundament voor het ontwikkelen van preventiestrategieën.

Doelstelling

Het doel van deze review was om een overzicht te geven van bestaand onderzoek naar het gebruik van machine learning modellen om suïcidale ideatie of mensen die het risico lopen om zelfmoord te plegen te detecteren op sociale media, waarbij verschillende methodologieën vergeleken werden en aanbevelingen werden gedaan voor toekomstig onderzoek.

Methode

Er werd gezocht op PubMed, IEEE Xplore, Embase and Web of Science voor studies die gebruik maakten van machine learning modellen om suïcidale ideatie te detecteren of gebruikers met een suïciderisico te identificeren. Studies werden vervolgens gescreend en geselecteerd op basis van de geschiktheidscriteria. De geïncludeerde studies werden geanalyseerd en belangrijke karakteristieken werden verzameld en gepresenteerd in een tabel. Vanwege de diversiteit in methoden onder de studies werd een categorisatiemodel voorgesteld om de onderzoeken op basis van hun methoden in vier groepen in te delen om zo vergelijking te vergemakkelijken.

Resultaten

In totaal werden 32 studies opgenomen in deze review, die elk gebruik maakten van verschillende machine learning modellen en veelbelovende maar uiteinlopende resultaten toonden. Door gebrek aan homogene methoden en manieren van het rapporteren van de resultaten, was een exacte vergelijking moeilijk.

Conclusie

Hoewel het huidige onderzoek nuttige inzichten biedt in online gedrag met betrekking tot suïcidaliteit, is verdere verkenning essentieel. Het is van cruciaal belang om ethische, privacy en representatiegerelateerde problemen aan te pakken vooraleer machine learning modellen gebruikt

kunnen worden voor screening en preventie. Deze review benadrukt de noodzaak van gestandaardiseerde methodologieën en diverse, klinisch geïnformeerde benaderingen om de complexiteit van geestelijke gezondheid binnen sociale media in volledigheid aan te pakken.

Introduction

Suicide, Suicidal Ideation, and Suicide Prevention

Suicide is a major public health concern. Every year, over 700,000 people die worldwide, making it one of the most common causes of premature death (1). Furthermore, according to a survey from 2013, around 64% of people know someone who has attempted suicide, indicating the widespread impact of suicide on the larger community (2). Despite its prevalence, suicide prevention has not received the attention it deserves due to a lack of awareness of suicide and the stigma that surrounds it in many societies (1). This stigma and the lack of an open dialogue can be particularly challenging for individuals who struggle with suicidal thoughts and for suicide survivors (3). This introduction will first explore the various aspects of the complexity of suicide, discussing risk factors and existing prevention strategies. Then, social media will be discussed, and its potential in contributing to suicide research and prevention. Lastly, machine learning will be addressed, and the opportunities provided by this emerging technology in enhancing the detection and prevention of suicide will be examined.

Suicide itself is preceded by a process that can be broken down into different stages. One framework that tries to conceptualize this process is called 'the ideation-to-action framework'. Development of suicidal ideation, defined by the DSM-5 as "thoughts about self-harm, with deliberate consideration or planning of possible techniques of causing one's own death", is the first step of the three-step theory based on the previously stated framework (the two other steps being 'strong vs moderate ideation' and 'progression from ideation to attempts') (4) (5). Suicidal ideation is a relatively common experience, almost 1 in 10 people will experience suicidal thoughts at some point in their lives (6). It is particularly prevalent in young adults, with research indicating that individuals aged 18 to 25 are at higher risk of experiencing suicidal ideation compared to other age groups (7) (8).

Suicidal ideation, however, does not always lead to a suicide attempt. Most people who experience suicidal ideation do not follow through with suicidal actions (6). A major factor that could predict whether someone with suicidal ideation is actually at risk of attempting suicide is their sense of social connectedness and their sense of hopelessness and pain (5). In addition, other factors that may contribute to the risk of suicide include a history of mental illness (especially depression), a history of trauma, substance abuse, chronic illness, or poverty (9). There is also a difference between gender, women are more likely to experience suicidal ideation however male suicide rates are higher than female suicide rates. This could potentially be explained by the traditional Western idea of masculinity in comparison to femininity. It is perceived as less socially acceptable for a man to openly talk about anxiety or discuss their emotions. In contrast, women tend to communicate more easily and stay connected socially, which is an important protective factor against suicide (10). There are also gender-specific risk factors that should be taken into account. Women may be more likely to attempt suicide if they struggle with an eating disorder, a posttraumatic stress disorder, a bipolar disorder, interpersonal issues, or have a history of previous abortion. On the other hand, men may be at greater risk if they show disruptive behavior, feel hopeless, if their parents are divorced, if they have a friend who is suicidal, or because they have access to means to commit suicide (11). Some social groups are more vulnerable when it comes to suicide than others. For example, transgender people are at a higher risk of attempting suicide compared to the general population (12). This disparity is also found among sexual minorities, with bisexual individuals being particularly vulnerable (13). Discrimination based on race is also associated with an increased risk (14). Social groups that experience more discrimination, stigmatization, and social rejection are more likely to express suicidal ideation and attempt suicide indicating the importance of addressing the unique needs of these groups and developing targeted strategies for suicide prevention.

Although there is extensive literature on the risk factors associated with suicide, prevention is not an easy task. One major strategy that is deemed effective in decreasing the prevalence of suicide is limiting access to lethal methods of self-harm (15). However, it should be noted that the most common methods of suicide differ depending on geographic location. For example, firearms are the most commonly used method in the United States, while poisoning with pesticides is more prevalent in rural areas in Latin America and Asia. This is why these specific methods of prevention should be tailored to the specific methods of suicide prevalent in a particular region (16). School-based awareness programs are also effective in reducing suicide attempts, they have been shown to reduce the risk of suicide by 55% (15).

Treating suicidal ideation before it turns into suicidal behavior is also a vital part of suicide prevention. There is evidence that suggests that psychotherapeutic interventions, like Cognitive Behavioral Therapy (CBT) and Dialectical Behavior Therapy (DBT), are effective treatments for suicidal ideation (17). CBT is a form of talk therapy that helps to break negative thought patterns and change behavior. DBT focuses more on emotion regulation and the development of mindfulness skills in addition to addressing maladaptive behaviors and thought patterns. Ketamine could also prove to reduce suicidal ideation, especially in urgent settings, although more research is needed when it comes to long-term safety (18).

One aspect that could contribute to more effective prevention programs is the timely detection of those at risk, however, this is not always straightforward in a clinical setting. Almost half of the people who die by suicide have contact with a primary care physician within 1 month of death, but a mental health diagnosis is often not documented. This is why greater efforts should be made to evaluate mental health and suicide risk (19). Despite numerous efforts, there are currently no cost-effective screening tools (15). The Columbia- Suicide Severity Rating Scale (C-SSRS) has been considered the 'golden standard' for assessing suicidal ideation. It consists of multiple sections and aims to predict whether a suicidal or non-suicidal individual is at risk of committing suicide. This tool is not without its criticism as it does not seem to fully address the spectrum of suicidal ideation or behavior and has the potential to miss many combinations of suicidal ideation (20). Another criticism is that the C-SSRS can only be filled in by a trained interviewer, whereas the Beck Scale for Suicidal Ideation (BSSI), an older tool, can be self-administered. The BSSI is more focused on evaluating the severity of the suicidal ideation itself. Even though they have some limitations, both tools are examples of tools that can be used to support healthcare professionals in evaluating suicide risk. However, it is worth noting that these tools are typically used for individuals who are already considered at risk of suicidal behavior (21). There is still a gap when it comes to effectively detecting individuals who are potentially at risk. One of the reasons that it is difficult to identify someone who is suicidal or has suicidal ideation is that these thoughts are often not expressed directly (22). Data gathered from social media posts could potentially provide a solution to this problem by timely identifying at-risk individuals who may not express suicidal ideation directly in other contexts.

Social Media

Social media has gained a lot of popularity in the past decade with about 4.62 billion social media users around the world or 58.4% of the world's total population (23). Social media websites or programs allow people to connect with each other through the internet from their computers or mobile phones. They provide a medium to share text and other media like pictures and videos. The time we spend on social media has been increasing over the past couple of years, with an average of 2 hours and 27 minutes per day, social media accounts for the largest single share of our time connected to the internet (23). Over the years, several social media platforms have emerged as key players in this rapidly evolving digital world. One of the first globally adopted platforms that has the most active users across all age groups is Facebook (23). It is a platform that allows users to connect with friends and family, share updates, photos, and videos, and join interest-based groups (24). Another popular platform, especially for a younger audience, is Instagram. This platform, which is also owned by Facebook, focuses more on media like photo and video-sharing. Users can post visual content and follow friends, celebrities, or brands (25). TikTok, which also gained significant popularity among younger audiences, is a rapidly growing platform where people can share and watch short videos, often set to music (26). Twitter, now also known as X, is a platform where users can connect and share short posts, also known as tweets. It is very popular for engaging in public conversations and sharing opinions and ideas (27). Weibo can be considered a Chinese alternative to Twitter (28). YouTube is not always considered a social media platform although it shares a lot of similar features. It is a video-sharing platform where users can upload, view, and comment on videos and follow channels (29). These social media platforms are most used by people aged 15 to 29 (23). Crucially, this is about the same age group that is most at risk for suicidal ideation (cf. supra). Furthermore, suicide ranks are the fourth leading cause of death for this age group (1). Social media platforms allow people to express their emotions and thoughts, and as a result, some people may express suicidal ideation on these platforms. Current research even suggests that there is a significant association between suicide-related posts and suicide rates. This is why social media platforms offer a promising opportunity for identifying at-risk individuals and putting suicide prevention programs into action (30).

There are several benefits that social media platforms can offer in the context of suicide prevention. Social media platforms can provide an effective way to spread more awareness and information about suicide to the general public, as most people use social networks on a daily basis (31). Social media also offers the opportunity to reach specific target groups. As previously stated, it is very important to take into account that certain social groups are more vulnerable to experiencing suicidal ideation and attempting suicide. These individuals often join online social media networks where they can connect with like-minded people, share their problems, and offer mutual support. These social networks offer an opportunity to deploy prevention programs that can reach more people from these specific target groups (32). For instance, a study that evaluated the reach of a potential prevention program targeting lesbian, gay, and bisexual adolescents found that with the use of a social networking service like MySpace, they were able to reach more than 18.000 individuals, which is significantly more than current traditional methods allow (30). Social media could also be an opportunity for suicide research. Current research focuses a lot on traditional risk factors and does not fully acknowledge the multifaceted nature of suicide. For example, one aspect that has not yet received adequate attention is the role of social context (33). Data from social media has already been used to identify risk factors and online activity patterns for detecting at-risk individuals (34). Another study tried to identify suicide-related risk factors on Twitter, a social media platform where users can post and react to short messages known as "tweets". They studied conversations and compared them to geographic suicide rates from vital statistics data. Their findings supported the fact that Twitter could serve as a potential dataset for future suicide research (35).

While social media can offer valuable opportunities for suicide research and prevention, there are several other aspects that should be considered. Social media itself can have a negative impact on suicidal ideation or suicidal behavior. These platforms can provide more straightforward methods for someone to be personally attacked and bullied. Victims of cyberbullying, mostly young people, are almost twice as likely to attempt suicide, which makes it a serious risk factor (31). It also affects other risk factors associated with suicidal ideation and behavior, like depression (36). Furthermore, people who participate in cyberbullying have also reported an increased likelihood of attempting suicide (31). Another negative aspect of social media that should be considered is the facilitation of the endorsement of suicidal behavior. Even though some studies found that social networks provide opportunities to receive help and support from others, the opposite effect has also been reported. If someone expresses their suicidal intentions on unmediated platforms or private chatrooms within social networks, there exists a risk that the suicidal behavior is further endorsed or that multiple individuals form a pact and all agree to commit suicide (31). These pacts are not a new phenomenon. Studies from before the time of social media have reported on the occurrence of suicide pacts. One study concluded the suicide pacts were more common in the first half of the 20th century (37). It is difficult to find a lot of information regarding the global occurrence of suicide pacts and whether the internet has contributed to a global increase in such incidents. Instead of a decrease, a Japanese study found that there now might be an increase in suicide pacts and that this can be attributed to the rise of the internet and social media (38). Twitter has been described as one of the platforms where people easily make suicide pacts. It is a platform where you can easily stay anonymous and create multiple accounts, which makes it more difficult to implement prevention strategies on time (39). Closely linked to suicide pacts is the phenomenon of copycat suicides or suicide contagion. There is a lot of research stating that the suicide of a celebrity can cause a sudden surge in suicide rates. A South-Korean study found evidence that the suicide of a national celebrity significantly increased national suicide rates with media coverage playing a huge role in this phenomenon (40). This effect is called the “Werther effect”, named after Goethe’s “The Sorrows of Young Werther” from 1774, which had caused several copycat suicides (41). The World Health Organization (WHO) has provided guidelines on responsible reporting of suicide to try and minimize the detrimental effect media reportings can have on people who are suicidal (42). Social Media platforms however can be used to discuss and report on these topics freely and can undermine these guidelines.

People who access the internet or social media platforms may come across imagery and videos of self-harm acts, sometimes normalizing suicidal behavior. One study found that people aged 10-17, were 11 times more likely to express suicidal ideation after visiting websites that encouraged self-harm or suicide (43). Luckily there has been a lot of progress in recent years in addressing the issue of suicidal content on social media platforms. Platforms such as Facebook, Twitter, and Instagram have implemented guidelines for identifying and removing such content. Users can also report certain content if they feel it does not conform to the platform’s guidelines. Additionally, these platforms provide resources such as links to crisis hotlines and support groups, for users who may be struggling with mental health issues. While these efforts might be a step in the right direction, there is still a lack of research evaluating the effectiveness of these types of strategies (31).

