Suicidal Ideation Detection from Social Media Posts through Machine Learning Algorithms

Bachelor's Thesis

Maya Deniz Yilmaz

mayadeniz.yilmaz@uzh.ch

Methods of Plasticity Research
Department of Psychology
University of Zürich

Supervisor:
Prof. Dr. Nicolas Langer

01.06.2022

Abstract

Suicide is one of the biggest public health challenges worldwide, taking the lives of more then 700 000 people each year. The wide use of social media platforms present an opportunity to automatically detect suicidal ideation posts, which may help suicide prevention. In this thesis, I give a brief overview of the text document classification process for suicidal ideation detection (SID) from social media posts by presenting some feature extraction and selection methods and the most popular machine learning classifiers, including traditional and deep learning algorithms. A review of the literature on this topic showed that suicidal posts can be successfully detected with algorithms achieving F1 scores between 0.53-0.96. Therefore, machine learning algorithms are an efficient way to detect suicidal ideation posts on social media, which can be further implemented for suicide prevention.

Contents

Chapter 1

Introduction

1.1 Suicide as an Emerging Public Health Problem

Suicide is one of the leading causes of death worldwide and in Switzerland. According to the World Health Organisation, it is estimated that 703 000 people died by suicide in 2019 [1]. When looking at age groups, suicide was the fourth leading cause of death among young people aged between 15-29 [2], which raises significant concerns. In Switzerland, suicide is also the cause of death in 15 out of 1000 deaths. Between the ages of 15 to 45, accidents and suicide are the cause of majority of deaths [3]. These numbers reflect a major public health challenge, which makes it essential to understand the warning signs of suicide and to identify individuals who are at risk of taking their own lives.

Possible causes of suicidal ideation and behaviours are thought to be a complex interaction of psychological, biological, environmental, and cultural factors. Some of the current theories on the etiology of suicidal ideation are the Diathesis-Stress Model, The Interpersonal-Psychological Theory of Suicide (IPTS), Three Step Theory (3ST) and The Integrated Motivational-Volition Theory (IMV) [4]. The Diathesis-Stress model suggests that a combination of biological and/or psychological factors create a vulnerability for some individuals, who are then predisposed to suicidality when encountering particular life stressors [5]. IPTS predicts that the combination of "perceived burdensomeness" to others and "thwarted belonging", which is when the need to belong is unmet, can result in suicidal ideation [6]. IMV on the other hand suggests that the pathway to suicidal ideation is through defeat and entrapment [7]. 3ST proposes pain and hopelessness as factors that lead to suicidal ideation [8]. Some of the possible biological factors that may lead to suicidal ideation are dysregulated hypothalamic-pituitary-adrenal (HPA) axis function, neuroinflammation and immune system dysfunction [4]. In terms of genetic predisposition, variants of the FKBP5 gene have been implicated in depression and suicidal ideation and behaviors [9].

1.2 Suicidal Ideation Detection and the Importance of Social Media Communication

Suicidal ideation is defined by the ICD-11 as "thoughts, ideas, or ruminations about the possibility of ending one's life, ranging from thinking that one would be better off dead to formulation of elaborate plans" [10] The transition from suicidal ideation to the first suicide attempt usually occurs within a year after the ideation onset [11]. For this reason, suicidal ideation detection (SID) plays a key role in preventing suicide. Figure 1.1 presents an overview of the methods and domains of SID.

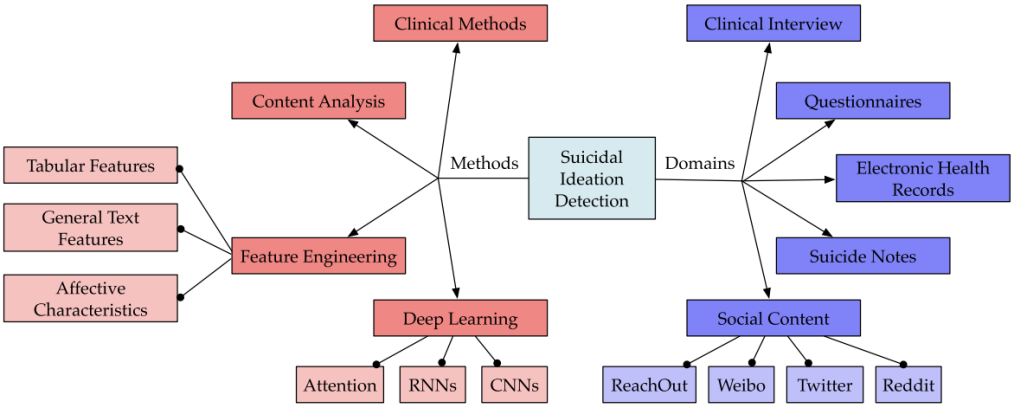

Figure 1.1. Methods and Domains of Suicidal Ideation Detection. From: Ji et al., 2021, p.215

Traditional forms of SID make use of clinical methods such as self-reports and face-to-face interviews, which are not applicable as screening instruments due to their high cost, whereas machine learning techniques are generally applied for automatic detection [12]. Social media's increasing popularity among young people, who are usually harder to engage in traditional forms of treatment, make it a promising tool to monitor suicidal thoughts [13].

1.3 What is Machine Learning?

Machine learning is "a branch of artificial intelligence that systematically applies algorithms to synthesize the underlying relationships among data" [14]. It was defined by Arthur Samuel as the "field of study that gives computers the ability to learn without being explicitly programmed" [15].

1.3.1 Unsupervised and Supervised Learning

Machine learning can be broadly divided into two categories by whether the algorithm's learning is "supervised" or "unsupervised" [16]. Unsupervised learning is when the algorithm tries to identify natural groupings within the data without any reference to a specific outcome or a "right answer" [17]. Supervised learning on the other hand makes use of prior training examples to map each input to an output, where the training data is already "labeled". This means that the value of the outcome is already specified for each observation [16]. If the research goal is to classify data or make predictions, it is better to opt for supervised learning, whereas unsupervised learning methods are suitable for understanding relationships within a data set [18].

1.4 Aim of the Thesis

The aim of this thesis is to provide the reader a brief introduction to machine learning to understand the literature around SID with machine learning algorithms from social media posts and a comparison of different feature and classifier combinations in terms of accuracy. While doing that, I will be focusing more on binary classifications of suicidal ideation. I will also be discussing the ethical implications of social media research and how these findings can be applied for future suicide interventions.

Chapter 2

Theoretical Background

The text document classification process involves firstly pre-processing techniques that cleans and organizes the raw data to prepare it for the machine learning model [19]. Feature extraction and feature selection methods are used to reduce dimensionality (i.e. number of covariates), which is when an increasing number of features (i.e. independent variables) "relative to cases result in lower accuracy and generalizability" [20].

2.1 Feature Extraction & Selection Methods

Feature extraction is the first step following the pre-processing where the data is transformed into a new feature space to reduce the amount of data to be processed, while still accurately and fully representing the original data set [19]. Feature selection obtains a subset of features according to particular criteria, where the original features are maintained [21].

2.1.1 Word Embedding Techniques

In natural language processing (NLP), word embedding techniques are unsupervised learning techniques that convert words in a document into a vector space representation by analyzing their co-occurrences [22]. Common word embedding techniques are bag of words (BoW), term frequency–inverse document frequency (tf-idf), word2vec and GloVe.

BoW is the simplest way of representing words as numerical values. It is a feature extraction method that calculates the term-frequency of words in a document, which is the amount of time a word occurs relative to the number of all words in that document. Terms like "the" occur frequently in all types of documents, which reduces the importance of other words, though in NLP these prepositions are generally not accounted for [23].

The tf-idf is a popular term-weighting method that calculates how rare a word

is across all documents. It is commonly used to reduce the importance of regularly used words, while also giving higher weight to words that occur infrequently [23].

Word embeddings like word2vec and GloVe are able to assess the syntactic and semantic relationships of words by using their vector space representations [24]. The main difference between them is that word2vec only considers the local context of words [25] while Glove makes use of the global co-occurance counts [26].

2.1.2 Linguistic Inquiry and Word Count (LIWC)

LIWC is a text analysis program that counts words in categories and gives them a score for each. These categories are "Word count, 4 summary language variables (analytical thinking, clout, authenticity, and emotional tone), 3 general descriptor categories (words per sentence, percent of target words captured by the dictionary, and percent of words in the text that are longer than six letters), 21 standard linguistic dimensions (e.g., percentage of words in the text that are pronouns, articles, auxiliary verbs, etc.), 41 word categories tapping psychological constructs (e.g., affect, cognition, biological processes, drives), 6 personal concern categories (e.g., work, home, leisure activities), 5 informal language markers (assents, fillers, swear words, netspeak), and 12 punctuation categories (periods, commas, etc)" [27]. In the context of SID, it can be used as a tool to find correlations between choices of words and suicidal ideation [23].

2.1.3 Sentiment Analysis (SA) Approach

SA is "the process of determining the emotional tone behind a series of words using NLP techniques" [23]. It produces a polarity score, which reflects whether the document is negative (-1) or positive (1), and a subjectivity score that ranges from factual (0) to totally subjective (1). Selected values are then combined and used for prediction modeling [23].