Another aspect that has received a lot of criticism is the algorithms social media use to show certain content to a user. In the early days of social media, platforms like Facebook mainly showed you content shared by other people you decided to connect with. Then came the introduction of an algorithm that showed you the content that this algorithm speculated you would find most interesting. With the evolution of technology, this algorithm became more and more complex and started using more variables, for example how long a user looks at a certain post or how much they interact with certain content (44). The primary objective of these algorithms is to show personalized content to users, to retain their engagement with the social media platform. The potential for these algorithms to promote harmful content has recently gained a lot of attention. YouTube, mainly used for posting and viewing videos, is another platform that has become very reliant on algorithms with more than 70% of all views on the platform being a result of the algorithm’s recommendations (45). The algorithm has been criticized for tending to show videos similar to those the user has previously watched. A study has identified this “homophily effect” as potentially harmful when it comes to providing scientific information during a pandemic. For example, the algorithm's tendency to suggest anti-vaccine videos to individuals who have previously shown an interest in this topic may be detrimental (46). This “filter bubble” where users are only exposed to content that confirms their existing beliefs and interests could also be harmful to individuals at risk for suicidal behavior, especially because there is a lack of validation of the preventative measures these platforms have implemented over the years. However, there is currently not a lot of research on this specific topic, which could be explained by the constant and rapidly evolving nature of these algorithms. The rapid evolution of these algorithms can partly be attributed to the integration of new machine-learning techniques. These new techniques can help algorithms in managing the vast amount of data generated daily by billions of users.

Machine Learning

Artificial intelligence (AI) is defined as the capability of a computer or a computer-controlled robot to carry out tasks that are usually associated with intelligent beings (47). Machine learning is a subcategory of artificial intelligence that focuses on using data to train computers and make them function without the need of being programmed. Machine learning involves gathering and preparing data, such as numbers, images, or text, which are used as training data. Programmers select a suitable model, provide the data, and let the model learn patterns or make predictions. The model can be fine-tuned for improved accuracy. It can then be evaluated with the use of data that was put aside. The result is a model that can be applied to different data sets in the future (48).

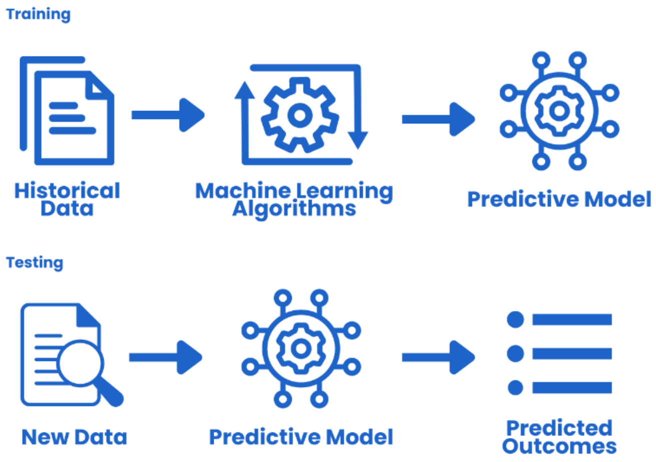


Figure 1. A visual representation of the mechanism of machine learning with both training and testing. Adapted from: Sarker IH. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Comput Sci. 2021;2:160. doi:10.1007/s42979-021-00592-x

There are several types of machine learning, with the main difference being between supervised and unsupervised machine learning (49). While there exist numerous other types, such as reinforcement learning and semi-supervised learning, delving into those would go beyond the scope of this introduction. Supervised learning is currently the most used technique. The term “supervised” refers to the fact that it needs external assistance to function. The algorithm receives input data with the corresponding output labels, which allows the algorithm to learn the relationship between the two. There are two fundamental categories of supervised learning. If the model has to categorize input data into specific predefined groups, it is called a classification problem, with examples such as logistic regression and support vector machines. If the output involves continuous variables, e.g. blood pressure, then this is called a regression problem, with examples like linear regression and decision tree regression (50). Unsupervised machine learning does not use labeled data for categorizing information, unlike supervised machine learning. Its main goal is to analyze and cluster unlabeled datasets to find similarities among different data points. The model does not have any prior knowledge about the organization of the clusters. Therefore, clustering may lead to the uncovering of unforeseen relationships among data points (49). K-means clustering is one of the simplest examples of clustering. In this method, the algorithm tries to categorize data into different groups where each data point belongs to the cluster with the nearest mean value (51). While supervised learning often has a more specific goal, unsupervised learning helps to uncover hidden patterns and relationships in the data that may not be immediately apparent.

Several models can also be combined, resulting in an ensemble method. This combination of individual outputs produces a model with improved predictive performance (52). An example is Random Forest, an existing ensemble learning method that combines predictions from multiple decision trees. This method can also be referred to as a ‘consensus model’.

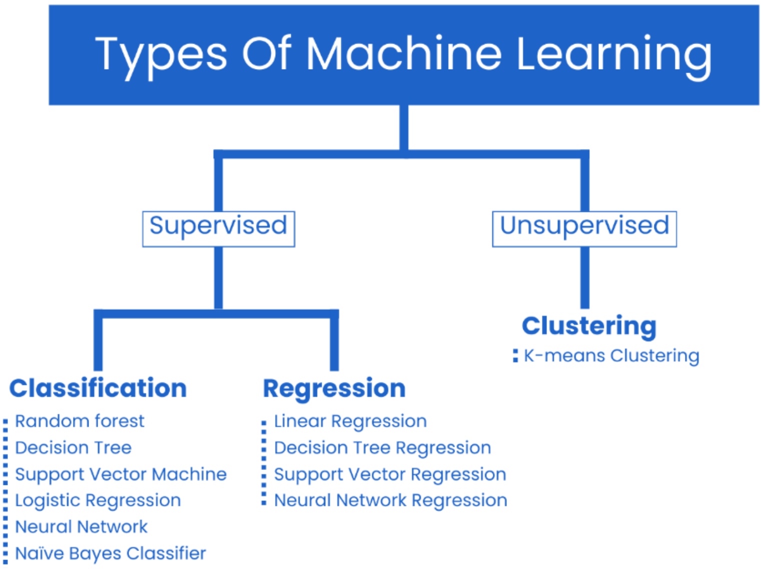


Figure 2. Types of machine learning with some examples of the most common models used in research.

In order to evaluate the effectiveness of machine learning models, it is important to understand the different performance metrics commonly used in research (53). These results are used to interpret the performance of a machine learning model. Some of the frequently used metrics in the context of machine learning are accuracy, precision, recall, F1 score, and AUC. Here we briefly explain and illustrate these metrics.

Accuracy is one of the simpler ways to measure the overall performance of a model. It refers to the proportion of true positive and true negative predictions out of all the predictions made by the model (54).

$$\frac{True\ Positives + True\ Negatives}{Total\ Predictions}$$

Even though it is most frequently used, it can be misleading if there is an imbalance in the distribution of classes in the dataset (55). Here is a hypothetical example: In a dataset of 20 at-risk individuals and 900 controls, the model could miss half of the at-risk people and still have an accuracy of almost 98.91% if it manages to correctly classify all the controls as ‘not at risk’.

Precision refers to the proportion of true positive predictions out of all the positive predictions made by the model (56). If the precision in a model is high, it was able to identify people at risk without wrongly classifying individuals as at risk when they were not at risk. In other words, the model has a low rate of false positives.

$$\frac{True\ Positives}{False\ Positives + True\ Positives}$$

Recall, also referred to as sensitivity, is the proportion of true positive predictions out of all actual positive cases in the dataset (56). It tells us something about the ability of a model to correctly identify at-risk individuals as at risk. A low false-negative rate results in a high recall rate.

$$\frac{True\ Positives}{False\ Negatives + True\ Positives}$$

F1 score consists of both recall and precision. It is the harmonic mean of these two metrics, balancing both in a single metric (57). This can be valuable if the accuracy is not reliable due to an imbalance in the data, as stated previously. A high F1 score indicates that it has both high precision and recall, indicating good performance.

$$\frac{2 * Precision * Recall}{Precision + Recall}$$

Area Under the Curve (AUC) is another important performance metric. It is derived from the Receiver Operating Characteristic (ROC) curve, which is a visualization of a model's ability to discriminate between positive and negative cases across various threshold settings. In this graph the true positive rate (sensitivity) is plotted on the Y-axis and the false positive rate (1-specificity) is plotted on the X-axis (58). By varying the threshold, you can control the trade-off between the true positive rate and the false positive rate. You can make it very strict with a false positive rate of zero, but then the true positive rate will be lower. The AUC is a single value that summarizes the overall performance of the model across all threshold levels, it is the probability of correctly distinguishing between a randomly chosen positive instance and a randomly chosen negative instance. An example of a ROC curve is given below with two different outcomes (see Fig. 3). In this instance, we can see that the area enclosed by the orange line and X- axis is smaller than the area enclosed by the green line and the X-axis. The AUC of orange is smaller than the AUC of green, meaning that the orange model would make more mistakes in detecting true positives at a faster rate compared to the green model. An AUC of 0.5 indicates that the model does not perform better than random chance and therefore means the model has no discriminative power. If you take into account that the perfect AUC is 1, an AUC of 0.8 can be considered very good.

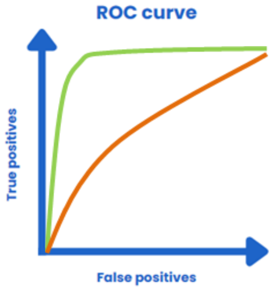


Figure 3. Illustration of ROC curves. The green line represents a high AUC, while the orange line represents a low AUC.

In recent years, the implementation of machine learning has been widespread across numerous domains, including healthcare. One of the reasons machine learning is considered promising in healthcare is because of the accessibility of data, driven by the digitalization and centralization of health records. This increased data availability allows for the creation of stronger and more precise algorithms because more data can increase the accuracy of a machine-learning model (59). Historically, the medical field of infectious diseases has been known for recording a significant amount of epidemiological data, because of the importance of trying to understand and predict the spread of certain diseases. For example, machine learning models have been shown to be successful in predicting the outcome of patients infected with Ebola Virus based on a limited amount of clinical symptoms and laboratory results (60). The use of certain models could also help predict a hospital patient's likelihood of contracting hospital-acquired infections like Clostridium difficile infection (61). As we have recently witnessed how fast a new infectious disease can emerge, machine learning can help us understand and manage novel global outbreaks. There have been models developed that can discover antibodies that can bind and potentially neutralize SARS-CoV-2 based on data from other known virus-antibody sequences. These models can find potential antibodies for various viruses, based on the amino acid sequence alone, something that would typically take months using conventional experimental methods (62). Furthermore, machine learning has been demonstrated to be effective in diagnosing patients with COVID-19 through CT scans (63). Deep learning, a subset of machine learning that uses multi-layered artificial neural networks to learn more complex patterns, has already been widely adopted in many medical imaging settings and will continue to expand its impact (64). Another field that greatly benefits from machine learning is genetics, as it aids in the analysis of the immense volume of data this field produces. It can help in the process of annotating genes along entire chromosomes and could also find and characterize genes when it comes to susceptibility to diseases and is vastly superior to traditional statistical methods (65) (66). Another promising area for machine learning in healthcare is personalized medicine. It is already widely known that because of individual differences among patients, not everyone processes medication in the same manner (67). Taking this into account is crucial for certain medications that can have toxic effects at the wrong dosage. However, predicting the precise impact is difficult, due to the vast amount of largely unknown factors that play a role in pharmacokinetics. Several studies explored the implementation of machine learning to predict the correct dosage of Warfarin, a coagulation drug with a narrow therapeutic window (68) (69). They found that machine-learning methods could outperform traditional approaches to assessing the correct initial dosage.

These examples represent only a small part of the numerous potential uses for machine learning in the realm of medicine. As demonstrated, machine learning can serve as a very helpful tool and resource for both medical research and practical clinical applications. Compared to traditional statistical analysis, machine learning offers several advantages when it comes to research, for instance, both methods can be used for prediction based on data but traditional statistical methods sometimes struggle with handling enormous quantities of data and complex relationships (70).

There are, however, some challenges that have been described when it comes to implementing machine learning in healthcare. Firstly, it is important to note that even though there is a rising interest in machine learning medical research and data is becoming more widely available, there still is a lack of meaningful contributions to clinical care in comparison with other industries (71). One difficulty is the data itself. The importance of high-quality and accurately labeled data in machine learning cannot be overstated. It forms the foundation upon which machine learning models are built (72). This can be a challenge when it comes to healthcare. In contrast to fields where data is well-organized, healthcare data tends to be diverse, filled with noise, and often incomplete. It can be difficult for machine learning models to deal with these large and varied datasets that can include sparse data and missing values (73). It can be very time-consuming and costly to acquire and standardize large amounts of unbiased data (71). The incompleteness of electronic health records (EHR) data has been described by several studies to be one of the main lasting reasons for data quality issues (74) (75). As a result, more efforts in this area are necessary to develop more accurate machine-learning models that can be deployed in real clinical settings. Another common problem when it comes to data and data mining is data leakage. It occurs when the machine learning model accidentally gets information about the answer it’s supposed to predict. This can lead to the development of a model that seems more accurate than it is (76). Data leakage can result from various factors such as improper data preprocessing, issues with the timing of data used (temporal leakage), or incorrect data sampling methods. If the model finds surprising data patterns, like a strong correlation between the “patient ID” and a diagnosis, it’s reasonable to suspect data leakage may be occurring. However, it is also possible that the patient ID number is assigned based on specific factors that can be linked to a certain diagnosis, which makes it more difficult to suspect leakage. The best way to detect data leakage is to cross-validate the model with independent datasets, preferably collected or generated using a different process. If this causes a significant drop in performance, it could be a clear sign of data leakage (77). Another challenging factor is the fact that these complex machine-learning models are considered “black-box technology”. These models possess several hidden layers that autonomously recognize characteristics of the dataset. However, it is not always possible to reveal how the classification process takes place or what these characteristics are (78). This is why there has been some hesitation in using these models because the outcome of potentially misclassifying a patient could be detrimental. There recently has been a high demand for interpretable machine learning models that could provide reasoning behind given outcomes so a clinician can still choose to accept or reject certain predictions (79). The problem of interpretability has been demonstrated in a study that investigated the application of machine learning models for predicting whether a pneumonia patient was at a higher risk of death (80). Their model indicated that pneumonia patients with a history of asthma had a lower risk of dying, which opposes general knowledge of asthma being a significant risk factor for severe pneumonia. The model had failed to interpret the nuance that asthma patients when admitted to the hospital for pneumonia, received more aggressive and specialized care due to their pre-existing condition. This example highlights the need for the ability to understand the reasoning behind a model's prediction. There is still an ongoing discussion about whether making more understandable models decreases the accuracy of these models (81). As a model increases in complexity and takes into account more parameters and layers of data, its potential for accurate predictions can also increase. However, this can often result in outcomes that are difficult to understand and interpret.

Since the beginning of machine learning and artificial intelligence, people have been interested in utilizing these technologies to help detect individuals at risk of suicide. Using algorithms with the help of computers to detect suicide risk has been first described in 1974 by Greist et al (82). After this publication, little research was done to further investigate the implementation of machine learning and computers when it comes to suicide prediction. It is only in the last decade, with the exponential development of this type of technology, that there has been substantial growth of interest in this field. Several studies have used machine learning with data from electronic health records to accurately predict suicidal behavior. One study achieved AUC varying from 0.81 to 0.86 across various prediction timeframes (83). The authors found some predictive risk factors, including symptoms that involve the patient's emotional state and the history of a depressive episode. Similar results were found in another study that concluded that longitudinal clinical data combined with machine learning could prove to be effective as a broad screening tool to detect whether someone is at risk of attempting suicide (84). However, most research limits itself to clinical data and EHR because these are scalable sources that offer structured data. This does have its drawbacks, for instance, this type of data includes many variables that are not related to suicide, which can introduce noise and could decrease the accuracy of machine learning models (85). People who are at risk of suicide also do not always have contact with a clinical environment, which means they could be overlooked when using this type of data for prediction purposes. Furthermore, it has also been noted that EHR data could include certain biases due to incomplete data for example (86). There are several examples of studies that try to go past medical health record data. In one research project, interviews were conducted and recorded, and these recordings were then used as data for machine learning to try to determine whether someone was suicidal or not (87). Similarly, as previously discussed, machine learning can be highly effective when working with visual data, such as medical imaging, and this is also applicable to suicide research. Functional magnetic resonance imaging (fMRI) has gained a lot of interest when it comes to psychiatric research as scientists strive to understand functional brain pathways and their relationship to mental health. One study has tested a machine learning model to try and identify adolescents with suicidal thoughts by analyzing their neural signatures in fMRI data (88). The scientists presented some concepts that were either positive, negative or something regarding suicide. The model was able to make a distinction between controls and people who engage in suicidal ideation with an accuracy of 91%.

Although these examples showcase significant progress in utilizing machine learning to assist in detecting at-risk individuals and potentially preventing suicide, there still is a lack of scalable cost-effective screening tools. The majority of studies tend to enlist test subjects through medical networks. While this is an effective way of gathering data to test certain models in different applications, it may not be able to reach every individual or at-risk group. Data from social media might provide a more scalable strategy, potentially helping us develop tools that can have a broader impact. To effectively utilize social media data, several data-gathering methods can be employed (89). Some media platforms offer Application Programming Interfaces (APIs,), that enable developers to access and interact with the platform’s data and functionality. An API acts as a bridge between the social platform and the programmers or researchers who want to access the platform’s data. This makes it fairly easy to gather relevant data as you can set parameters and define what you're looking for. Facebook and Twitter are some examples of platforms that offer APIs. It is important to note that the companies behind these platforms have the authority to change, restrict or even shut down their API. Another method that is used if an API is not available or does not provide the necessary data is Web Scraping. It is the process of extracting specific data from websites by looking at the page’s layout. Nowadays, there are specific tools available to simplify the web scraping process, eliminating the need for users to have a programming background. Furthermore, there are third-party social media datasets and data resellers that have gathered and organized social media data that can be used for research.

Because of the vast amount of social media data available, there is a lot of research that looks into different applications of machine learning in different fields. Sentiment analysis is a popular technique that has already been used for business. It entails trying to map out the emotions or opinions of people online. This has already been described as an effective tool to monitor opinions on certain products for marketing campaigns (90). For instance, the Linguistic Inquiry and Word Count (LIWC) method counts and categorizes words into psychological groups to understand their emotional context (91). Similarly, VADER, another tool, not only considers psychological dimensions but also evaluates emotional intensity and context, expanding the scope of analysis (92). Social media also provides a lot of data that can help in identifying epidemics, a field that has already used machine learning in several other ways as we discussed earlier. One study developed and tested a model to detect influenza epidemics based on Twitter messages (93). Their model achieved a correlation of 0.78 with the Center for Disease Control and Prevention statistics. Another study used Twitter to track misconceptions regarding COVID-19 and concluded that machine-learning models could evaluate the public's understanding regarding health-related subjects which could aid in improving health communication (94).

In conclusion, there is a rising interest in using the available social media data in combination with machine learning models. Social media platforms offer a wealth of information about individuals' thoughts, feelings, and behaviors, which can potentially be used to identify patterns and signals associated with suicidality. This is why social media, in combination with machine learning, could contribute to the development of scalable detection tools that have a wider reach than current strategies.

This review examined existing research investigating the implementation of machine learning models for the detection of suicidal ideation, suicidality, or social media users at-risk. Firstly, an overview was given of the method used to identify the studies included in this review, followed by a general overview of their findings. Lastly, the methods and applicability of these studies were discussed, along with some critical and ethical considerations.

Methods

Eligibility Criteria

This systematic review included studies that investigate the use of any machine learning model for detecting suicidality and/or suicidal ideation on social media platforms such as Facebook, Twitter, Instagram, Weibo, and TikTok. Studies that make use of community-based websites or forums, such as Reddit, and focused on subgroups or communities dedicated to mental health and suicide were excluded. These platforms consist of subreddits or threads, which are individual forums or communities dedicated to specific topics. Analyzing subgroups that are explicitly created for discussing mental health and suicide may not provide an accurate representation of the platform’s overall population. Studies that focus on platforms such as WhatsApp, Messenger, Telegram, and Discord were also disregarded since they are primarily used for private one-on-one or small-group communication rather than public postings.

Several limitations on study design, environment, or participant characteristics were applied in this review. Case studies were excluded due to their limited statistical power and generalizability compared to other study designs. Studies that merely describe protocols or those for which the full text was not attainable. Only studies that were published in English, Dutch, or French were included. Studies that did not investigate suicidality as a primary endpoint but included it in their analysis were also added to this review. The inclusion criteria were kept intentionally broad to capture all relevant research on this emerging issue.

Information Sources

Studies were identified by searching the following databases:

- PubMed
- IEEE Xplore
- Embase
- Web of Science

These databases were consulted until the 16th of April 2023. References cited in pre-existing systematic review reports on the current or a similar topic were also analyzed.

Search Strategy

For PubMed, we used the Advanced Search builder with the following query: ‘(suicidality OR depression OR suicide OR suicidal OR Suicidal ideation) AND (social media OR twitter OR facebook OR instagram OR tiktok OR weibo) AND (machine learning or deep learning or algorithms)’.

For IEEE Xplore we searched in All Metadata and created the following search query in Advanced Search: ("suicidality" OR "Depression" OR "Suicide" OR "Suicidal" OR "Suicidal Ideation") AND ("Social Media" OR "Twitter" OR "Facebook" OR "Instagram" OR "TikTok" OR “Weibo”) AND ("Machine learning" OR "Deep Learning" OR "Algorithms")’.

We used the following search query for Embase: ‘('suicidality'/exp OR 'depression'/exp OR 'suicide'/exp OR 'suicidal' OR 'suicidal ideation'/exp) AND ('social media'/exp OR 'twitter'/exp OR 'facebook'/exp OR 'facebook' OR 'instagram'/exp OR 'tiktok'/exp OR ‘weibo’/exp) AND ('machine learning'/exp OR 'deep learning'/exp OR 'algorithms'/exp)’.

For Web Of Science we looked at documents in all fields and used the following query: ‘ALL=(suicidality OR depression OR suicide OR suicidal OR Suicidal ideation) AND ALL=(social media OR twitter OR facebook OR instagram OR tiktok OR weibo) AND ALL=(machine learning or deep learning or algorithms)’.

On Google Scholar we used different search queries to find several existing reviews on the subject: ‘Suicide on social media machine learning review’, ‘detecting suicidality on social media machine learning review’, ‘suicide risk social media machine learning review’. We sorted the results by relevance and examined them. The references of multiple reviews were examined.

All results were added into Endnote (95).

Selection Process

For the first selection, one reviewer screened the title and abstracts of all the selected articles. The primary focus was to ensure that each paper involved the use of machine learning to analyze a social media platform, with an emphasis on suicidality or suicide. One reviewer then analyzed the full text of these remaining studies. The objective was to identify and remove any papers that did not align with the review’s goal or meet the criteria. This step further refined the selection of studies, narrowing down the list to those most relevant to the topic at hand.

Data Collection and Data Extraction

One reviewer was responsible for extracting the relevant data from the remaining studies. A pre-defined Excel template was used to collect and organize the extracted data, allowing for an efficient and structured approach to the analysis. The Excel template created by the reviewer consisted of several columns, each representing a key aspect of the methodologies employed in the selected studies, allowing for a systematic and consistent comparison of the various approaches taken by the researchers.

The following information was extracted from the selected articles:

- Author
- Publication year
- Social Platform
- Data/sample size
- Direct contact with subjects (y/n)
- Data gathering method
- Annotation method
- Machine Learning model + results
- External validation (y/n)
- Part of suicide that was assessed (risk, complete suicide, attempt, ideation)
- Clinical professional involved (y/n)
- The main goal of the study

Results

Study Selection

A total of 1137 records were found in databases searching and an additional 25 records were found by searching references of existing systematic reviews on the subject. After removing duplicates, 922 records were screened from which 66 full-text articles were reviewed. Ultimately, a total of 32 studies were included in this review.

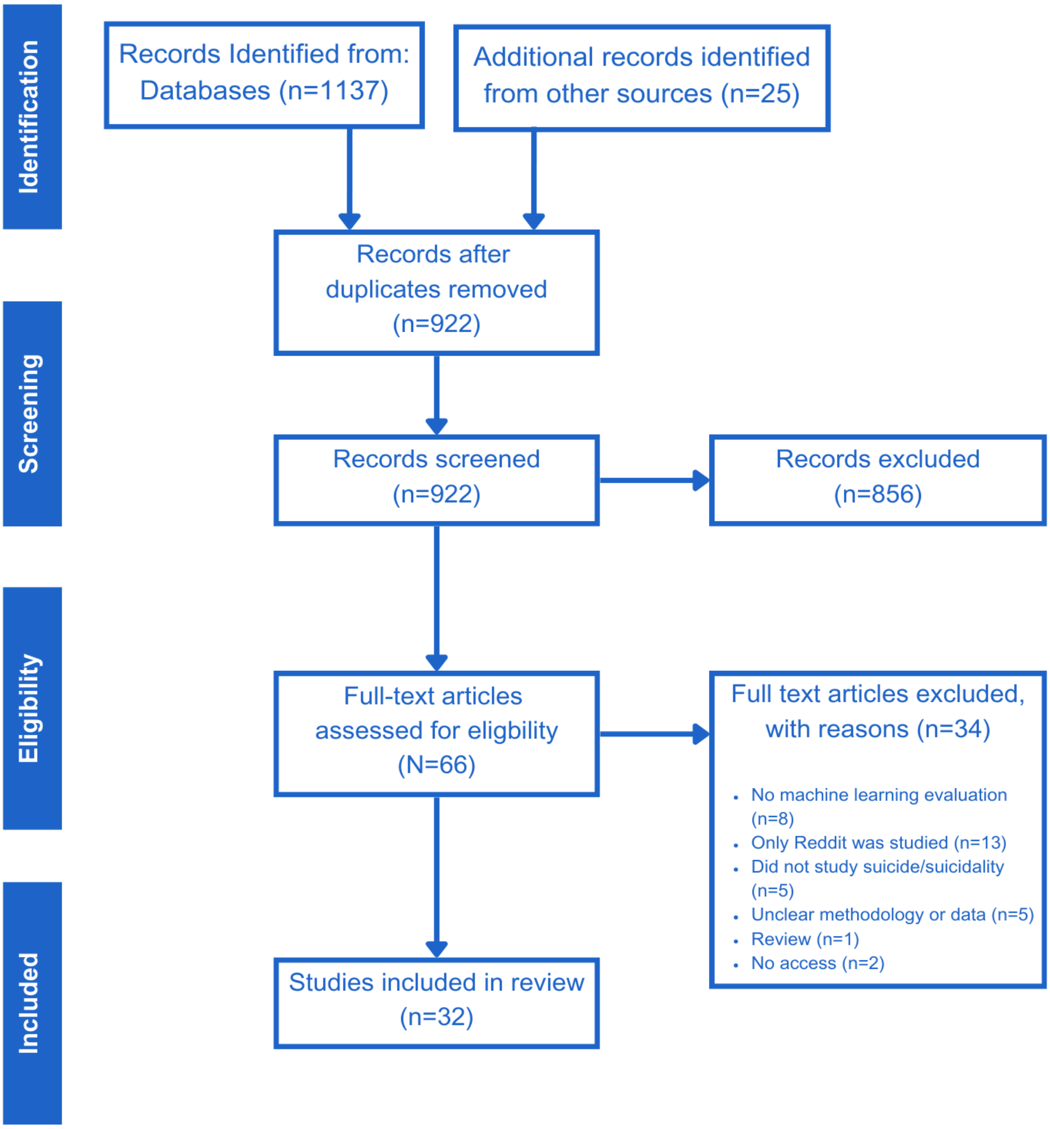

Identification
Screening
Eligibility
Included
Records Identified from: Databases (n=1137)
Additional records identified from other sources (n=25)
Records after duplicates removed (n=922)
Records screened (n=922)
Records excluded (n=856)
Full-text articles assessed for eligbility (N=66)
Full text articles excluded, with reasons (n=34)
• No machine learning evaluation (n=8)
• Only Reddit was studied (n=13)
• Did not study suicide/suicidality (n=5)
• Unclear methodology or data (n=5)
• Review (n=1)
• No access (n=2)
Studies included in review (n=32)

Figure 4. Flowchart of study selection (96).

Study Characteristics

For a comprehensive summary of each study's essential characteristics, please refer to Supplementary Table 1 in the addenda section of this review. The details of these characteristics are also elaborated upon below.