2.1.4 Principal Component Analysis (PCA)

Principal Component Analysis is a statistical procedure that reduces the dimensionality of large data sets by creating new uncorrelated variables that maximize variance. This technique increases interpretability while also minimizing information loss [28]. Figure 2.1 is a visualised example of extracted features in a two-dimensional space using PCA, where the points represent the data and their colour represents the class that they belong to. Here, the extracted features successfully separate the data in suicidal and non-suicidal classes.

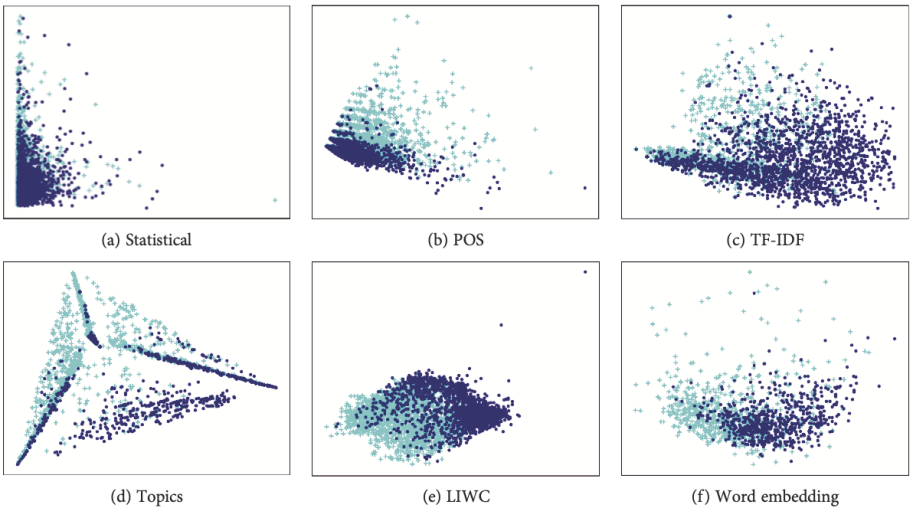

Figure 2.1. Visualisation of Extracted Features Using PCA. From: Ji et al., 2018, p.6

2.1.5 Topic Modeling

Topic modeling is a methodology used for scanning documents for word and phrase patterns, which are then clustered into an underlying set of topics (see Figure 2.2). This method can be used to "classify, index and summarize the content of documents" [29]. Latent Dirichlet Allocation (LDA) is a popular example of a topic model [30]. In SID, this can be used to identify topics related to suicide such as depression, anxiety, stress and hopelessness [31].

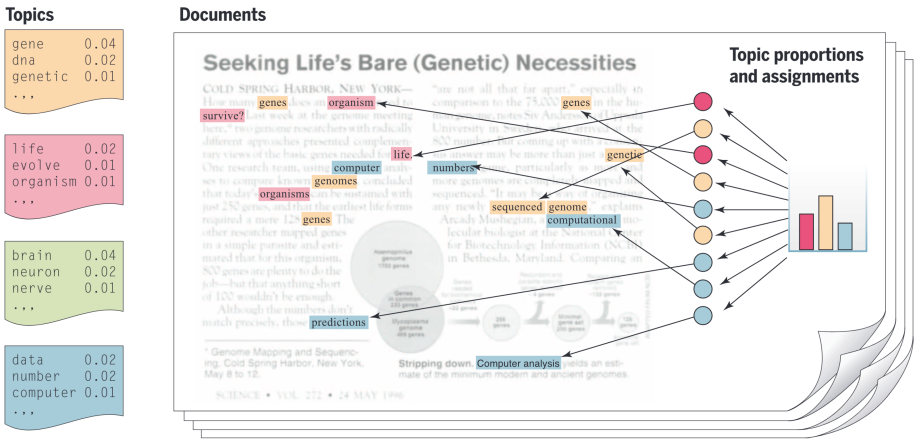

Figure 2.2. Topic Modeling. From: Jordan & Mitchell, 2015, p. 258

2.2 Classification Algorithms

2.2.1 ZeroR

ZeroR is the simplest classifier. It simply predicts the majority class and is used as a benchmark for other classification methods [32].

2.2.2 Logistic Regression

Logistic regression is one of the most widely used supervised learning classification algorithms that predicts the probability of a binary outcome [33].

2.2.3 Decision Tree

The decision tree constructs a series of decision rules in the form of a tree structure [17] by applying "a splitting rule on successively smaller partitions of data, with each partition being a node on the tree" [16]. Here, nodes represent "attributes in a group that is to be classified and each branch represents a value that the node can take" [18]. Figure 2.3 is an example for a classification decision tree that predicts type 2 diabetes mellitus [16]. Decision trees provide advantages for a range of target audiences, as they are simple to understand and interpret [19].

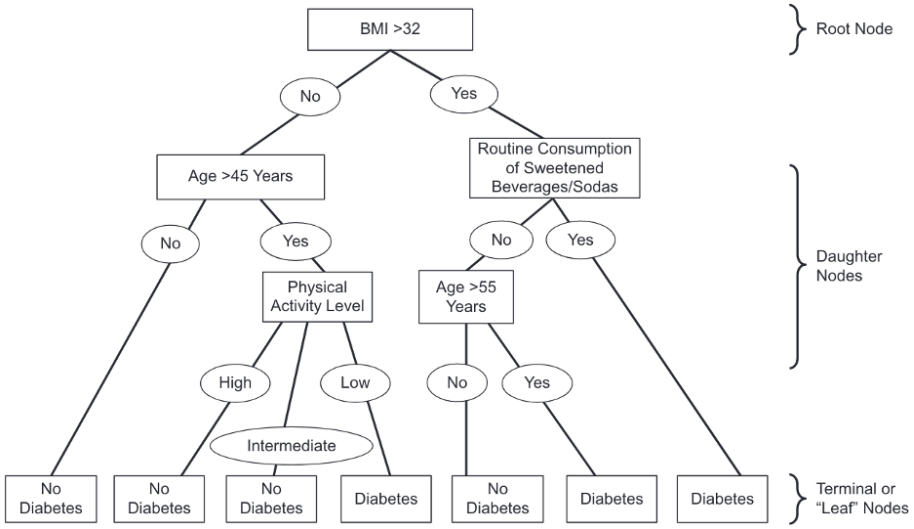

Figure 2.3. Classification Decision Tree for type 2 diabetes mellitus. From: Bi et al., 2019, 2226

2.2.4 Random Forest

Random Forest is an ensemble classifier that makes use of a combination of decision trees, where the output is the class selected by the most trees [34].

2.2.5 Naïve Bayes

Naïve Bayes is a "probabilistic classifier based on applying Bayes' Theorem" [19] that assumes the independence of predictive variables, which means that "the presence of a particular feature in a class is unrelated to the presence of any other feature" [18]. It is a popular method because of its simplicity but it requires a smaller training data set because of its independence assumption [35].

2.2.6 Support Vector Machine (SVM)

Support Vector Machines are supervised learning algorithms that construct an optimal boundary called a hyperplane, which maximizes the margin between two classes (see Figure 2.4) and thus minimizing the classification error [18].

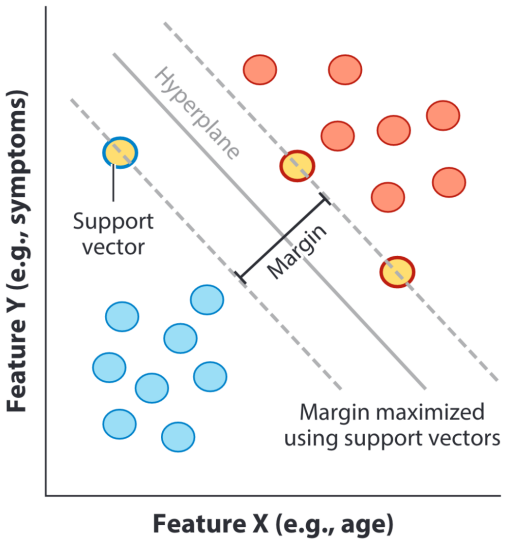

Figure 2.4. Support Vector Machine. From: Dwyer et al., 2018, p.101

2.2.7 Deep Neural Networks

Deep learning is a class of machine learning techniques, which makes use of multilayered "neural networks". These artificial neural networks function similar to the human brain, where interconnected "neurons" work together to solve a

specific problem [36]. The internal layers of deep networks provide learned representations of the input data and adjust parameters according to the errors at its output [29].

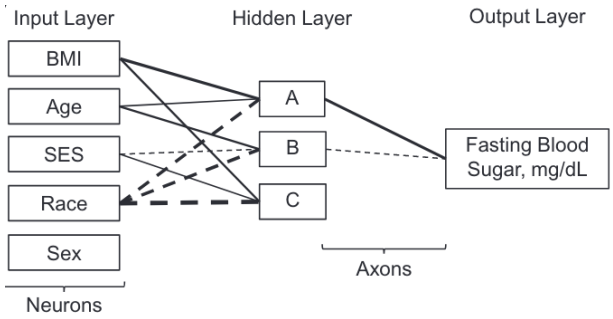

Figure 2.5. Neural Network Example. From: Bi et al., 2019, p.2225

Figure 2.5 is an example of a neural network with one hidden layer "that examines the relationship between clinical and demographic predictors and a numerical outcome, fasting blood sugar level" [16]. Some examples to most used deep neural networks are convolutional neural networks (CNN) and recurrent neural networks (RNN) [12].

2.3 Classifier Performance Metrics

Commonly used metrics for the evaluation of classifier performance are accuracy, precision, recall (i.e. sensitivity), F1 score, receiver operating characteristics (ROC) and area under the ROC curve (AUC) [23] [37]. Table 2.1 is a two-by-two confusion matrix in the context of binary suicidal post classification, which visualises the performance of the algorithm and depicts the four possible outcomes: number of true positives (TP), number of true negatives (TN), number of false positives (FP), and number of false negatives (FN). Most performance metrics are derived from these four values [37].

Table 2.1

Confusion Matrix: Binary Classification of Suicidal Ideation.

		Predicted Class	
		Suicidal Ideation	Not Suicidal Ideation
Actual Class	Suicidal Ideation	True Positive (TP)	False Negative (FN)
	Not Suicidal Ideation	False Positive (FP)	True Negative (TN)

Here, accuracy is the ratio of the correctly classified posts to all posts [37]:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{2.1}$$

Recall calculates the fraction of suicidal ideation posts that are correctly classified as suicidal ideation [23]:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.2}$$

Precision is the fraction of actual suicidal ideation posts among posts that are classified as suicidal ideation [23]:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2.3}$$

The F1 score (i.e. F-measure) is a metric based on Precision and Recall, as shown in Equation 2.4. It ranges between 0 and 1, where a perfect classifier yields an F1 score of 1 [37].

$$F_1 = \frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{\text{TP}}{\text{TP} + \frac{1}{2}(\text{FP} + \text{FN})} \tag{2.4}$$

The ROC plots the TP rate against the FP rate and visualises the trade-off between them [38]. Figure 2.6 is an example of an ROC curve, where the AUC is also calculated.

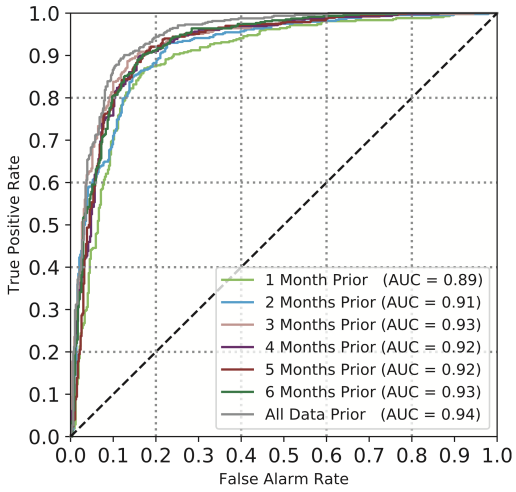

Figure 2.6. ROC Curve. From: Coppersmith et al., 2018, p.5

The AUC is a single value measurement that ranges from 0 to 1 [38]. A classifier that yields a large AUC is usually preferable over one that yields a smaller AUC [37]. AUC is also sensitive to the distribution of the data, because it assumes that suicidal and non-suicidal posts are distributed equally in the data set. This is generally not the case in SID, so the threshold should be adjusted to the prevalence of suicidal posts in the training data [39].

In this thesis, I will mostly be focusing on the F1 score to compare the performance of the classifiers, because it gives a balanced evaluation as the harmonic mean of Precision and Recall [23].

Chapter 3

Methods

3.1 Literature Selection

To select the literature that will be included in this review, I first cross-referenced the papers provided by Prof. Langer on the topic of SID from social media posts. Thus, I identified the key literature on this topic. Further, I decided to focus on the binary classification of suicidal ideation and compare the F1 scores achieved by different machine learning models. For this I did a further literature search on Google Scholar and Web of Science and identified more papers reporting F1 scores of classifiers for SID on social media. I searched for the key words "suicide", "suicidal ideation", "social media", "machine learning", "text classification" and "natural language processing". Here, I also gave importance to the inclusion of non-English data sets to increase generalizability, especially the Chinese microblogging platform Sina Weibo.

Chapter 4

Results

To present the results, I decided to categorize the literature to Reddit 4.1 and microblogs 4.2 and report the performance of the classifiers of each section in a joint table. The reasoning behind this categorization was the post length: Posts on Reddit are much longer then posts on microblogs, which are usually under 300 words [40]. Because of this, feature combinations that yield the best results may differ. Also, the performance of the classifiers are not comparable between short and long posts, as longer posts contain more information and are thus easier to classify correctly [23].

While describing the characteristics of the control groups in the studies, I make use of vocabulary defined by [41]. From here on, "generic" control group refers to a data set of randomly selected non-suicidal posts and "focused" control group refers to a data set of non-suicidal posts making use of suicide-related phrases, which make them harder to classify.

Studies that will be presented also differ in terms of classification. Most studies make a binary classification, where they classify posts as either "suicidal" or "non-suicidal", whereas some of them define multiple classes and label the posts accordingly.

4.1 Reddit

Reddit is a popular online forum which was founded in 2005 [23]. Posts on Reddit are organized in topic categories which are called "subreddits" [42]. One of the subreddits called "SuicideWatch" is a platform where users talk about their suicidal thoughts and share their feelings [23]. It is intensively used for mental health research [42] and to collect suicidal ideation posts [12] for machine learning based SID.

4.1.1 Detecting Suicidal Ideation on Forums: Proof-of-Concept Study (Aladag et al., 2018)

In their study, Aladag et al. [23] selected 785 random posts from Reddit, specifically from the SuicideWatch, Depression, Anxiety and ShowerThoughts subreddits, which were then manually annotated as suicidal if the author of the post clearly appeared to have suicidal ideations or non-suicidal. Posts from the SuicideWatch subreddit were not annotated and were directly classified as suicidal. Four experiments were conducted, where a custom data set out of these posts were generated for each.

In the first experiment, the authors compiled a data set that consisted of 175 posts from SuicideWatch and 210 posts from ShowerThoughts. It can be said that the posts from ShowerThoughts acted as a generic control group. The F1 scores of the classifiers are reported in Table 4.1.

The second experiment included 200 posts from the Anxiety and Depression subreddits each, on top of the data set from the first experiment. It can be said that the addition of these posts resulted in a focused control group. The resulting F1 scores are reported in Table 4.1.

The aim of the third experiment was to check if all of the posts in the SuicideWatch subreddit can be classified as suicidal and all of the posts in the ShowerThoughts subreddit can be classified as non-suicidal. When annotated by psychiatrists, 85.7% of posts on SuicideWatch were identified as having been posted by people with suicidal ideation. For this, they created a training set containing 5000 non-annotated posts from each of the subreddits and tested the models against previously annotated 175 SuicideWatch and 210 ShowerThought posts. In this experiment, the best F1 score was reached by the SVM algorithm with 0.92. This performance indicates that the assumption of all posts in SuicideWatch belonging to the suicidal class and all posts in ShowerThoughts in non-suicidal class is a valid one.

The final experiment aimed to test the trained model from experiment three on the full data set of 785 posts. The classifier yielding best results was again SVM with an F1 score of 0.79. The lower performance compared to the third experiment was attributed to the lack of "edge cases" (posts from Depression and Anxiety subreddits) in the training data set.

In conclusion, this study showed that classifiers can achieve high performance in SID from long format social media posts and that it is safe to assume that all posts in the SuicideWatch subreddit are suicidal and that all posts in the ShowerThoughts subreddit are non-suicidal.

4.1.2 Discovering Shifts to Suicidal Ideation from Mental Health Content in Social Media (De Choudry et al., 2016)

For their study, De Choudry et al. [42] collected data from 14 mental health subreddits and SuicideWatch and separated the data into two time periods: t1 being 11.02.2014 to 11.08.2014 and t2 being 12.08.2014 to 11.11.2014. They then categorized users that posted on these subreddits into two sets: The first set consisted of users that posted on the mental health subreddits during t1 but never posted on SuicideWatch. The second set of users posted on mental health subreddits in t1 but switched to posting on SuicideWatch during t2. This resulted in 440 users in the second set and they also randomly sampled another 440 users from the first set. This resulted in a data set of 13049 posts and 101035 comments.

The authors observed that the second set of users use a greater number of first person singular pronouns and lower numbers of second person pronouns, first person plural pronouns and third person pronouns when compared to the first set of users. They also use more verbs and adverbs but less nouns, have a lower readability index and adjust their language less to the general style on the mental health subreddits. They have longer but fewer posts and receive fewer comments on them.

De Choudry et al. also found that the use of phrases like “depression”, “useless”, “suicide” , “anxiety”, “no friends”, “have nothing”, “kills” and “to cry” significantly increases a users chances of posting on SuicideWatch in the future. The use of phrases like “counseling”, “relationship that”, “intimate”, “hope it”, “i agree” and “and enjoy” significantly lowered the probability of posting on SuicideWatch in the future.

The authors then extracted thematic clusters from the posts in an unsupervised way and examined the most dominant themes. They found that themes of hopelessness, anxiety, impulsiveness, self-esteem, loneliness and severe or stigmatized illness were associated with the heightened probability of posting on SuicideWatch in the future.