Publication Year:

Figure 5 represents a visualisation of the publication years of the studies. The majority of the studies, approxamitaley 72%, were published from 2020 and onwards with relatively fewer papers from 2019 and earlier. For year 2023, data remains incomplete because the search process was concluded in April 2023, with only a limited number of studies available online up to that point.

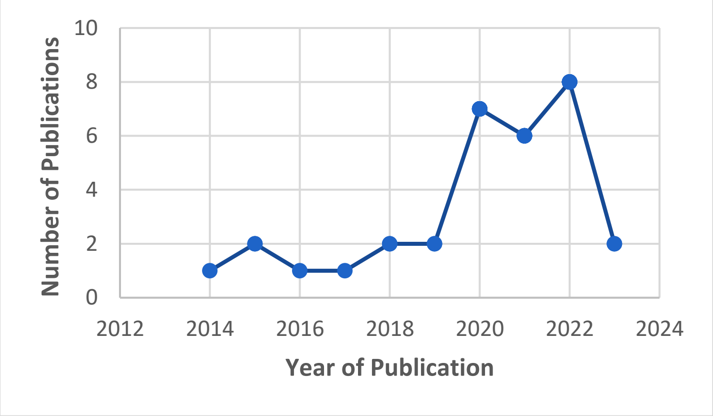


Figure 5. Year-wise distribution of studies.

Data:

The most used social media platform is Twitter, featuring in roughly 69% of the studies (97-118). Weibo is the second most used social media platform, featuring in 7 studies (119-125). Instagram, Facebook, and other social media platforms such as Tumblr, YouTube and Vkontakte, a Russian social media platform, are less frequently used (101) (107) (110) (126) (127) (128). Figure 6 shows a visualization depicting the frequency of each social media platform's usage. Two studies adopted a user-centric approach that involved actively identifying and collecting data from multiple social media platforms on which a specific user had posted content (107) (127).

If studies did not use an equivalent control group to match their study group, the specific individual count is provided. In terms of sample size, it is important to note that most studies only looked at isolated individual social media postings/messages. Only 15 studies or roughly 47% took a broader perspective, considering not just the content but also linking them to the users behind the posts (99) (100) (101) (103) (107) (114) (116) (119)

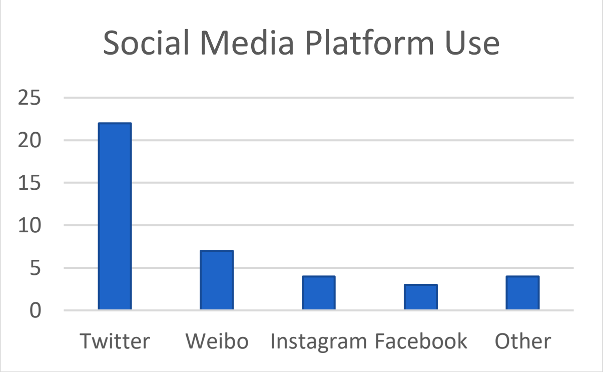


Figure 6. Social media platform usage in included studies.

(121) (122) (124) (125) (126) (127) (128). The variation in data volume is also quite noticeable, ranging from 102 to 4,031,020 postings.

Detailed demographic information about gender or age of users studied was not provided in Supplementary Table 1. Out of the included studies, only four studies provided data regarding the age of the subjects, with an average age of approximately 26 years within this set of studies (101) (121) (126) (128). Similarly, gender distribution information was limited, with just five studies mentioning it and indicating an average distribution of roughly 3:1 (female:male) (99) (101) (121) (126) (128).

Method:

In examining the methodology, the first consideration was whether the studies included any form of psychological evaluation of the subjects, through a questionnaire or interview. Only five studies have mentioned incorporating a psychological evaluation of the subjects (99) (119) (121) (126) (128). Four of the included studies incorporated data of people who had attempted suicide or passed away due to suicide (101) (107) (122) (127). These studies did not incorporate a psychological evaluation.

Regarding the annotation process, a significant portion of the studies, roughly 53%, manually annotated the data (97) (98) (100) (102) (104) (105) (108) (109) (111) (112) (113) (114) (115) (117) (120) (122) (124). This was carried out by the researchers themselves or by clinicians or psychologists. In addition, some studies used automatic linguistic analysis tools, such as LIWC and HowNet, or a sentiment analysis tool like VADER (99) (103) (104) (106) (117) (119) (120) (121) (122) (123) (125) (126). Notably, LIWC stood out as the most common method among these three options. Two studies used pre-analyzed data sourced from previous research (110) (118). One study used a neural network to analyze and annotate their data (116). Lastly, four studies did not disclose a specific annotation method. Among these, three studies relied on categorizing data based on whether the social media user had a personal history of suicide attempts or had passed away due to suicide (101) (107) (127). Additionally, one study categorized their data based on the results of a user's psychological evaluation (128).

The five most common machine learning models used are listed in figure 7. Support vector machine was the most common machine learning model being used in 15 studies (98) (100) (102) (108) (109) (112) (113) (114) (107) (119) (122) (124) (127) (115) (118). It is important to note that many of these studies didn't limit themselves to just one model. Instead, they explored multiple machine learning models. Seven studies developed custom machine learning models, either by building upon existing classification models or by combining several commonly used machine learning models to create an ensemble model (97) (101) (108) (117) (120) (123) (128).

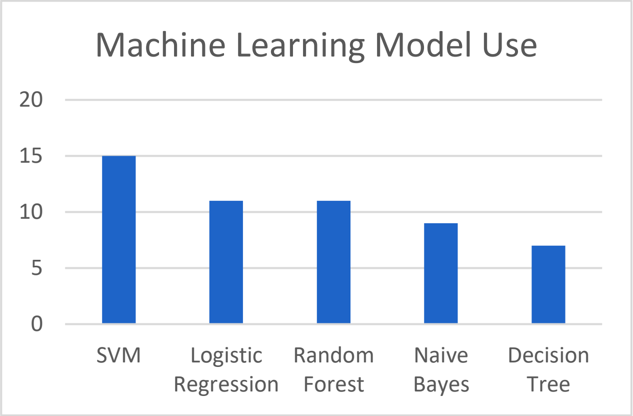


Figure 7. Machine learning model usage in included studies.

Outcome:

In terms of study outcomes, 22 studies, or roughly 69%, primarily focused on detecting instances of suicidal ideation (97) (98) (100) (102) (103) (104) (106) (108) (110-118) (120) (122) (123) (125) (126). Additionally, seven studies solely evaluated suicide risk (101) (107) (109) (119) (124) (127) (128). One study took a dual approach, examining both suicidal ideation and suicide risk (121). Lastly, two studies addressed suicidality as the core point of their research (99) (105).

Categorization Model:

Due to heterogeneity of the included studies, a proposed categorization is listed below in figure 8. This categorization is not based on existing literature but instead centers around key differences in study methodologies. This proposed approach of classifying the studies into groups facilitates the comparison of results of all included studies. Figure 9 shows the distribution of the number of studies in each group. Group 1 consists of studies that focused on utilizing machine learning to

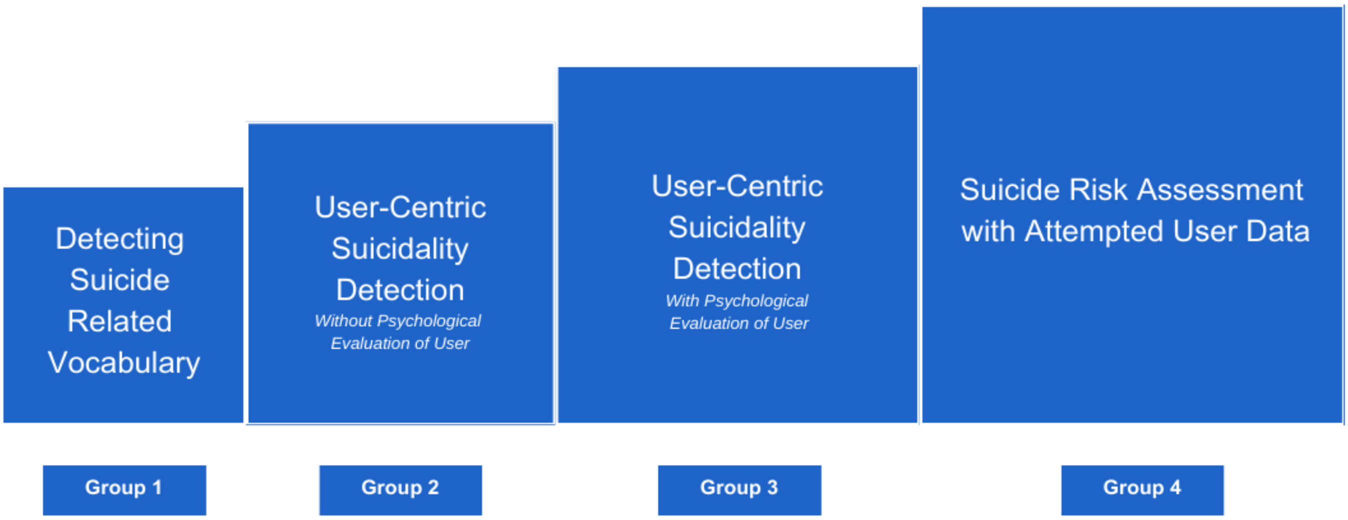

Detecting Suicide Related Vocabulary
User-Centric Suicidality Detection
Without Psychological Evaluation of User
User-Centric Suicidality Detection
With Psychological Evaluation of User
Suicide Risk Assessment with Attempted User Data
Group 1
Group 2
Group 3
Group 4

Figure 8. Classification of studies into four methodological groups.

detect whether a post is a form of suicidal ideation or contains linguistic characteristics related to suicide. Studies within group 2 employed machine learning models to detect whether a social media user is suicidal or at risk of attempting suicide, based on their social media postings. Similarly to group 2, group 3 also attempted to assess the risk of a social media user, additionally employing psychological evaluations of users an interview or web-based surveys. In contrast, group 4 consists of studies that evaluated the user's risk of suicide using data from individuals who had previously attempted suicide or had passed away due to suicide.

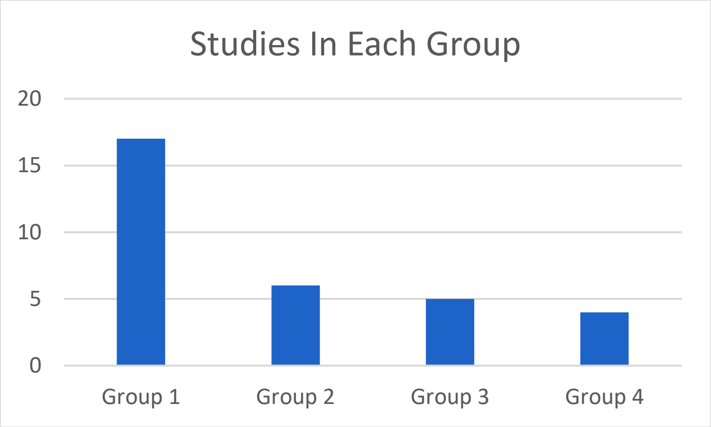


Figure 9. Distribution of studies in each group.

Results of Individual Studies

In Supplementary Table 2 the best performing model mentioned in each study is listed with its corresponding results. The details of all outcomes are discussed per group.

Group 1

More than half of the studies, about 53%, belong to this group. Out of these studies, only three reported an AUC. Baghdadi et al. (2022) reported the highest AUC (97). Employing the BERT model, they achieved an AUC, accuracy, and F1-score of 0.911, 0.9606 and 0.9586, respectively. Haque et al. (2022) used a Random Forest model to achieve an AUC, accuracy and F1-score of 0.92, 0.93 and 0.92, respectively (104). Jung et al. (2023), using a Gradient Boosting Machine, reported the lowest AUC score at 0.907 and an F1-score of 0.846 (105). When considering accuracy and F1-score, Rabani et al. (2020) achieved the highest results in this category (111) (112). They reached an accuracy of 0.985 using the Random Forest model and an F1-score of 0.97 with a Decision Tree model in another study. Rezig et al. (2021) also reported a high accuracy of 0.974 using a Logistic Regression model (115). Kancharapu et al. (2022) used an LSTM, while

Yatapala et al. (2021) employed an ANN, resulting in an accuracy of 0.87 and 0.8676 respectively (106) (115). Liu et al. (2022) achieved an accuracy of 0.8061 by using an ensemble model (123). O'Dea et al. (2015) used an SVM model while Ramachandran et al. (2020) employed a Logistic Regression model, and they achieved an accuracy of 0.76 and 0.763 respectively (109) (113). Parraga et al. (2019) reported the lowest accuracy in this category, reaching 0.48, alongside an F1-score of 0.87 (110). When looking at F1-score, Bhattacharya et al. (2021) reports the second highest F1-score of 0.94 using a GRU/LSTM model (98). Schoene et al. (2022) used a Feature GCN and reported an F1-score of 0.91 (117). Kancharapu et al. (2022) and Du et al. (2018) both achieved an F1-score of 0.83 using an LSTM and CNN, respectively (102) (106). Liu et al. (2022) and Fu et al. (2021) both used an Ensemble Model and achieved an F1-score of 0.792 and 0. 7798, respectively (123) (120). Metzler et al. (2022) achieved an F1-score of 0.55 using XLNet, while the lowest F1-score was reported by Ramachandran et al (2020) (108) (113).

Group 2

The highest AUC in this group was achieved by Ramirez et al. (2020) (114). They employed an SVM model which resulted in an AUC, accuracy, and F1-score of 0.95, 0.86 and 0.86, respectively. The remaining two studies that reported an AUC are from Pan et al. (2023) and Roy et al. (2020), where they obtained an AUC of 0.82 with a Logistic Regression model and 0.88 with a Random Forest model, respectively (125) (116). Ma et al. (2020) reported both the highest accuracy and F1-score, reaching 0.9181 and 0.9154, respectively, using a dual attention-based suicide risk model (DAM) (124). Chatterjee et al. (2021) reported a slightly lower F1-score of 0.81 alongside an accuracy of 0.87, employing a Logistic Regression model (100). Similarly, Pan et al. (2023) also employed the Logistic Regression model, achieving the lowest F1-score of 0.78 (125). Fodeh et al. (2019) was the only study within this group reporting their results in specificity and sensitivity, with values of 0.839 and 0.912, respectively (103).