Finally, the authors used a logistic regression model to differentiate between the first and second set of users defined above. The classifier achieved an F1 score of 0.80.

All in all, the authors found many indicators, which characterize the transition from discussing mental health topics not related to suicide to talking about suicide. They also developed a machine classification algorithm that can predict this shift.

4.1.3 Supervised Learning for Suicidal Ideation Detection in Online User Content (Ji et al., 2018)

For their study, Ji et al. [43] collected data from both Reddit and Twitter. From Reddit, the authors created two data sets, the first one consisting of the subreddit SuicideWatch and popular posts in Reddit. The second data set contained the subreddits SuicideWatch, gaming, jokes, books, movies and AskReddit. Overall it included 3549 suicidal ideation posts and a 3652 non-suicidal posts.

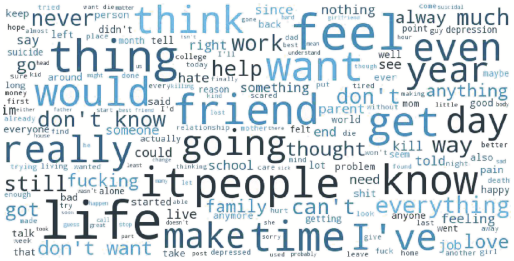

Figure 4.1. Reddit Word Cloud. From: Ji et al., 2018, p.4

The authors made use of word clouds to visualise the frequently used words in suicidal ideation posts (See Figure 4.1). They observed that the most used words differ in Reddit and Twitter, as people on Reddit describe their life events and stories about their friend circle, while Twitter users tend to be more straightforward, making use of expressions such as "want kill", "going kill", and "wanna kill". While analyzing the language of the posts, the authors found that users with suicidal ideation tend to use more personal pronouns, communicate more negative emotions, use more words related to death, use the present tense to talk about their suffering, pain and depression and use future tense to express their hopelessness and suicide intentions.

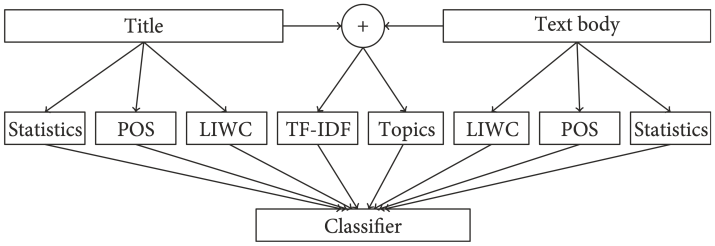

Figure 4.2. Model Structure for Reddit Data Set. From: Ji et al., 2018, p.4

Figure 4.2 depicts the machine learning model for the Reddit data set. Here the authors extracted features using statistical features like number of words, characters, sentences and paragraphs in the title and text body, POS, LIWC,

tf-idf and LDA for topic features. POS and LIWC featuers were applied to both title and text body of the posts. LDA and tf-idf was applied to the combination of title and text body as one piece of text.

The F1 scores for the first data set are presented in 4.1. In the second data set, posts on SuicideWatch were used in combination with one of the five other subreddits to evaluate the classifiers on specific online communities. The best performing classifiers in terms of accuracy for each topic were LSTM and XGBoost for gaming, XGBoost for jokes, Random Forest for books, XGBoost for movies and Random Forest and XGBoost for AskReddit.

The data collection process of Twitter and the results of the Twitter data set will be presented in the subsection 4.2.5.

4.1.4 Detection of Suicide Ideation in Social Media Forums Using Deep Learning (Tadesse et al., 2019)

For their experiment, Tadesse et al. [44] used the previously built Reddit data set by [43] introduced in the subsection 4.1.3. The authors designed two frameworks for feature extraction: The first one made use of tf-idf, BoW and statistical features in combination with traditional machine learning algorithms. The second one combined word2vec with deep learning classifiers.

When examining the data set, the authors found expressions of hopelessness and frustration, a sense of urgency, anxiety, sense of guilt, regret and signs of loneliness in posts from the subreddit SuicideWatch. Users posting on this subreddit also make more references to themselves, use more question marks, use more negations, tend to be preoccupied with their feelings, and use more words related to death. When looking at non-suicidal posts, they observed that users use more words describing happy moments, positive attitude and feelings, mention social relations activities and strive towards maintaining positive spirits.

The F1 scores of both frameworks are presented in Table 4.1.

In conclusion, the authors provide a comparison of SID through traditional machine learning algorithms and deep learning methods from Reddit.

4.1.5 Binary Suicidal Ideation Classification Results for the Reddit Data Set

The F1 scores of the machine learning models tested by [23], [43] and [44] are presented in Table 4.1. Out of these studies, the best F1 score of 0.938 was achieved by Ji et al. with a model that integrated statistical features, LDA, tf-idf, POS and LIWC in combination with Random Forest classifier.

Table 4.1

F1 scores of the Reddit data set for suicidal ideation classification.

Study	Data	Sample size	Model	Classifier	F1
Aladag et al., 2018	generic	385	LIWC, SA	ZeroR	0.66
			LIWC, SA	SVM	0.92
			LIWC, SA	LR	0.92
			LIWC, SA	Random Forest	0.89
	focused	785	LIWC, SA	ZeroR	0.66
			LIWC, SA	SVM	0.73
			LIWC, SA	LR	0.81
			LIWC, SA	Random Forest	0.80
Ji et al., 2018	generic	7201	Statistics	SVM	0.8116
			Statistics+LDA	SVM	0.8603
			Statistics+LDA+tf-idf	SVM	0.8634
			Statistics+LDA+tf-idf+POS	SVM	0.8727
			Statistics+LDA+tf-idf+POS+LIWC	SVM	0.9123
			Statistics	Random Forest	0.7653
			Statistics+LDA	Random Forest	0.9001
			Statistics+LDA+tf-idf	Random Forest	0.8954
			Statistics+LDA+tf-idf+POS	Random Forest	0.9031
			Statistics+LDA+tf-idf+POS+LIWC	Random Forest	0.938
			Statistics	GBDT	0.7513
			Statistics+LDA	GBDT	0.9017
			Statistics+LDA+tf-idf	GBDT	0.899
			Statistics+LDA+tf-idf+POS	GBDT	0.8955
			Statistics+LDA+tf-idf+POS+LIWC	GBDT	0.9478
			Statistics	XGBoost	0.7664
			Statistics+LDA	XGBoost	0.9028
			Statistics+LDA+tf-idf	XGBoost	0.9049
			Statistics+LDA+tf-idf+POS	XGBoost	0.9133
			Statistics+LDA+tf-idf+POS+LIWC	XGBoost	0.9133
			Statistics	MLFFNN	0.7742
			Statistics+LDA	MLFFNN	0.8631
			Statistics+LDA+tf-idf	MLFFNN	0.8385
			Statistics+LDA+tf-idf+POS	MLFFNN	0.9133
			Statistics+LDA+tf-idf+POS+LIWC	MLFFNN	0.9295
			word2vec	LSTM	0.9239
Tadesse et al., 2019	generic	7201	Statistics	Random Forest	0.751
			tf-idf	Random Forest	0.809
			BoW	Random Forest	0.786
			Statistics + tf-idf + BoW	Random Forest	0.841
			Statistics	SVM	0.79
			tf-idf	SVM	0.827
			BoW	SVM	0.811
			Statistics + tf-idf + BoW	SVM	0.838
			Statistics	Naïve Bayes	0.713
			tf-idf	Naïve Bayes	0.761
			BoW	Naïve Bayes	0.784
			Statistics + tf-idf + BoW	Naïve Bayes	0.815
			Statistics	XGBoost	0.761
			tf-idf	XGBoost	0.841
			BoW	XGBoost	0.826
			Statistics + tf-idf + BoW	XGBoost	0.831
			word2vec	LSTM	0.926
			word2vec	CNN	0.928
			word2vec	LSTM-CNN	0.934

4.2 Microblogs

Microblogs are short textual posts written continuously by an individual that are usually personal [45]. The readers of these posts are called "followers" [45]. Microblogging platforms like Twitter and Sina Weibo allow people to share short textual messages on the internet [46]. Twitter users share more then 200 million posts per day, which are called "tweets" [47]. Founded in 2009, Sina Weibo is the most popular microblogging service in China, since Twitter is unavailable [46] [48].

On Sina Weibo, people share their inner thoughts and suicidal ideations on microblogs called "tree holes" whose authors have committed suicide [49]. One of these tree holes with over a million posts is under the suicide note of a microblogger called Zoufan, who committed suicide on 17.03.2012 [50]. This microblog group is also used for the collection of suicidal ideation posts [49] [50].

4.2.1 Detection of Suicidal Ideation on Social Media: Multimodal, Relational, and Behavioral Analysis (Ramírez - Cifuentes et al., 2020)

In their study , Ramírez-Cifuentes et al. [41] firstly collected suicide-related sentences from Reddit's Suicide Watch forum, which were then translated to Spanish and reviewed by clinical psychologists. To generate a reliable Twitter data set, they then selected a random sample of 1200 users, who had at least 2 tweets containing the search phrases that were collected from Reddit. A clinician was then asked to classify these users into 3 categories: "control" (users who do not seem to manifest suicidal ideations), "suicidal ideation risk" (users with suicidal ideation signs) and "doubtful" (cases that the psychologists were unsure about). A second labeling process was then conducted for the users in the suicidal ideation risk category, where annotators were provided with the summarized version of a users profile (SPV), which mainly contained tweets related to suicide and its risk factors. In order to create the SPV, they developed a short profile version classifier (SPVC) that applied a BoW model, PCA and logistic regression analysis and achieved an F1 measure of 0.90. The SPVC was used to identify the top 15 suicide related tweets with the highest predicted values by the SPVC of each user. These set of tweets were then evaluated by two additional annotators as either "suicidal ideation risk" or "control". The authors defined 2 different control groups of the same size: focused and generic control group (see above for description).

When comparing the suicidal ideation risk and focused control groups, they found significant differences in number of friends, median tweet length, overall ratio, median time between tweets, verbs, verbs conjugated in singular of the first person ("I"+verb), cognitive mechanisms, anxiety-related terms, usage of personal

pronouns, usage of the pronoun “I,” negations, terms to express feelings, coursing terms, the usage of suicide explicit terms, depression-related terms, self-loathing, substance abuse, self-injuries, and terms expressing lack of social support. There were also significant differences between the groups for n-grams such as I feel, sad, kill myself, cry/crying, depression, to die, horrible, anxiety, die, pills, among others.

When comparing the suicidal ideation risk and generic control groups, they found significant differences in the median classifier score, the number of tweets generated, and the median time between tweets. The suicidal ideation group tended to use terms more related to health and biological aspects, while the generic control group tended to have more discussions around topics such as money and work. The use of self-references was also higher in the suicidal ideation risk group. Both the generic and focused control group tend to tweet more on weekdays and during the day, whereas the suicidal ideation risk group tended to tweet more on the weekends and at night.

The authors also developed a classifier trained on images that were extracted from Instagram which made use of a subset of the phrases and keywords that had been identified from Reddit. To define the image classifier,they used CNN and have found significant differences between the suicidal ideation risk and generic control group.

When comparing the control groups, they found significant differences in suicide-related lexicons, such as suicide methods, suicide explicit terms, bullying, discrimination, and substance abuse–related terms. They also found significant differences in the number of tweets, number of friends, number of followers, median favorites and retweet counts, overall ratio, polarity score, median time between tweets, among others.

The classification task results are presented in Table 4.3.

Overall, the authors have made use of information from multiple platforms (Twitter, Instagram, Reddit) to build a model for SID, defined a SPV to minimize the noise of writings unrelated to suicide and have shown that the classifiers yield better results when using generic control groups instead of focused control groups.

4.2.2 Multi-class machine classification of suicide-related communication on Twitter (Burnap et al., 2017)

Firstly, Burnap et al. [51] collected posts from Tumblr, www.recoveryourlife.com, www.enotalone.com, www.experienceproject.com and www.takethislife.com, which either have dedicated sections or are designed for discussing suicide. The resulting 1800 posts were then annotated using the crowdsourcing online service Crowdflower, where the annotators answered the question "Is this person suicidal?" for binary classification. The authors used the tf-idf method to identify

terms that appear frequently in tweets that contain suicidal ideations. They then collected data from Twitter for 6 months using the identified search terms and at the same time identified names of reported suicide cases in England retrieved another data set from Twitter using their name and surname as search terms. Combining both data sets, they produced a random sample of 1000 tweets that resulted in a data set of 816 tweets after they were annotated.

The authors then defined 5 feature sets: The first one consisted of POS, structural features such as total number of sentences with negations, general and affective lexical domains, sentiment score for positivity and negativity and 62 keywords that were derived from the Web. The second feature set made use of the LIWC text analysis software. The authors then developed a set of regular expression (RegEx) and pattern matching rules from the posts that were collected from Tumblr for the third feature set. For the fourth feature set by incorporating all of the features in the first three feature sets. Finally, they applied PCA to the fourth set to achieve a new set of linearly uncorrelated features.

A baseline experiment was conducted to test the machine classifiers and according to the authors, the results were "reasonable but required refining". To improve the results, they then used Rotation Forest (RF) ensemble classifier, which combined Naïve Bayes and SVM classifiers. The machine classification results for the suicidal ideation class are presented in Table 4.2.

Table 4.2

F1 scores for the suicidal ideation class.

	Classifier			
Feature Sets	Naïve Bayes	Decision Tree	SVM	Random Forest
Set 1	0.603	0.435	0.607	0.525
Set 2	0.579	0.384	0.612	0.493
Set 3	0.588	0.486	0.603	0.519
Combined Set	0.586	0.466	0.614	0.690
Combined Set (PCA)	0.477	0.482	0.411	0.516

After these experiments, the authors made use of the best performing classifiers for a 12 month case study on Twitter, where they had a binary classification task and a 7-class task, which included suicidal ideation and other suicide-related topics such as reporting of a suicide, memorial, campaigning and support. For the 7-class task, out of the systematic sample of 2000 tweets, 3 out of 4 annotators only agreed on 1121. The classifier achieved 65.29% accuracy with this sample. For the binary task another systematic sample of 2000 tweets were annotated and the resulting data set consisted of 1731 tweets. The classifier achieved 85% accuracy with this sample.

The classifier was then applied to a data set of 1,884,248 tweets from this 12 month period, which contained the 62 keywords and seemed to be in the London

time zone. The most common category was "flippant reference to suicide", which means that 48% of these tweets had many of the linguistic features of suicidal ideation but were written in relation to trivial matters. It was also observed, that suicidal ideation tweets peaked around the time of widely publicized celebrity suicides.

Lastly, the authors derived the age and gender of the users that posted suicidal ideation tweets and found that when compared to the baseline of Twitter users, there was an under-representation of men using suicidal ideation language, over-representation of users with unisex names and an over-representation of users aged between 13-20. More precisely, the age group of 13-20 made out 81.2% of the users that posted suicidal ideation tweets. They also found a significant correlation (r=0.11) between the rate of suicidal tweets and female daily suicide rate.

To conclude, Burnap et al. have developed a machine classifier that successfully distinguishes between suicidal ideation and other kinds of communication related to suicide.

4.2.3 Suicide Related Text Classification with Prism Algorithm (Chiroma et al., 2018)

For their experiment, Chiroma et al. [52] made use of a previously defined data set by Burnap et al. (Burnap et al., 2015), which consisted of 1060 tweets that were categorized into 7 classes. This set of tweets were organized into 3 data sets: A binary, a three-class and a seven-class data set. The aim was to compare the Prism algorithm, a classification rule-learning algorithm that was developed by Cendrowaski in 1987, with other popular machine learning algorithms.

The binary data set contained 289 tweets and the performance of the machine learning algorithms are presented in 4.3. The three-class data set contained "suicide", "flippant" and "non-suicide" classes and Prism, Random Forest and SVM achieved an F-measure of 0.65 for the "suicide" class. The seven-class data set consisted of "suicide", "campaign", "flippant", "support", "memorial", "reports" and "other" and the the best F-measure for the "suicide" class was achieved by Prism with 0.65.

To conclude, the results indicate that the Prism algorithm is also applicable for SID.

4.2.4 Extracting psychiatric stressors for suicide from social media using deep learning (Du et al., 2018)

To create a data set from Twitter, Du et al. [53] first selected tweets that included 21 keywords/phrases such as “suicide”, “kill myself”, “want death”, etc. These

tweets were filtered by stop words such as “hotline”, “suicide bomb”, “suicide attack”. The tweets were then annotated "positive" if they were related to suicidal ideation and "negative" if they were discussion of suicide of other people, news or reports, not related to suicidal ideation, negation of suicide ideation, or other non-positive tweets. The tweets labeled "positive" were further annotated for psychiatric stressors.

The authors then trained a CNN based binary classifier, which was used to further select 3000 tweets with "positive" labels to be annotated for psychiatric stressors. The CNN model was compared to other widely used machine learning algorithms, where the models included Glove Twitter embedding for future extraction. The F1 measures of these classifiers are presented in 4.3.

To extract the mention of stressors in the "positive" labeled tweets, the authors leveraged an RNN based framework. The best F1 score of 0.5325 was achieved with the GloVe Twitter embedding.

Du et al. also made use of a data set of psychiatric notes, which were annotated for stressors. These 946 sentences were used for transfer learning to be used on Twitter. Also initialized with GloVe Twitter embedding, the best model trained achieved an F1 score of 0.5014. When the best model trained on the clinical notes was transferred to initialize the RNN model, an F1 score of 0.549 was achieved.

All in all, the authors made us of deep neural networks for the binary classification of suicidal ideation on Twitter and also for the extraction of psychiatric stressors from suicide related tweets.

4.2.5 Supervised Learning for Suicidal Ideation Detection in Online User Content (Ji et al., 2018)

Here, the data collection process of the Twitter sample and the experiments on this data set will be presented. The authors collected tweets from Twitter using search phrases like "suicide”, “die” and “end my life”. These tweets were then labeled as "suicide text" if they expressed suicidal thoughts or included potential suicidal actions. Tweets that were formally discussing suicide, referring to other’s suicide or were not relevant to suicide were labeled as "non-suicide text". This resulted in a data set of 10288 tweets.