Group 3

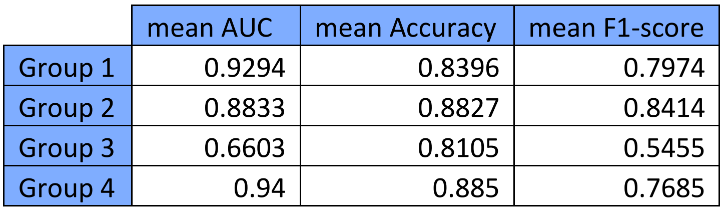
The highest AUC and F1-score within this group was achieved by Lekkas et al. (2021) by utilizing a consensus ensemble machine learning model (126). They achieved an AUC, accuracy, and F1-score of 0.755, 0.702 and 0.741, respectively. Ophir et al. (2020) reported an AUC of 0.746 by employing a Multi Task Model, while the lowest AUC of 0.48 was achieved by Cheng et al. (2022) using an SVM model (128) (119). They also reported a sensitivity of 0.64 and a specificity of 0.32. The group's highest accuracy was achieved by Braithwaite et al. (2016) with an accuracy of 0.919

(99). Braithwaite et al. (2016) also reported a high specificity of 0.97 and relatively lower sensitivity of 0.53. Finally, Guan at al. (2023) reported the lowest F1-score at 0.35 (121).

<u>Group 4</u>

Coppersmith et al. (2018) was the only study to report an AUC, achieving an impressive score of 0.94 through a self-developed model (101). Huang et al. (2014) reached an accuracy of 0.94 by using an SVM model, whereas Meraliyev et al. (2021) achieved an accuracy of 0.83 with a Multinomial Naïve Bayes model (122) (127). Finally, Mbarek et al. (2022) reported an F1-score of 0.854 by employing a Random Forest model, while Huang et al. (2014) achieved an F1-score of 0.683 (107).

Table 3 presents a summary of the mean AUC, Accuracy and F1-score for each group. Group 4 has the highest mean AUC, achieving a mean AUC of 0.94. It is important to note that only one study reported findings in AUC within this group. Group 1 has a mean AUC of 0.9294 and group 2 and 3 have a mean AUC of 0.8833 and 0.6603, respectively. The mean accuracy scores across all four groups are relatively close. Group 4 secures the highest mean accuracy at 0.855, while group 2 closely follows with a mean accuracy of 0.8827. Group 1 and 3 achieve a mean accuracy of 0.8396 and 0.8105, respectively. In terms of mean F1-scores, group 2 achieves the highest score of 0.8414. Group 1, 4 and 3 record a mean F1-score of 0.7974, 0.7685 and 0.5455, respectively.



	mean AUC	mean Accuracy	mean F1-score
Group 1	0.9294	0.8396	0.7974
Group 2	0.8833	0.8827	0.8414
Group 3	0.6603	0.8105	0.5455
Group 4	0.94	0.885	0.7685

Table 1. Mean AUC, accuracy, and F1-score of each group.

Discussion

The main goal of this review was to explore and analyze existing research on the use of machine learning models for the detection of suicidal risk on social media. This review looked at a wide variety of studies that investigated suicidality, suicidal ideation, and risk of suicide, offering insights into the potential use of machine learning as a tool for timely detecting social media users at risk. Most studies succeeded in developing models with relatively high accuracies, yet they differ significantly in terms of methodology. The purpose of this discussion is to highlight differences in data, methodological considerations, the inherent heterogeneity among the studies and the interpretation of study results. Broader implications for public health and ethical considerations are also discussed. Lastly, recommendations for future research are proposed, along with an assessment of the limitations encountered during the review process.

Data

Various social media platforms were used as data sources, yet one platform, Twitter, stands out as the most used platform. Even though Facebook is the most popular platform with the most active users across all age groups, Twitter seems to be a more favored choice for data sourcing. This could be explained by the fundamental differences of these platforms. Twitter functions as a microblogging platform where users share short, public messages, while Facebook primarily focuses on personal interactions among users and has stricter privacy controls, making it more challenging to access data. Even though both platforms offered free API access, there are a lot more free-to-use third-party applications available for Twitter to collect specific data, which makes it easier to source data from Twitter. Weibo was the second most frequently used for data sourcing. This is largely because many of the included studies were conducted by Chinese researchers. China has banned major social media platforms such as Facebook, Twitter, Instagram, and others, which means researchers aiming to access social media data from Chinese population can only turn to Weibo. It is also worth noting that popularity of social media platforms changes over time, and new platforms can emerge rapidly, as seen with TikTok. Additionally social media platforms can alter their policy regarding data usage and API access. For instance, Twitter, now known as X, recently ended free access to their API (129). As a result, researchers will now need to pay to collect data from Twitter. Given the global differences in social media usage and the evolving nature of these platforms, developing a universal system for data collection and creating and evaluating machine learning models for social media platforms can become challenging.

In Supplementary Table 1 the distinction between social media users and social media postings was made because not all studies included users in their dataset. Studies from group 1 solely focused on individual postings without linking these postings to the person who created them. Most of these studies acquired their dataset by employing specific keywords related to suicide, searching through numerous social platform posts. Depending on the chosen keywords, the resulting dataset could vary considerably. Additionally, subtler expressions of suicidal thoughts or suicidality could potentially be missed and thus remaining absent from the dataset. Other ways of achieving a dataset included using pre-annotated datasets from previous research. Studies from group 2, 3 and 4 did include users in their sample and adopted a more user-centric approach. Nevertheless, assembling a dataset of social media users who are either suicidal or have a history of suicide attempts poses a considerable challenge. Various approaches were pursued in this regard. Some studies examined comments on posts related to the suicide of a public figure, as such posts sometimes serve as forums for individuals to express their own suicidal feelings (120) (124). This is, however, a highly specific method and may lack generalizability. Another approach involved collecting data from social media postings of well-known celebrities who have committed suicide (107). This method, while limited in scope due to its focus on a specific small subset of the population, also can be criticized since celebrities might behave differently on social media compared to the general public due to their public status. Another method included utilizing external databases containing data from individuals who have committed suicide. One study made use of a project like this called OurDataHelps, which collected social media data to support mental health research (101). Conversely, some studies did not actively search for users with suicidal tendencies or risks, instead, they took a general sample of users. Subsequently, they did a psychological evaluation or manual annotation. Depending on the sample size, this approach could potentially yield a dataset that is more representative of the typical user base of a social media platform when compared to other methods.

Demographic information plays a vital role when it comes to evaluating the representativeness of a sample. Notably, among the 15 studies that adopted a user-centric approach, only six offered demographic insights regarding the study population. Four studies provided details on both age and gender, while one study only shared information about gender, and another shared information about age. The average age reported by the studies that included information about age was 26 years. Consequently, this research may have overlooked a portion of younger social media users, potentially impacting the assessment of at-risk individuals. In terms of gender, on average, females were three times more prevalent than males in the studies that provided information about gender.

This trend underscores the higher representation of females in these samples, which correlates with the fact that females are more likely to experience suicidal ideation. However, it remains crucial to work towards achieving a balanced gender distribution, as an equal representation can help uncover gender-specific insights and enhance the overall validity and applicability of the findings. The limiting reporting of demographic data can be attributed to the challenge of collecting certain information without direct contact with the study population, primarily due to information scarcity on social media and privacy considerations. Nevertheless, the inclusion of detailed demographic information can enhance the interpretation of findings and contribute to a more comprehensive and nuanced analysis.

When considering sample size, it is important to note substantial variations among studies. The number of postings and social media users used by each study varies a lot. The scale of the dataset can greatly affect the learning process of a machine learning model and therefore the outcome of the study. Given the novelty of this research, there is currently no established standard sample size, and with an almost endless pool of potential social media users and postings, determining the correct sample size can be challenging. It's worth noting that a small sample size carries the risk of greater data variability and a limited representation of the broader social media platform’s content and user base. On the other hand, a sample that is excessively large presents its own set of challenges, such as more extensive annotation requirements and the potential risk for the machine learning model to capture noise and anomalies rather than general patterns in the data. These variations in sample size further complicate the comparability of study outcomes.

Methodological Considerations

As mentioned previously the methodologies of all included studies are unique. Each study adopted a distinct approach to annotating data, assembling, and employing machine learning models and reporting the outcomes. This is why an attempt was made to categorize the studies, to facilitate the comparison of the results and, more importantly, to discuss the strengths and limitations of each approach. In this section of the discussion, the characteristics of each method will be further elaborated upon, explaining potential advantages and limitations, while comparing the different approaches among the four groups.

Studies falling within **group 1** solely focused on examining individual social media postings, determining whether machine learning models could accurately classify these postings as suicidal ideation or not. At the outset of the process, data annotation played a key role, especially since there was no direct psychological evaluation of the users who created the individual postings.

Most studies performed a manual annotation, a process in which researchers were tasked with flagging posts as an expression of suicidal ideation or not. This was accomplished through various means, including the utilization of psychological dictionaries to detect suicidal word use or by reaching out to psychologists and clinical professionals to help with the annotation process. Consequently, the annotation of data exhibited variability and was subject to the annotator's personal experience, making replication of the exact annotation difficult. Some studies made use of tools such as LIWC, VADER or TextBlob for annotating their datasets. These tools can analyze the sentiment of written text by examining various linguistic elements and patterns. However there seems to be a lack of guidelines concerning the minimum number of words needed by these tools to provide a consistent and correct analysis. This can prompt questions about whether social media postings offer sufficient linguistic context for these tools to provide a valid sentiment analysis. These represent various approaches to obtaining an annotated dataset. However, distinguishing whether a text genuinely indicates suicidal ideation rather than a sarcastic, non-serious remark, can be highly challenging, especially without further context. This challenge persists even with the help of professionals or automated programs. Nevertheless, having a consistent and validated method of annotating a dataset is very important to assemble a machine learning model capable of correctly identifying social media postings expressing suicidal ideation. Additionally, there are other critical points to consider in this approach. Firstly, as mentioned earlier, these studies do not link any posting to the user who created it. Consequently, a lot of potential context is lost, including whether the user frequently expresses other signs that are linked to suicidality. Furthermore, suicidality is a very complex phenomenon, leading to questions about whether a single instance of suicidal ideation in a post genuinely corresponds to a heightened suicide risk or if it might simply be an impulsive form of emotional expression.

While this general method presents an approach with relatively low dimensional data, which can lead to more consistent machine learning results, it suffers from a notable deficit of context and information. This deficiency makes it difficult to evaluate whether the employed machine learning model can be used to correctly detect suicidal ideation or suicidality in social media postings and determine whether the social media user is at-risk.

Studies within **group 2** went a step further, exploring not only isolated postings, but a series of posts linked to their respective authors. Taking on this more user-centric approach, these studies attempted to develop machine learning models capable of identifying a social media user is suicidal or at-risk based on their posts. Given the absence of a psychological evaluation of the users, these studies relied on annotation methods similar to those in group 1. Half of these studies manually marked users as suicidal idolators or suicidal based on posted content and profile information. Additionally, one study utilized a Chinese version of the LIWC tool, and two studies employed machine learning models to further analyze social media postings for annotation (103) (125) (116). Machine learning can help in efficiently navigating through a large dataset and clustering vast datasets and provides a broader perspective for analyzing unannotated data. However, it is worth noting that this is highly dependent on the quality and amount of data. Although these models may unveil new patterns, it’s essential to validate them before considering them for practical use.

In this group several studies aimed to expand their focus beyond linguistic features related to suicide. Notably, two studies included image postings into their dataset, and furthermore, two studies delved deeper into leveraging the user-centric approach (124) (114). For instance, Chatterjee et al not only examined the mean number of words per posting but also explored other user-based characteristics such as time of posting (100). Additionally, Ramirez et al. looked at relational characteristics, investigating user interactions through metrics such as number of followers and users followed (114). This means that a user-centric approach offers more context and information for machine learning models to take advantage of when classifying users. Nonetheless, the effectiveness of this method still heavily relies on the quality of data analyzing and annotating, given the absence of validated psychological evaluations for individual users.

Studies in **group 3** introduced contact with the observed social media users to perform psychological evaluations on each user. Various screening tools were used such as the Suicide Probability Scale, Depression Anxiety Stress Scales-21, Depressive Symptom Inventory-Suicide Subscale, and more. These screening tools can be employed to provide more comprehensive clinical insights into individual users, ensuring a more accurate categorization and enhancing the quality of training data for machine learning models. Furthermore, this was often combined with the LIWC text analysis tool, offering a multitude of features to train the machine learning models with. However, it is essential to recognize the potential limitations of these screening tools, particularly when they are used without face-to-face interaction. The questionnaires were often deployed as online surveys, lacking the context provided by clinical professionals. However, one study conducted an online interview via the examined social media platform, presenting a more personal yet labor-intensive approach (128). These limitations might affect the learning data, as its reliability hinges on the effectiveness of the psychological screening tool used. As previously noted, the development of accurate tools to screen social media users for suicidality depends on obtaining correct and validated training data. Furthermore, this method relies on the voluntary participation of social media users in research, potentially excluding individuals with depressive symptoms or suicidal thoughts who might lack the motivation to participate in these studies.

The last group adopted a different approach compared to the prior three in terms of data sourcing. Studies in **group 4** collected social media data from individuals who had attempted suicide or had passed away due to suicide. Most studies within this group, apart from Huang et al., refrained from annotation because the data was already linked to individuals involved in suicide attempts or completed suicides. Conversely, Huang et al. did conduct an annotation process, separating true suicidal posts from normal posts (122). Their objective was to develop a machine learning model that can detect true suicidal posts, which they sourced from confirmed suicidal cases.