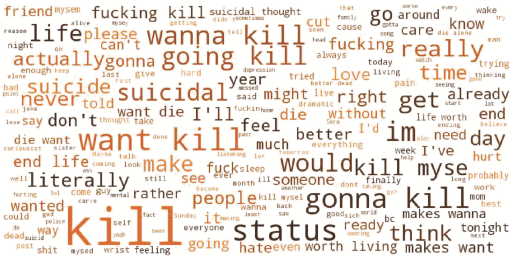

Figure 4.3. Twitter Word Cloud. From: Ji et al., 2016, p.4

Figure 4.3 is a word cloud that visualises the frequently used words in the suicidal posts. The model for the Twitter data set excluded the number of paragraphs in statistical features, POS, and LIWC features of the text body. The F1 scores achieved by the classifiers are presented in Table 4.3.

Overall, the authors compare the performance of six different feature sets in combination with six supervised learning algorithms in terms of SID on posts from Reddit and Twitter.

4.2.6 Validating Machine Learning Algorithms for Twitter Data Against Established Measures of Suicidality (Braithwaite et al., 2016)

For their experiment, Braithwaite et al. [54] first selected 135 participants (85 female, 50 male) from Amazon's Mechanical Turk (MTurk, www.mturk.com), who provided their Twitter handle and had their psychological functioning assessed through the Depressive Symptom Inventory–Suicide Subscale (DSI-SS), Acquired Capability for Suicide Scale (ACSS) and The Interpersonal Needs Questionnaire (INQ). All of the tweets of each participant were combined into a single file and analyzed with LIWC. To increase interpretability, the authors opted for decision tree learning and achieved an F1 score of 0.62 with their model (see Figure 4.4).

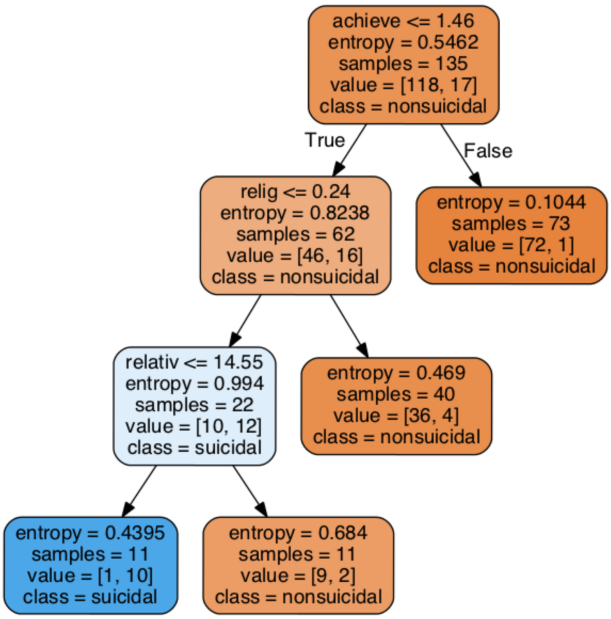

Figure 4.4. Decision Tree Learning Model. From: Braithwaite et al., 2016, p.5

In the first split, individuals that had a score higher then 1.46 in the "achieve" category of LIWC were labeled as non-suicidal. These were individuals that made use of achievement-related words more then the average person. The next node categorized the participants that had a score higher then 0.24 in the "religion" category of LIWC as non-suicidal. These participants used religion-related words more then usual. The final split was the "relativity" category of LIWC. Individuals who used words like "area", "bend", "exit", "walk", "down", "day", etc. more often were labeled as non-suicidal.

In conclusion, Braithwaite et al. made use of a decision tree model to show that machine learning algorithms are able to differentiate people in terms of suicide risk by making use of Twitter data.

4.2.7 Detecting suicidality on Twitter (Odea et al., 2015)

In their study, Odea et al. [55] first collected data from Twitter using previously defined search phrases related to suicide, such as "suicide", "kill myself", "my suicide note", "want to die", etc. These tweets were then annotated into three categories by three mental health researchers and two computer scientists. The categories were "strongly concerning", which meant that the tweets reflected strong suicidal ideation, "possibly concerning", which was the default category when the annotators were in doubt, and "safe to ignore", which was chosen when there was no indication of suicide risk.

For the model training, three different variations of feature extraction were explored: One of them was similar to a BoW approach, the other one made use of tf-idf and the last one filtered out words that had a document frequency of over 0.7. SVM and Logistic Regression algorithms were tested in combination with these features on a data set of 1820 tweets. The best performing model was the SVM algorithm in combination with the tf-idf method, which achieved an F1 score of 0.64 for the "strongly concerning" class. For the classes "possibly concerning" and "safe to ignore", the model achieved an F1 score of 0.83 and 0.62 respectively.

To conclude, the authors showed that machine learning algorithms are a viable way to differentiate the level of concern among suicide-related tweets.

4.2.8 Natural Language Processing of Social Media as Screening for Suicide Risk (Coppersmith et al., 2018)

For their study, Coppersmith et al. [38] have combined data from two sources: The first one is data donated through OurDataHelps.org, where people allow access to data from their social media, wearables, etc. and fill out questionnaires regarding demographic information, mental health history and previous suicide attempts. From OurDataHelps.org, the authors collected data from 186 users who have attempted suicide and 186 more that did not with similar demographics. The second data source is users that talk about their past suicide attempts on social media. The authors examine these statements to deduce the date of their suicide attempt and extract public data prior to it. They also found matched controls to this data by examining Twitter for users with the same demographics. The final data set had 418 users who had previously attempted suicide. With the addition of the control data, the final data set consisted of 395 230 posts.

The authors made use of deep learning model that included GloVe embeddings to analyse the language of the posts. This model achieved an AUC of 0.89, when it only used data for the month prior to the suicide attempt to make the classification. The AUC was 0.93, when it used data 6 months prior to the attempt. The fact that the performance is comparable suggests that the algorithm captures trait-level risk of suicide (i.e. risk of suicide over a long period of time) rather then state-level information (i.e. short period risk).

In conclusion, the authors showed that deep learning algorithms are able to successfully detect users at risk of attempting suicide.

4.2.9 Latent Suicide Risk Detection on Microblog via Suicide-Oriented Word Embeddings and Layered Attention (Cao et al., 2019)

In their study, Cao et al. [49], first construct a data set of 252,901 posts from the Sina microblog to generate their own suicide-oriented word embeddings. For this, they enrich preexisting word embeddings with domain information, for which they employ a masked-classification task. This task involves replacing all suicide related words in a suicidal ideation post with the word "mask" and labeling it as "non-suicidal". The sentences labeled as "suicidal" only contain randomly inserted "mask"s. This also prevents the classifier from labeling sentences just by whether it includes the word "mask".

They then develop a Suicide Detection Model (SDM), which employs suicide-oriented word embeddings, an LSTM layer to extract text features, a layer that extracts image features and a layer that extracts user features such as gender, screen name length, post count, follower count, following count, number of posts with picture and posting time.

The authors then constructed two data sets from Sina Weibo. The first one consisted of 190087 pieces of text from the microblog group under Zoufan's online suicide note. These were annotated by four doctoral students majoring in computational mental healthcare as "at suicide risk" if they expressed suicidal thoughts more then 5 times on different days and "not at suicide risk" otherwise. This data set was not considered in the suicide risk detection model. The second data set consisted of 252,901 suicidal and 491,130 non-suicidal posts.

Three sets of experiments were conducted: The first one compared the performance of LSTM and SDM in combination with seven different word embeddings. The results of the first experiment are presented in Table 4.3. Then, a comparison of LSTM, SDM, SVM and Naïve Bayes was made on the full test set of 6000 pieces of text (600 users' most recent 100 posts) and another harder subset of 1300 pieces of text, which was filtered from the original test set. This contained 130 users at suicide risk, who did not show obvious suicidal ideation on their posts that were not under Zoufan's online suicide note, creating a focused control group. The machine learning classifier results are presented in Table 4.3. Finally, the authors made an ablation test of SDM and the results of the test are presented in Table 4.3.

All in all, the authors developed a suicide-oriented word embedding and a new SDM and compared their performance to other popular word embedding methods and classifiers.

4.2.10 Topic Model for Identifying Suicidal Ideation in Chinese Microblog (Huang et al., 2015)

For their study, Huang et al. [31] identified publicly reported suicide cases between 2011 and 2014 and collected their data from Sina Weibo. This data was then annotated by six experts as suicidal if it was voted by at least 3 of them. It was annotated as "non-suicidal" otherwise. This resulted in a total of 7314 posts, 664 of them being "suicidal".

The authors then constructed an extended suicide lexicon by running word2vec on 100 million microblogs. They categorized the collected suicide words and phrases corresponding to 12 suicide warning signs. They also extended their lexicon by adding self-reference words such as "I", "me", etc. These would co-occur more often with suicidal ideation words in suicidal posts then non-suicidal posts.

For their machine learning model, they included several features: Knowledge based features (number of positive, negative, suicide words and reference words in line with the extended lexicon), syntactic features extracted using POS, topic modeling using LDA, advanced topic model which incorporates sentiment dictionaries as extra layers, posting type (original creation or retweet), posting time, social relationships and n-grams.