This proposed method offers a notable advantage similar to the method of group 3. It does not rely on visible expressions of suicidal ideation, but rather enables machine learning models to potentially discover subtle patterns and behaviors in the online activities of at-risk individuals. These patterns might not be obvious to researchers or linguistic analysis tools. However, the particular advantage of this method over that of group 3 lies in its independence from screening tools, which might not fully capture all nuanced facets and expressions of suicidality. Screening tools might indicate suicide risk without certainty of actual suicide, presenting a potential limitation in comparison to the approach of group 4, which directly examines individuals who have attempted or committed suicide. Furthermore, it does not hinge on the participation of individuals as it relies on a more retrospective approach. However, a significant challenge mentioned by each is the scarcity of data. As previously mentioned, assembling a dataset featuring confirmed cases of suicide and their corresponding social media data is incredibly difficult. Various methods were utilized to obtain such data, including collecting online suicide notes and matching them with social media accounts, identifying accounts that publicly disclose suicide attempts on their profile or posts, using external databases such as OurDataHelps, employing data from previous research, or collecting data from well-known public figures who committed suicide. It's evident that these approaches differ in terms of data's reliability and representativeness. Additionally, the retrospective nature of this approach might overlook the rapidly evolving landscape of social media platforms.

One aspect all studies had in common was the lack of external evaluation of the machine learning models. Typically, during model development, a section of the dataset was employed to train the model, and another portion was dedicated to assessing its performance. Cross-validation, as previously mentioned, plays an important role in determining the model's validity. Evaluating whether the machine learning model remains capable of accurately identifying suicidal ideation or suicidality in other data sets is crucial for determining its potential to be deployed in a real-world scenario.

When it comes to the machine learning model being used, it is difficult to conclude any trend of a more frequently used model in either group. It's important to note that most studies seem to employ and compare various models. This diversity suggests an exploratory approach, testing different models for their effectiveness. This highlights the novelty of machine learning technology given that there is no golden standard within this field of research.

In summary, the methodologies employed by the four groups vary significantly. While studies from group 1 develop machine learning models that can accurately detect text related to suicide, it remains a challenge to correctly label it as suicidal ideation without further context. Group 2 faces a similar problem, exhibiting advancements by considering patterns in user behavior, primarily linked to linguistic markers, yet the classification remains rather subjective and can lack contextual information. Group 3 introduces deeper insights through psychological evaluations and offers a lot of potential as it explores more clinical context. However, these screening tools might have limitations requiring further evaluation for the application in this context. Finally studies within group 4 examine actual suicidal data showing a lot of potential in developing machine learning models to timely detect at risk individuals, but they encounter challenges in acquiring data for model training.

Study Results

Due to the heterogeneity of the methodologies and performance metrics across all studies, it is difficult to compare the outcomes of each study. For example, determining the overall best-performing machine learning model is difficult, as different studies present different top performing models. This can be explained by the fact that the outcome of each model is highly influenced by the training data and model utilization methods, elements that significantly differ between the studies. Comparing the mean outcomes of all groups is also a challenge, especially due to the variety of performance metrics. While some patterns are observable in mean AUC across the groups, suggesting a decreasing trend from group 1 to 3, a lot of studies did not report this metric. Furthermore, group 4 still reported the highest AUC, yet this metric was only used by one study. The mean accuracy and F1-score did however exhibit no consistent pattern across the groups. Notably, studies in group 3 reported relatively lower outcomes in comparison to the other groups. This could potentially be attributed to the complex dimensions introduced by the use of psychological screening tools. While most studies showed promising results in detecting suicidal posts or classifying at-risk users, the heterogenous methodologies and metrics used make it challenging to definitively determine conclusions.

Recommendations for Future Research

As it stands now, there is a lot of potential in developing machine learning models for the deployment on social media platforms to detect suicidal ideation or help in identifying suicidal and at-risk users, which in term could lead to the development of prevention programs. Yet, more research is needed to validate the best approaches. Firstly, increased efforts should be made to collect more representative data. This involves gathering data from a wider demographic range, including age, gender, and ethnicity, ensuring a more comprehensive understanding of potential risk factors and nuances within various social media user groups. Collecting data across multiple platforms from a social media platform might help overcome specific platform differences. Additionally, more focus on data from confirmed suicidal cases and matched controls, despite its challenges, could offer a more robust approach. Furthermore, standardizing annotation protocols is vital if an approach requires annotation. Involving clinical professionals such as psychiatrists and psychologists can significantly enhance the training data for machine learning models. Current research predominantly adopts an informative technological perspective. It is therefore important to augment this approach with a more profound clinical perspective in future studies. Using validated screening tools might also offer a better chance in correctly categorizing the social media users for training data, however more exploratory research might be needed to assess which screening tool is most effective in this context. Future research should continue to explore various machine learning models, given the absence of a golden standard model, encouraging comparisons and assessments. Moreover, more effort is required to employ similar performance metrics that facilitate cross-study comparisons. Additionally, validating each model demands external validation using data from different datasets. Lastly, selecting the optimal research focus is essential. First of all, more effort is needed to assess whether the developed machine learning models can effectively discern expressions of suicidal ideation from neutral or sarcastic postings related to suicide. This highly depends on the correct annotation of the dataset. Additionally, further exploration is needed to understand the correlation between expressing suicidal thoughts on social media and the actual risk of attempting suicide, as critical risk factors beyond social media expressions might exist but remain less noticeable. Focusing on endpoints such as suicidality and the risk of suicide might offer more definitive outcomes compared to focusing solely on suicidal ideation, suggesting stronger foundations for the development of actual screening and prevention tools.

Implications and Ethical Considerations

Using machine learning models for early identification of users at risk of suicide presents the potential for the development of automatic screening tools. These tools could serve as the foundation for developing preventive programs that aim to identify and assist at-risk users, possibly by filtering out harmful content and providing easier access to support resources. However, the ethical use of social media data for research demands a thoughtful approach that prioritizes user consent and ensures the protection of individual privacy. Additionally, actively influencing the behavior of social media users should be approached with caution. Social media platforms currently use complex algorithms to influence the behavior of their users, a practice that has received substantial criticism. This not only raises serious privacy concerns but also led to legal action against companies like Meta, responsible for Facebook and Instagram, over allegations that their algorithms negatively affect the mental well-being of adolescents (130). Therefore, it is crucial for social media platforms to prioritize the mental health of their users. This situation offers an opportunity for collaboration between these platforms and researchers to address the adverse impacts of algorithms and strive toward more positive and supportive uses of these technologies.

Limitation of This Review

Due to the novelty of the topic, this review didn’t implement strict exclusion criteria. This could have resulted in the inclusion of studies with very differing methodologies, which might have affected the general consistency and comparability of the findings. Consequently, the main purpose of this study was to provide a comprehensive overview of existing methods, outlining their various strengths and weaknesses, and offering suggestions for future research. The variety in methodologies employed by the included studies might have introduced subjectivity that could have influenced the review's conclusions and recommendations. Additionally, this review's scope is limited to existing literature up to a specific timeframe, while the rapidly evolving nature of social media and technology could make the conclusions and recommendations time sensitive.

Conclusion

In conclusion, 32 studies that developed and evaluated machine learning models to detect suicidal ideation, suicidality or suicide risk were assessed in this review. The wide variety of methodologies employed by these studies led to a proposed classification model to understand their diverse approaches, highlighting the complexity of using machine learning models to interpret social media data to identify at-risk individuals. Current research already provides valuable insights into online behaviors associated with suicidality. Nevertheless, although showing substantial potential, further exploration and refinement are necessary, especially considering the ethical, privacy and representational concerns before considering the deployment of machine learning models as screening tools and for the development of preventative projects. This review emphasizes the need for more robust and standardized methodologies and diverse, clinically informed approaches to fully address the mental health complexities in social media landscapes.

References

1. Suicide: World Health Organization; 2021 [updated June 17. Available from: https://www.who.int/news-room/fact-sheets/detail/suicide.

2. Cerel J. Exposure to Suicide and Identification as Survivor. 2013 Jan:413-9.

3. Carpiniello B, Pinna F. The Reciprocal Relationship between Suicidality and Stigma. Frontiers in psychiatry. 2017;8:35.

4. Association AP. Diagnostic and statistical manual of mental disorders: DSM-5™, 5th ed. Arlington, VA, US: American Psychiatric Publishing, Inc.; 2013. xliv, 947-xliv, p.

5. Klonsky ED, May AM, Saffer BY. Suicide, Suicide Attempts, and Suicidal Ideation. Annual Review of Clinical Psychology. 2016;12(1):307-30.

6. Nock MK, Borges G, Bromet EJ, Alonso J, Angermeyer M, Beautrais A, et al. Cross-national prevalence and risk factors for suicidal ideation, plans and attempts. The British Journal of Psychiatry. 2008;192(2):98-105.

7. Piscopo KD. Suicidality and Death by Suicide Among Middle-Aged Adults in the United States. The CBHSQ Report. Rockville (MD): Substance Abuse and Mental Health Services Administration (US); 2013. p. 1-27.

8. Rossom RC, Coleman KJ, Ahmedani BK, Beck A, Johnson E, Oliver M, Simon GE. Suicidal ideation reported on the PHQ9 and risk of suicidal behavior across age groups. J Affect Disord. 2017;215:77-84.

9. Mclean JaM, Margaret and Platt, Stephen and Harris, Fiona and Jepson, Ruth. Risk and Protective Factors for Suicide and Suicidal Behaviour: A Literature Review. In: Care HaC, editor.: The Scottish Government 2008.

10. Mark L, Samm A, Tooding LM, Sisask M, Aasvee K, Zaborskis A, et al. Suicidal ideation, risk factors, and communication with parents. An HBSC study on school children in Estonia, Lithuania, and Luxembourg. Crisis. 2013;34(1):3-12.

11. Miranda-Mendizabal A, Castellví P, Parés-Badell O, Alayo I, Almenara J, Alonso I, et al. Gender differences in suicidal behavior in adolescents and young adults: systematic review and meta-analysis of longitudinal studies. Int J Public Health. 2019;64(2):265-83.

12. Maguen S, Shipherd JC. Suicide risk among transgender individuals. Psychology & Sexuality. 2010;1:34 - 43.

13. di Giacomo E, Krausz M, Colmegna F, Aspesi F, Clerici M. Estimating the Risk of Attempted Suicide Among Sexual Minority Youths: A Systematic Review and Meta-analysis. JAMA Pediatr. 2018;172(12):1145-52.

14. Oh H, Stickley A, Koyanagi A, Yau R, DeVylder JE. Discrimination and suicidality among racial and ethnic minorities in the United States. J Affect Disord. 2019;245:517-23.

15. Zalsman G, Hawton K, Wasserman D, van Heeringen K, Arensman E, Sarchiapone M, et al. Suicide prevention strategies revisited: 10-year systematic review. Lancet Psychiatry. 2016;3(7):646-59.

16. Ajdacic-Gross V, Weiss MG, Ring M, Hepp U, Bopp M, Gutzwiller F, Rössler W. Methods of suicide: international suicide patterns derived from the WHO mortality database. Bull World Health Organ. 2008;86(9):726-32.

17. Méndez-Bustos P, Calati R, Rubio-Ramírez F, Olié E, Courtet P, Lopez-Castroman J. Effectiveness of Psychotherapy on Suicidal Risk: A Systematic Review of Observational Studies. Front Psychol. 2019;10:277.

18. Wilkinson ST, Ballard ED, Bloch MH, Mathew SJ, Murrough JW, Feder A, et al. The Effect of a Single Dose of Intravenous Ketamine on Suicidal Ideation: A Systematic Review and Individual Participant Data Meta-Analysis. Am J Psychiatry. 2018;175(2):150-8.

19. Ahmedani BK, Simon GE, Stewart C, Beck A, Waitzfelder BE, Rossom R, et al. Health Care Contacts in the Year Before Suicide Death. Journal of General Internal Medicine. 2014;29(6):870-7.

20. Giddens JM, Sheehan KH, Sheehan DV. The Columbia-Suicide Severity Rating Scale (C-SSRS): Has the "Gold Standard" Become a Liability? Innov Clin Neurosci. 2014;11(9-10):66-80.

21. Andreotti ET, Ipuchima JR, Cazella SC, Beria P, Bortoncello CF, Silveira RC, Ferrão YA. Instruments to assess suicide risk: a systematic review. Trends Psychiatry Psychother. 2020;42(3):276-81.

22. Morese R, Gruebner O, Sykora M, Elayan S, Fadda M, Albanese E. Detecting Suicide Ideation in the Era of Social Media: The Population Neuroscience Perspective. Frontiers in psychiatry. 2022;13:652167.

23. Kemp S. Digital 2022: Global Overview Report: DATAREPORTAL; 2022 [updated January 26 2022. Available from: https://datareportal.com/reports/digital-2022-global-overview-report.

24. Hall M. Facebook. Encyclopedia Britannica: Britannica; 2023.

25. Eldridge A. Instagram. Encyclopedia Britannica: Britannica; 2023.

26. Britannica TEoE. TikTok. Encyclopedia Britannica: Britannica; 2023.

27. Britannica TEoE. X. Encyclopedia Britannica: Britannica; 2023.

28. contributors W. Weibo. Wikipedia: Wikipedia, The Free Encyclopedia; 2023.

29. Hosch WL. YouTube. Encyclopedia Britannica: Britannica; 2023.

30. Robinson J, Cox G, Bailey E, Hetrick S, Rodrigues M, Fisher S, Herrman H. Social media and suicide prevention: a systematic review. Early Interv Psychiatry. 2016;10(2):103-21.

31. Luxton DD, June JD, Fairall JM. Social media and suicide: a public health perspective. Am J Public Health. 2012;102 Suppl 2(Suppl 2):S195-200.

32. Silenzio VM, Duberstein PR, Tang W, Lu N, Tu X, Homan CM. Connecting the invisible dots: reaching lesbian, gay, and bisexual adolescents and young adults at risk for suicide through online social networks. Soc Sci Med. 2009;69(3):469-74.

33. O'Connor RC, Portzky G. Looking to the Future: A Synthesis of New Developments and Challenges in Suicide Research and Prevention. Front Psychol. 2018;9:2139.

34. Lopez-Castroman J, Moulahi B, Azé J, Bringay S, Deninotti J, Guillaume S, Baca-Garcia E. Mining social networks to improve suicide prevention: A scoping review. Journal of Neuroscience Research. 2020;98(4):616-25.

35. Jashinsky J, Burton SH, Hanson CL, West J, Giraud-Carrier C, Barnes MD, Argyle T. Tracking suicide risk factors through Twitter in the US. Crisis. 2014;35(1):51-9.

36. Hamm MP, Newton AS, Chisholm A, Shulhan J, Milne A, Sundar P, et al. Prevalence and Effect of Cyberbullying on Children and Young People: A Scoping Review of Social Media Studies. JAMA Pediatr. 2015;169(8):770-7.

37. Brown M, Barraclough B. Epidemiology of suicide pacts in England and Wales, 1988-92. Bmj. 1997;315(7103):286-7.

38. Naito A. Internet Suicide in Japan: Implications for Child and Adolescent Mental Health. Clinical Child Psychology and Psychiatry. 2007;12(4):583-97.

39. Lee SY, Kwon Y. Twitter as a place where people meet to make suicide pacts. Public Health. 2018;159:21-6.

40. Choi YJ, Oh H. Does Media Coverage of a Celebrity Suicide Trigger Copycat Suicides?: Evidence from Korean Cases. Journal of Media Economics. 2016;29(2):92-105.

41. Kogler V, Noyon. A. The Werther effect – About the handling of suicide in the media: Open Access Government; 2018. Available from: https://www.openaccessgovernment.org/the-werther-effect/42915/.

42. World Health O, International Association for Suicide P. Preventing suicide: a resource for media professionals. Geneva: World Health Organization; 2017 2017. Contract No.: WHO/MSD/MER/17.5.

43. Mitchell KJ, Wells M, Priebe G, Ybarra ML. Exposure to websites that encourage self-harm and suicide: prevalence rates and association with actual thoughts of self-harm and thoughts of suicide in the United States. J Adolesc. 2014;37(8):1335-44.

44. Media W. Facebook News Feed Algorithm History [Internet]. UTAH (US)2023 [updated March 9 2023. https://wallaroomedia.com/facebook-newsfeed-algorithm-history/]. Available from: https://wallaroomedia.com/facebook-newsfeed-algorithm-history/.

45. Hoang LN. Science Communication Desperately Needs More Aligned Recommendation Algorithms. Frontiers in Communication. 2020;5.

46. Abul-Fottouh D, Song MY, Gruzd A. Examining algorithmic biases in YouTube's recommendations of vaccine videos. International Journal of Medical Informatics. 2020;140:104175.

47. Copeland BJ. What is artificial intelligence? Encyclopedia Britannica: Britannica; 2022.

48. Brown S. Machine learning, explained 2021. Available from: https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained.

49. Nasteski V. An overview of the supervised machine learning methods. HORIZONSB. 2017;4:51-62.

50. Bishop CM. Pattern Recognition and Machine Learning. 1 ed. New York: Springer New York, NY; 2006. 738 p.

51. F.Y O, AkinsolaJ.E. T, Awodele O, HinmikaiyeJ. O, Olakanmi O, Akinjobi J. Supervised Machine Learning Algorithms: Classification and Comparison. International Journal of Computer Trends and Technology. 2017;48:128-38.

52. Dietterich TG, editor Ensemble Methods in Machine Learning. Multiple Classifier Systems; 2000 2000//; Berlin, Heidelberg: Springer Berlin Heidelberg.

53. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Information Processing & Management. 2009;45(4):427-37.

54. Sokolova M, Japkowicz N, Szpakowicz S, editors. Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. AI 2006: Advances in Artificial Intelligence; 2006 2006//; Berlin, Heidelberg: Springer Berlin Heidelberg.

55. Megahed FM, Chen Y-J, Megahed A, Ong Y, Altman N, Krzywinski M. The class imbalance problem. Nature Methods. 2021;18(11):1270-2.

56. Powers D. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. Mach Learn Technol. 2008;2.

57. Zhang E, Zhang Y. F-Measure. In: Liu L, ÖZsu MT, editors. Encyclopedia of Database Systems. Boston, MA: Springer US; 2009. p. 1147-.

58. Fawcett T. Introduction to ROC analysis. Pattern Recognition Letters. 2006;27:861-74.

59. Wiens J, Shenoy ES. Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology. Clin Infect Dis. 2018;66(1):149-53.

60. Colubri A, Silver T, Fradet T, Retzepi K, Fry B, Sabeti P. Transforming Clinical Data into Actionable Prognosis Models: Machine-Learning Framework and Field-Deployable App to Predict Outcome of Ebola Patients. PLoS Negl Trop Dis. 2016;10(3):e0004549.

61. Wiens J, Campbell WN, Franklin ES, Guttag JV, Horvitz E. Learning Data-Driven Patient Risk Stratification Models for Clostridium difficile. Open Forum Infectious Diseases. 2014;1(2).

62. Magar R, Yadav P, Barati Farimani A. Potential neutralizing antibodies discovered for novel corona virus using machine learning. Scientific Reports. 2021;11(1):5261.

63. Gozes O, Frid-Adar M, Greenspan H, Browning PD, Zhang H, Ji W-B, et al. Rapid AI Development Cycle for the Coronavirus (COVID-19) Pandemic: Initial Results for Automated Detection & Patient Monitoring using Deep Learning CT Image Analysis. ArXiv. 2020;abs/2003.05037.

64. Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine Learning for Medical Imaging. RadioGraphics. 2017;37(2):505-15.

65. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet. 2015;16(6):321-32.

66. McKinney BA, Reif DM, Ritchie MD, Moore JH. Machine learning for detecting gene-gene interactions: a review. Applied bioinformatics. 2006;5(2):77-88.

67. Roden DM, Wilke RA, Kroemer HK, Stein CM. Pharmacogenomics: the genetics of variable drug responses. Circulation. 2011;123(15):1661-70.

68. Sharabiani A, Bress A, Douzali E, Darabi H. Revisiting Warfarin Dosing Using Machine Learning Techniques. Computational and Mathematical Methods in Medicine. 2015;2015:560108.

69. Nguyen VL, Nguyen HD, Cho YS, Kim HS, Han IY, Kim DK, et al. Comparison of multivariate linear regression and a machine learning algorithm developed for prediction of precision warfarin dosing in a Korean population. J Thromb Haemost. 2021;19(7):1676-86.

70. Bzdok D, Altman N, Krzywinski M. Statistics versus machine learning. Nature Methods. 2018;15(4):233-4.

71. Deo RC. Machine Learning in Medicine. Circulation. 2015;132(20):1920-30.

72. Alpaydin E. Introduction To Machine Learning: Third Edition: The MIT Press; 2014 August 22. 640 p.

73. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform. 2018;19(6):1236-46.

74. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. Summit Transl Bioinform. 2010;2010:1-5.

75. Fox F, Aggarwal VR, Whelton H, Johnson O, editors. A Data Quality Framework for Process Mining of Electronic Health Record Data. 2018 IEEE International Conference on Healthcare Informatics (ICHI); 2018 4-7 June 2018.

76. Kaufman S, Rosset S, Perlich C, Stitelman O. Leakage in Data Mining: Formulation, Detection, and Avoidance. Acm T Knowl Discov D. 2012;6(4).

77. Samala R, Chan H-P, Hadjiiski L, Koneru S. Hazards of data leakage in machine learning: a study on classification of breast cancer using deep neural networks2020. 39 p.

78. Zhang S, Bamakan SMH, Qu Q, Li S. Learning for Personalized Medicine: A Comprehensive Review From a Deep Learning Perspective. IEEE Rev Biomed Eng. 2019;12:194-208.

79. Ahmad MA, Teredesai A, Eckert C, editors. Interpretable Machine Learning in Healthcare. 2018 IEEE International Conference on Healthcare Informatics (ICHI); 2018 4-7 June 2018.

80. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. Kdd'15: Proceedings of the 21st Acm Sigkdd International Conference on Knowledge Discovery and Data Mining. 2015:1721-30.

81. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence. 2019;1(5):206-15.

82. Greist JH, Gustafson DH, Stauss FF, Rowse GL, Laughren TP, Chiles JA. Suicide risk prediction: a new approach. Life Threat Behav. 1974;4(4):212-23.

83. Su C, Aseltine R, Doshi R, Chen K, Rogers SC, Wang F. Machine learning for suicide risk prediction in children and adolescents with electronic health records. Transl Psychiatry. 2020;10(1):413.

84. Walsh CG, Ribeiro JD, Franklin JC. Predicting suicide attempts in adolescents with longitudinal clinical data and machine learning. Journal of Child Psychology and Psychiatry. 2018;59(12):1261-70.

85. Linthicum KP, Schafer KM, Ribeiro JD. Machine learning in suicide science: Applications and ethics. Behavioral Sciences & the Law. 2019;37(3):214-22.

86. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PR, Bernstam EV, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. Med Care. 2013;51(8 Suppl 3):S30-7.

87. Pestian JP, Sorter M, Connolly B, Bretonnel Cohen K, McCullumsmith C, Gee JT, et al. A Machine Learning Approach to Identifying the Thought Markers of Suicidal Subjects: A Prospective Multicenter Trial. Suicide Life Threat Behav. 2017;47(1):112-21.

88. Just MA, Pan L, Cherkassky VL, McMakin DL, Cha C, Nock MK, Brent D. Machine learning of neural representations of suicide and emotion concepts identifies suicidal youth. Nat Hum Behav. 2017;1:911-9.

89. Russell MA. Mining the Social Web. 2nd Edition ed. Sebastopol: O'Reilly; 2013.

90. Ducange P, Fazzolari M, Petrocchi M, Vecchio M. An effective Decision Support System for social media listening based on cross-source sentiment analysis models. Engineering Applications of Artificial Intelligence. 2019;78:71-85.

91. Tausczik YR, Pennebaker JW. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. Journal of Language and Social Psychology. 2009;29(1):24-54.

92. Hutto C, Gilbert E. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. Proceedings of the International AAAI Conference on Web and Social Media. 2014;8(1):216-25.

93. Culotta A. Towards detecting influenza epidemics by analyzing Twitter messages. Proceedings of the First Workshop on Social Media Analytics; Washington D.C., District of Columbia: Association for Computing Machinery; 2010. p. 115–22.

94. Gupta M, Bansal A, Jain B, Rochelle J, Oak A, Jalali MS. Whether the weather will help us weather the COVID-19 pandemic: Using machine learning to measure twitter users' perceptions. International journal of medical informatics. 2021;145:104340.

95. The EndNote Team. EndNote. EndNote 21 ed. Philadelphia, PA: Clarivate; 2013.

96. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. Journal of clinical epidemiology. 2009;62(10):e1-34.

97. Baghdadi NA, Malki A, Magdy Balaha H, AbdulAzeem Y, Badawy M, Elhosseini M. An optimized deep learning approach for suicide detection through Arabic tweets. PeerJ Comput Sci. 2022;8:e1070.

98. Bhattacharya D, S. H. K N, S A, editors. Early Detection of Suicidal Tendencies from Text Data using LSTM. 2021 Innovations in Power and Advanced Computing Technologies (i-PACT); 2021 27-29 Nov. 2021.

99. Braithwaite SR, Giraud-Carrier C, West J, Barnes MD, Hanson CL. Validating Machine Learning Algorithms for Twitter Data Against Established Measures of Suicidality. JMIR Ment Health. 2016;3(2):e21.

100. Chatterjee M, Samanta P, Kumar P, Sarkar D, editors. Suicide Ideation Detection using Multiple Feature Analysis from Twitter Data. 2022 IEEE Delhi Section Conference (DELCON); 2022 11-13 Feb. 2022.

101. Coppersmith G, Leary R, Crutchley P, Fine A. Natural Language Processing of Social Media as Screening for Suicide Risk. Biomed Inform Insights. 2018;10:1178222618792860.

102. Du J, Zhang Y, Luo J, Jia Y, Wei Q, Tao C, Xu H. Extracting psychiatric stressors for suicide from social media using deep learning. BMC Med Inform Decis Mak. 2018;18(Suppl 2):43.

103. Fodeh S, Li T, Menczynski K, Burgette T, Harris A, Ilita G, et al., editors. Using Machine Learning Algorithms to Detect Suicide Risk Factors on Twitter. 2019 International Conference on Data Mining Workshops (ICDMW); 2019 8-11 Nov. 2019.

104. Haque R, Islam N, Islam M, Ahsan MM. A Comparative Analysis on Suicidal Ideation Detection Using NLP, Machine, and Deep Learning. Technologies. 2022;10(3):57.

105. Jung W, Kim D, Nam S, Zhu Y. Suicidality Detection on Social Media Using Metadata and Text Feature Extraction and Machine Learning. Arch Suicide Res. 2023;27(1):13-28.

106. Kancharapu R, SriNagesh A, BhanuSridhar M, editors. Prediction of Human Suicidal Tendency based on Social Media using Recurrent Neural Networks through LSTM. 2022 International Conference on Computing, Communication and Power Technology (IC3P); 2022 7-8 Jan. 2022.

107. Mbarek A, Jamoussi S, Hamadou AB. An across online social networks profile building approach: Application to suicidal ideation detection. Future Generation Computer Systems. 2022;133:171-83.

108. Metzler H, Baginski H, Niederkrotenthaler T, Garcia D. Detecting Potentially Harmful and Protective Suicide-Related Content on Twitter: Machine Learning Approach. J Med Internet Res. 2022;24(8):e34705.

109. O'Dea B, Wan S, Batterham PJ, Calear AL, Paris C, Christensen H. Detecting suicidality on Twitter. Internet Interventions. 2015;2(2):183-8.

110. Parraga-Alava J, Caicedo RA, Gómez JM, Inostroza-Ponta M, editors. An Unsupervised Learning Approach for Automatically to Categorize Potential Suicide Messages in Social Media. 2019 38th International Conference of the Chilean Computer Science Society (SCCC); 2019 4-9 Nov. 2019.

111. Rabani S, Khan QR, Khanday A. Detection of Suicidal Ideation on Twitter using Machine Learning & Ensemble Approaches. Baghdad Science Journal. 2020;17:1328-39.

112. Rabani ST, Khan QR, Khanday AMUD, editors. Multi-Class Suicide Risk Prediction on Twitter Using Machine Learning Techniques. 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN); 2020 18-19 Dec. 2020.

113. Ramachandran A, Gadwe A, Poddar D, Satavalekar S, Sahu S, editors. Performance Evaluation of Different Machine Learning Techniques using Twitter Data for Identification of Suicidal Intent. 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC); 2020 2-4 July 2020.