The authors first compare the F1 scores of models using LDA with different number of topics ranging from 100 to 1000 and the suicide lexicon. LDA with 900 topics yielded the highest F1 score with 0.603. They then test out the advanced topic model with different numbers of topics and 500 topics yielded the highest F1 score with 0.762. This reflects a significant improvement from the original topic model with LDA.

They then test several classifiers using all of the features described above with 500 topics. The F1 scores of these classifiers are reported in Table 4.3.

In conclusion, the authors developed an extended suicide lexicon and an advanced topic model to increase the performance of classifiers for SID and compared different machine learning algorithms to identify the best performing one.

4.2.11 Proactive Suicide Prevention Online (PSPO): Machine Identification and Crisis Management for Chinese Social Media Users With Suicidal Thoughts and Behaviors (Liu et al., 2019)

For their study, Liu et al. [50] collected comments left on Zoufan's online suicide note and 10000 posts without suicidal thoughts and behaviors. The resulting 27007 comments were annotated by 5 psychology postgraduates to use as the training sample. A binary classification model was tested in combination with n-gram features, domain knowledge features (DKF) making use of a generic suicide-

related lexicon and theory motivated features (TMF) including personal traits and depression. The performance of the machine learning models are reported in Table 4.3.

After identifying the suicidal ideation posts from the test sample using the machine learning algorithms, the authors contacted the 12486 users who expressed suicidal ideation via direct message. 4318 of them completed the assessment protocol, which consisted of the two questions "Do you have a plan to commit suicide?" and "Have you ever attempted suicide?", and the 9-Item Patient Health Questionnaire (PHQ-9) [56]. Help-seeking behavior and the acceptability of this proactive help program was also measured. The participants were mostly female, students or unemployed, single and had a college degree.

The respondents showed moderately severe depressive symptoms on the PHQ-9, most of them thought they would be better of dead and nearly half of them had a suicide plan. Out of the 1403 valid samples, 545 had attempted suicide and two thirds of the participants had not received any psychological treatment in the past. Two thirds of them found the help through direct messaging acceptable.

To assess the effectiveness of the direct message, the authors analysed the language of the respondents social media posts one month after the completion of the program and compared it with one month prior the onset of the program. For this, they used the Simplified Chinese Linguistic Inquiry and Word Count (SCLIWC), which is the modified version of LIWC to function better in Simplified Chinese. They found that the use of death related words in their posts significantly declined and the use of future-oriented words significantly increased after completion of the program.

Overall, the authors tested out machine learning models for SID from Sina Weibo and created a proactive suicide prevention program. They found that by providing crisis management to individuals identified as at risk by the machine learning algorithms, they were able to change the language use of the users significantly, which indicates some primary evidence for the efficacy of the PSPO.

4.2.12 Binary Suicidal Ideation Classification Results for Microblog Data Set

The F1 scores of the machine learning models tested by [41], [52], [53], [43], [49], [31] and [50] are presented in Table 4.3. Out of these studies, the best F1 score of 0.9646 was achieved by Ji et al. with a model that integrated statistical features, LDA and tf-idf in combination with Random Forest classifier.

Table 4.3

F1 scores of the Microblog data set for suicidal ideation classification.

Study	Platform	Data	Sample size	Model	Classifier	F1
Ramírez-Cifuentes et al., 2020	Twitter	generic	448037	BoW, full profile	MLP	0.81
				word embeddings, full profile	CNN	0.82
				BoW, SPV	MLP	0.86
				word embeddings, SPV	CNN	0.82
				SNPSY	LR	0.87
				BoW+SNPSY	LR	0.87
		focused	1080228	BoW, full profile	MLP	0.79
				word embeddings, full profile	CNN	0.79
				BoW, SPV	MLP	0.83
				word embeddings, SPV	CNN	0.82
				SNPSY	LR	0.85
				BoW+SNPSY	LR	0.85
Chiroma et al., 2018	Twitter	generic	289	BoW	DT	0.79
				BoW	Random Forest	0.81
				BoW	Naïve Bayes	0.69
				BoW	SVM	0.80
				BoW	Prism	0.85
Du et al., 2017	Twitter	focused	3000	GloVe embedding	CNN	0.83
				GloVe embedding	SVM	0.81
				GloVe embedding	Extra Trees	0.79
				GloVe embedding	Random Forest	0.77
				GloVe embedding	LR	0.80
				GloVe embedding	Bi-LSTM	0.81
Ji et al., 2018	Twitter	generic	10288	Statistics+LDA+tf-idf	Random Forest	0.9646
				Statistics+LDA+tf-idf	GBDT	0.9503
				Statistics+LDA+tf-idf	XGBoost	0.9597
				Statistics+LDA+tf-idf	SVM	0.9497
				Statistics+LDA+tf-idf	MLFFNN	0.9421
				Statistics+LDA+tf-idf	LSTM	0.9059
Cao et al., 2019	Sina Weibo	generic	6000	suicide-oriented-FastText	LSTM	0.881
				suicide-oriented-FastText	SVM	0.69
				suicide-oriented-FastText	Naïve Bayes	0.701
		focused	1300	suicide-oriented-FastText	SVM	0.641
				suicide-oriented-FastText	LSTM	0.753
				suicide-oriented-FastText	Naïve Bayes	0.622
Huang et al., 2015	Sina Weibo	generic	7314	LDA	SVM	0.603
				LDA+SA	SVM	0.762
				LDA+SA+meta features	SVM	0.768
				LDA+SA+meta features	Logistic	0.53
				LDA+SA+meta features	J48	0.80
				LDA+SA+meta features	Random Forest	0.713
				LDA+SA+meta features	Random Tree	0.677
				LDA+SA+meta features	Decision Table	0.746
Liu et al., 2019	Sina Weibo	generic	387823	n-gram	SVM	0.83
				n-gram+DKF	SVM	0.84
				n-gram+DKF+TMF	SVM	0.85
				n-gram	Decision Tree	0.74
				n-gram+DKF	Decision Tree	0.76
				n-gram+DKF+TMF	Decision Tree	0.76
				n-gram	Random Forest	0.80
				n-gram+DKF	Random Forest	0.79
				n-gram+DKF+TMF	Random Forest	0.78
				n-gram	LR	0.83
				n-gram+DKF	LR	0.83
				n-gram+DKF+TMF	LR	0.84

Chapter 5

Discussion

In this section, I will first provide a summary and comparison of the studies presented in the results section. Further, I will be discussing some of the limitations of the presented studies, such as the lack of intention understanding, data deficiency, annotation bias and the lack of generalizability of the results. I will then go on to talk about some of the ethical concerns that come up from conducting mental health research from publicly available social media data, including informed consent, sensitive data and anonymity. Finally, I will be discussing how the findings from these studies can be applied for suicide prevention, which is the main goal of automatic SID from social media posts.

When looking at the studies that made use of data from Reddit, Tadesse et al. compared traditional machine learning algorithms and deep learning methods in terms of their performance on SID and found that deep learning methods yielded better results than the traditional classifiers. A combination of LSTM and CNN was the best performing out of the deep learning models and Random Forest was the best performing traditional classifier. Aladag et al. not only tested out traditional classifiers for binary classification but also validated the assumption that all posts in SuicideWatch are suicidal and all posts in the ShowerThoughts subreddit are non-suicidal, which is useful for further research on SID on Reddit. They not only tested their classification models on generic data sets but also on focused data sets where it is harder to distinguish between suicidal and non-suicidal posts. LR and SVM performed best on the generic data set, while LR yielded the best results for the focused data set. De Choudry et al. collected data from even more subreddits and identified different measures that signal the transition from discussing mental health topics not related to suicide to talking about suicide. They also developed a machine learning algorithm with the aim of projecting this shift. Ji et al. compared the performance of traditional and deep learning algorithms on data collected both from Reddit and Twitter.

Regarding the studies that collected data from Twitter, Ramírez-Cifuentes et al. included multi-platform information for their SID model and developed a SPV which is useful for reducing the noise caused by writings unrelated to suicide. The best performing classifier for both control groups was LR. They also identified

the differences between suicidal ideation risk groups, focused control groups and generic control groups. Burnap et al. also developed a machine learning model to differentiate between suicidal ideation and other types of communication related to suicide. Chiroma et al. proposed a new algorithm called Prism to be used for SID and compared it to other traditional machine learning algorithms and found that the Prism algorithm performed better than them. Du et al. used deep neural networks for binary classification of suicidal ideation and also identified psychiatric stressors from suicide related tweets. They reached the best results with the Bi-LSTM classifier. Braithwaite et al. used a decision tree model to distinguish between people in terms of their suicide risk where they achieved an F1 score of 0.62. Odea et al. used machine learning classifiers to differentiate the level of concern between suicide-related tweets. Coppersmith et al. also made use of deep learning algorithms and were able to successfully identify users at risk of suicide 6 months prior to their suicide attempt.

When looking at the studies that used data from Sina Weibo, Cao et al. developed a suicide-oriented word embedding and a new SDM to compare to other widely used word embedding methods and classifiers. They tested out their models on both generic and focused data sets and reached the best F1 scores with the LSTM classifier on both data sets. Huang et al. developed an extended suicide lexicon and an advanced topic model to improve the performance of classifiers for SID. Using these, the best performing model reached an F1 score of 0.8 with a J48 classifier. Liu et al. compared machine learning models in terms of performance for SID and the best performing classifier was SVM with an F1 score of 0.85. They also developed a proactive suicide prevention program, which successfully changed the language use of suicidal users.