114. Ramírez-Cifuentes D, Freire A, Baeza-Yates R, Puntí J, Medina-Bravo P, Velazquez DA, et al. Detection of Suicidal Ideation on Social Media: Multimodal, Relational, and Behavioral Analysis. J Med Internet Res. 2020;22(7):e17758.

115. Rezig AA, editor A Novel Optimizer Technique for Suicide Prediction In Twitter Environment. 2021 International Conference on Information Systems and Advanced Technologies (ICISAT); 2021 27-28 Dec. 2021.

116. Roy A, Nikolitch K, McGinn R, Jinah S, Klement W, Kaminsky ZA. A machine learning approach predicts future risk to suicidal ideation from social media data. npj Digital Medicine. 2020;3(1):78.

117. Schoene AM, Bojanić L, Nghiem MQ, Hunt IM, Ananiadou S. Classifying Suicide-Related Content and Emotions on Twitter Using Graph Convolutional Neural Networks. IEEE Transactions on Affective Computing. 2023;14(3):1791-802.

118. Yatapala KYDHT, Kumara BTGS, editors. Detection of Suicide Ideation in Twitter using ANN. 2021 6th International Conference on Information Technology Research (ICITR); 2021 1-3 Dec. 2021.

119. Cheng Q, Li TM, Kwok CL, Zhu T, Yip PS. Assessing Suicide Risk and Emotional Distress in Chinese Social Media: A Text Mining and Machine Learning Study. J Med Internet Res. 2017;19(7):e243.

120. Fu G, Song C, Li J, Ma Y, Chen P, Wang R, et al. Distant Supervision for Mental Health Management in Social Media: Suicide Risk Classification System Development Study. J Med Internet Res. 2021;23(8):e26119.

121. Guan L, Hao B, Cheng Q, Yip PS, Zhu T. Identifying Chinese Microblog Users With High Suicide Probability Using Internet-Based Profile and Linguistic Features: Classification Model. JMIR Ment Health. 2015;2(2):e17.

122. Huang X, Zhang L, Chiu D, Liu T, Li X, Zhu T, editors. Detecting Suicidal Ideation in Chinese Microblogs with Psychological Lexicons. 2014 IEEE 11th Intl Conf on Ubiquitous Intelligence and Computing and 2014 IEEE 11th Intl Conf on Autonomic and Trusted Computing and 2014 IEEE 14th Intl Conf on Scalable Computing and Communications and Its Associated Workshops; 2014 9-12 Dec. 2014.

123. Liu J, Shi M, Jiang H. Detecting Suicidal Ideation in Social Media: An Ensemble Method Based on Feature Fusion. Int J Environ Res Public Health. 2022;19(13).

124. Ma Y, Cao Y, editors. Dual Attention based Suicide Risk Detection on Social Media. 2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA); 2020 27-29 June 2020.

125. Pan W, Wang X, Zhou W, Hang B, Guo L. Linguistic Analysis for Identifying Depression and Subsequent Suicidal Ideation on Weibo: Machine Learning Approaches. Int J Environ Res Public Health. 2023;20(3).

126. Lekkas D, Klein RJ, Jacobson NC. Predicting acute suicidal ideation on Instagram using ensemble machine learning models. Internet Interventions. 2021;25:100424.

127. Meraliyev B, Kongratbayev K, Sultanova N, editors. Content Analysis of Extracted Suicide Texts From Social Media Networks by Using Natural Language Processing and Machine Learning Techniques. 2021 IEEE International Conference on Smart Information Systems and Technologies (SIST); 2021 28-30 April 2021.

128. Ophir Y, Tikochinski R, Asterhan CSC, Sisso I, Reichart R. Deep neural networks detect suicide risk from textual facebook posts. Scientific Reports. 2020;10(1):16685.

129. Willingham AJ. Why Twitter users are upset about the platform's latest change | CNN Business. CNN. 2023.

130. Stempel J, Bartz D, Raymond N, Bartz D, Raymond N. Meta's Instagram linked to depression, anxiety, insomnia in kids - US states' lawsuit. Reuters. 2023;Sect. Legal.

Addenda

Supplementary Table 1. Study characteristics.

Study ID	Data			Method			Outcome	Group
Author (Year) (Ref.)	Social Platform	Sample Size (Users)	Sample Size (Postings)	Psychological Evaluation (y/n)	Annotation & Text Analysis	Machine Learning Model	Suicidality/ SI/ Suicide Risk	
Baghdadi et al. (2022) (97)	Twitter	/	2,030	No	Manual	Universal Sentence Encoder, BERT	SI	1
Bhattacharya (2021) (98)	Twitter	/	14,472	No	Manual	Naive Bayes, SVM, Random Forest, Logistic Regression, GRU, LSTM	SI	1
Braithwaite et al. (2016) (99)	Twitter	135	27,000	Yes	LIWC	Decision Tree	Suicidality	3
Chatterjee et al. (2021) (100)	Twitter	Suicidal: 445 Controls: 724	Suicidal: 74,125 Controls: 114,579	No	Manual	Logistic Regression, Random Forest, SVM, Xgboost	SI	2
Cheng et al. (2022) (119)	Weibo	974	/	Yes	SC-LIWC	SVM	Suicide Risk	3
Coppersmith et al. (2018) (101)	Twitter, Facebook, Instagram, Tumblr	418	197,615	No	/	Word Embeddings, Bidirectional LSTM layer, Self-Attention Layer, Linear layer with softmax output	Suicide Risk	4

Du et al. (2018) (102)	Twitter	/	3,262	No	Manual	CNN, SVM, Extra Trees, Random Forest, Logistic Regression, Bi-LSTM	SI	1
Fodeh et al. (2019) (103)	Twitter	3,873	12,066	No	LSA, LDA, NMF	Decision Tree, K-means Clustering	SI	2
Fu et al. (2021) (120)	Weibo	/	15,316	No	Manual + Automatic	BERT, Fine tuning model, Psychology+	SI	1
Guan et al. (2023) (121)	Weibo	909	/	Yes	SCMBWC	Random Forest, Simple Linear Regression	SI + Suicide Risk	3
Haque et al. (2022) (104)	Twitter	/	49,178	No	Manual + VADER, TextBlob	Multinomial Naive Bayes, Logistic Regression, SGD, Random Forest, SVC	SI	1
Huang et al (2014) (122)	Weibo	/	6,714	No	Manual + HowNet	Naive Bayes, Logistic Regression, J48, Random Forest, SMO, SVM	SI	4
Jung et al. (2023) (105)	Twitter	/	20,000	No	Manual	Random Forest, Gradient Boosting Machine	Suicidality	1
Kancharapu et al. (2022) (106)	Twitter	/	20,250	No	VADER	LSTM	SI	1
Lekkas et al. (2021) (126)	Instagram	52	/	Yes	LIWC	Xgboost, Logitboost, KNN, NNET, AVNET	SI	3

Liu et al. (2022) (123)	Weibo	/	40,222	No	LIWC	Single feature classification Ensemble feature classification Multi-feature classification	SI	1
Ma et al. (2020) (124)	Weibo	5,000	500,000	No	Manual	SVM, Naive Bayes, CNN, LSTM, SDM, DAM	Suicide Risk	2
Mbarek et al. (2022) (107)	Twitter, YouTube, Tumblr	111	/	No	/	Random forest, Bayes Net, SVM, Decision Tree, Databoost	Suicide Risk	4
Meraliyev et al. (2021) (127)	Instagram, Vkontakte	50	/	No	/	Multinomial Naive Bayes, KNN, Gaussian Naive Bayes, SVM, Decision Tree	Suicide Risk	4
Metzler et al. (2022) (108)	Twitter	/	3,202	No	Manual	Majority Classifier, TF-IDF, SVM, BERT, XLNet	SI	1
O'Dea et al. (2015) (109)	Twitter	/	14,701	No	Manual	SVM, Logistic Regression	Suicide Risk	1
Ophir et al. (2020) (128)	Facebook	1,002	83,292	Yes	/	Single Task Model Multi Task Model	Suicide Risk	3
Pan et al. (2023) (125)	Weibo	Depression Group: 3,196 Controls: 5,167	Depression Group: 487,251 Controls: 357,939	No	SC-LIWC	Logistic Regression, Linear Regression	SI	2
Parraga et al. (2019) (110)	Facebook, Twitter, Instagram	/	102	No	Pre-analyzed	K-means, PAM, H. single/complete/average /Ward.D/Ward.D2	SI	1

Rabani et al. (2020) (111)	Twitter	/		18,756		No	Manual	Multinomial Naive Bayes, Naive Bayes, Decision Tree, Logistic Regression, SMO, Random Forest	SI	1
Rabani et al. (2020) (112)	Twitter	/		19,523		No	Manual	Logistic Regression, Multinomial Naive Bayes, SVM, Decision Tree	SI	1
Ramachandran et al. (2020) (113)	Twitter	/		4,443		No	Manual	SVM, Logistic Regression, Random Forest, Multinomial Naive Bayes	SI	1
Ramirez et al. (2020) (114)	Twitter	252		1,214,474		No	Manual	Multilayer Perceptron, Convolutional Neural Network, SVM	SI	2
Rezig et al. (2021) (115)	Twitter	/		193,720		No	Manual	SVM, Naive Bayes, Logistic Regression, Decision Tree	SI	1
Roy et al. (2020) (116)	Twitter	Suicidal Ideators: 283	Controls: 2,655	Suicidal Ideators: 512,526	Controls: 3,518,494	No	Neural Network	Random forest	SI + Suicide Risk	2
Schoene et al. (2022) (117)	Twitter	/		112,969		No	Manual + LIWC	Maximum Entropy Classifier, LSTM, ALBERT, GCN	SI	1
Yatapala et al. (2021) (118)	Twitter	/		SI: 4,062	Controls: 5,144	No	Pre-analyzed	ANN, SVM	SI	1

Notes: Ref.: reference number, SI: suicidal ideation, LSTM: long short-term memory, LIWC: linguistic inquiry and word count, KNN: K-nearest neighbors, NNET: Neural Network, AVNET: averaged random seed neural nets, SVM: support vector machine, SC-LIWC: simplified Chinese linguistic inquiry and word count, SCMBWC: simplified Chinese microblog word count, ALBERT: a lite bidirectional encoder representations from transformers, GCN: graph convolutional network, SMO: sequential minimal optimization, TF-IDF: term frequency- inverse document frequency, *BERT: bidirectional encoder representations from transformers, CNN: convolutional neural network, LDA: latent Dirichlet allocation, LSA: latent semantic analysis, NMF: non-negative matrix factorization, VADER: valence aware dictionary for sentiment reasoning, SGD: stochastic gradient descent, SVC: stochastic gradient descent, SDM: suicide detection model, DAM: dual attention based suicide risk detection model, PAM: partition around medoids, GRU: gated recurrent unit, ANN: artificial neural network*

Supplementary Table 2. Results of individual studies.

Study ID	Result						Group
Author (Year) (Ref.)	Best Performing Machine Learning Model	AUC	ACC	Sens	Spec	F1	
Baghdadi et al. (2022) (97)	BERT	0.9611	0.9606	/	/	0.9586	1
Bhattacharya (2021) (98)	GRU/LSTM	/	/	/	/	0.94	1
Braithwaite et al. (2016) (99)	Decision Tree	/	0.919	0.53	0.97	/	3
Chatterjee et al. (2021) (100)	Logistic Regression	/	0.87	/	/	0.81	2
Cheng et al. (2022) (119)	SVM	0.48	/	0.64	0.32	/	3
Coppersmith et al. (2018) (101)	Word Embeddings, Bidirectional LSTM layer, Self-Attention Layer, Linear layer with softmax output	0.94	/	/	/	/	4
Du et al. (2018) (102)	CNN	/	/	/	/	0.83	1
Fodeh et al. (2019) (103)	Decision Tree	/	/	0.912	0.829	/	2
Fu et al. (2021) (120)	Ensemble Model: Psychology+	/	/	/	/	0.7798	1
Guan et al. (2023) (121)	Simple Lineair Regression	/	/	/	/	0.35	3
Haque et al. (2022) (104)	Random Forest	0.92	0.93	/	/	0.92	1
Huang et al (2014) (122)	SVM	/	0.94	/	/	0.683	4
Jung et al. (2023) (105)	Gradient Boosting Machine	0.907	/	/	/	0.846	1
Kancharapu et al. (2022) (106)	LSTM	/	0.87	/	/	0.83	1
Lekkas et al. (2021) (126)	Consensus Ensemble Model	0.755	0.702	0.769	0.654	0.741	3

Liu et al. (2022) (123)	Ensemble Model	/	0.8061	/	/	0.792	1
Ma et al. (2020) (124)	DAM	/	0.9181	/	/	0.9154	2
Mbarek et al. (2022) (107)	Random Forest	/	/	/	/	0.854	4
Meraliyev et al. (2021) (127)	Multinomial Naive Bayes	/	0.83	/	/	/	4
Metzler et al. (2022) (108)	XLNet	/	/	/	/	0.55	1
O'Dea et al. (2015) (109)	SVM	/	0.76	/	/	/	1
Ophir et al. (2020) (128)	Multi Task Model	0.746	/	/	/	/	3
Pan et al. (2023) (125)	Logistic Regression	0.82	/	/	/	0.78	2
Parraga et al. (2019) (110)	H. Average	/	0.48	/	/	0.87	1
Rabani et al. (2020) (111)	Random Forest	0.9972	0.985	/	/	/	1
Rabani et al. (2020) (112)	Decision Tree	/	/	/	/	0.956	1
Ramachandran et al. (2020) (113)	Logistic Regression	/	0.763	/	/	0.17	1
Ramirez et al. (2020) (114)	SVM	0.95	0.86	/	/	0.86	2
Rezig et al. (2021) (115)	Logistic Regression	/	0.974	/	/	/	1
Roy et al. (2020) (116)	Random Forest	0.88	/	/	/	/	2
Schoene et al. (2022) (117)	Feature GCN	/	/	/	/	0.91	1
Yatapala et al. (2021) (118)	ANN	/	0.8676	/	/	/	1

Notes: ACC: accuracy, Sens: sensitivity, Spec: specificity, F1: F1-score