Overall, deep learning algorithms performed better than traditional classifiers in terms of SID on both long and short format posts. Including more features in the model also generally improved the performance of the classifiers. Further, using a focused control group resulted in lower F1 scores compared to generic control groups.

5.1 Limitations

One of the major limitations of the studies on suicidal ideation detection is that the exact mental health status of the users who are classified by the algorithms as expressing suicidal ideation are unknown [53]. The fact that a user made a suicidal ideation post does not indicate with certainty that the person is suicidal. Also, because in most studies there is no data on whether the user who made a suicidal ideation post actually attempted suicide [23], the algorithms can not predict an actual suicide attempt, but only suicidal ideation. This results in a lack of intention understanding. The causes behind suicidal behavior are an interaction of many different factors, including age, culture, cognitive and

social factors, personality and negative life events [57]. The machine learning algorithms learn statistical clues from online textual content to predict suicidal ideation but fail to understand the reasoning behind the posts [12]. Making use of temporal information (i.e. users posts over time) could help better understand the stages that lead to a suicide attempt [12], as it would integrate the change in mental status of a user in the prediction model. Future research on SID should make an effort to differentiate between suicidal ideators and suicide attempters and incorporate the psychology of suicide in their models. This way, we can improve our understanding of what leads to behavioral enactment, which would assist the development of effective prevention and intervention methods [57].

Another limitation of these studies is data deficiency. Most of the prediction models make use of supervised learning, which requires manually annotated data. Unfortunately, the amount of labeled data available is not sufficient for further research [12]. Future efforts should be made towards creating a reliable labeled data set, which ideally also includes demographic information, multimodal data and data with social relationships.

Annotation bias is also a concern when it comes to labeling data for supervised learning. Using crowdsourcing based annotation makes it harder to minimize the biases caused by manual annotation, even though it is a practical way of creating a big data set. Studies try to counteract this by using predefined annotation rules [43], having multiple annotators [42] and having clinician annotators [41] but the annotators still do not reach a full agreement in most cases, although the agreement rate is generally around 70-90%. Defining standardised annotation rules for SID could help negate annotation bias and provide a basis for future studies.

Finally, in many cases, the demographic information of the users are unknown, since they are not provided by the social media platform. Users who do not have social media accounts or do not make their accounts public are also excluded from the analysis, which presents another limitation caused by difference in the nature of the users [41]. Also, in the studies that have demographic information included in their data, it is observed that there is an over-representation of females [38] [50] [51]. These factors lower the representativeness of the data and make it harder to generalize the results.

5.2 Ethical Concerns

Social media platforms present a great opportunity for automatic SID research with massive amounts of data being available online for use. The availability of public data also comes with a series of ethical concerns regarding consent, anonymity and the protection of sensitive data when conducting mental health research.

The mental health research community has existing guidelines for ethical practices in human participant research based on past experiences and public debate. In the USA, all human participant research "must be approved by a committee of at least five persons, with at least one member from outside of the institution". This committee is the Institutional Review Board (IRB). For countries that are part of the European Union, ethics committees serve the same purpose [58]. Since observational social media research from publicly available data is not considered human subject research, it is exempt from a full review by the IRB [59]. Because of this, it is the researchers responsibility to obtain and use the publicly available data to the highest possible ethical standards [60].

In traditional research approaches, informed consent for example is built into the research design [61]. In social media research however, the participant is generally not aware of their participation. A key argument against this is that the users had to agree with the terms and conditions of the social media platform, which states that their data can be accessed by thirds parties. Nevertheless, some aspects of informed consent such as the right to withdraw become complicated when a user deletes a post or their account and the researcher is not aware of it [60]. In this scenario, it is not clear if this means that user withdraws from the study. Still, it is not feasible to obtain consent from each individual when the research involves millions of users [62]. It could also effect the users behavior [53], which would change the results of the study.

SID research involves sensitive data that entails information about peoples mental health, which might cause the individual embarrassment and reputation damage when revealed to new audiences. In particular, republishing of quotes may expose the identity of users, as they can easily be traced back by search engines [59]. This might also be a problem if the sensitive information are used by employers or insurance companies against the interest of people [41]. Because of this, proper protection of the data by restricting access to sensitive data and separating annotations from user data is essential [58].

In traditional research approaches, it is also more straightforward to make the data anonymous, whereas with social media data the anonymisation procedure is also more complex [60]. Users may be aware that their information is public, but they still might have intended it to be for a small audience [58]. To minimize the risk for users, the reproduction of posts in publications and during presentations should be avoided [60] by removing usernames and profile pictures, paraphrasing the original posts and using synthetic samples if possible. It is also considered good practice to leave out the identity of the user or other sensitive information if it is not needed for the analysis [58].

Overall it is reasonable to say that social media users' expectation of privacy should be on par with the intent and statement of privacy of a platform [59]. An example of this is the privacy policy of Twitter: "Twitter is primarily designed to help you share information with the world. Most of the information you

provide us through Twitter is information you are asking us to make public... Our default is almost always to make the information you provide through the Services public for as long as you do not delete it, but we generally give you settings or features, like protected Tweets, to make the information more private if you want." [63]. Taking this into account, it is the researchers responsibility to handle the social media data accordingly and minimize the risk of harm for the user.

5.3 Application to Suicide Prevention

The principal goal of automatic SID from social media is identifying individuals who are at risk of committing suicide to prevent them from acting on their suicidal thoughts. A longitudinal study conducted in the US between 2009-2001 shows that the majority of people who attempted suicide made a healthcare visit within a year before their attempt [64]. When an individual engages with a mental health professional, the clinician administers a standardized risk assessment, which is usually the Beck's Scale for Suicide Ideation [38]. The first 5 items of this questionnaire screens the patients attitude towards dying and items no. 6-19 are used only for patients who express an active or passive wish to attempt suicide [65]. The screening is designed to identify suicidal individuals and direct them to treatment [66]. This infrastructure of suicide risk assessment and intervention requires firstly the interaction of the individual with the healthcare system, secondly for the healthcare professional to decide that a suicide risk screening is a worthy use of time and finally for the individual to be willing to to disclose their suicidal thoughts [38]. Since improving suicide risk assessment is a part of suicide prevention [66], this process emphasises the requirement of a way to screen individuals outside the context of the healthcare system, especially the identification of signals associated with suicidality that are less obvious then explicit disclosure [38].

Automatic SID could be incorporated in the existing healthcare system by giving access to the healthcare provider the results of the risk assessment made by the machine learning algorithms from the individuals social media data, where the individual authorises the analysis of their data and the access of the healthcare professional to the results of it [38]. This could help identify at risk individuals who do not explicitly disclose their suicidal thoughts. Another way of integrating automatic SID in the healthcare system would be through a screening tool that proactively identifies suicidal individuals in the general population from publicly available data, which would require public discussion of the ethical considerations before they can be implemented [38].

Another method of intervention proposed by [12] and [50] is proactive conversational intervention. Here, individuals who are detected by the machine learning algorithms as expressing suicidal ideation are contacted for example via a direct

message. The contents of this message would include emotional and informational support, URLs for assessment protocols on suicidal thoughts and behaviors, depressive symptoms, and help-seeking behaviors and a way to contact a mental health professional. Liu et al. have shown that PSPO was an effective way of providing crisis management, which contacted users who were identified as at risk by their model through a direct message. For messages that are tailored to the individual, natural language generation techniques could be employed that generate counseling responses. Automatic response generation could be a promising method of intervention that could be implemented in the future.

Bibliography

[1] "Suicide." [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/suicide

[2] World Health Organization, *Suicide worldwide in 2019: global health estimates*. Geneva: World Health Organization, 2021.

[3] *Assisted suicide and suicide in Switzerland*. Neuchâtel: Bundesamt für Statistik (BFS), Oct 2016, no. 3902308. [Online]. Available: https://dam-api.bfs.admin.ch/hub/api/dam/assets/3902308/master

[4] B. Harmer, S. Lee, T. v. H. Duong, and A. Saadabadi, "Suicidal Ideation," in *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2022. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK565877/

[5] K. van Heeringen, "Stress–Diathesis Model of Suicidal Behavior," in *The Neurobiological Basis of Suicide*, ser. Frontiers in Neuroscience, Y. Dwivedi, Ed. Boca Raton (FL): CRC Press/Taylor & Francis, 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK107203/

[6] T. E. Joiner, K. A. Van Orden, T. K. Witte, E. A. Selby, J. D. Ribeiro, R. Lewis, and M. D. Rudd, "Main Predictions of the Interpersonal-Psychological Theory of Suicidal Behavior: Empirical Tests in Two Samples of Young Adults," *Journal of abnormal psychology*, vol. 118, no. 3, pp. 634–646, Aug. 2009. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2846517/

[7] R. C. O'Connor and O. J. Kirtley, "The integrated motivational–volitional model of suicidal behaviour," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 373, no. 1754, p. 20170268, Sep. 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6053985/

[8] E. D. Klonsky, A. M. May, and B. Y. Saffer, "Suicide, Suicide Attempts, and Suicidal Ideation," *Annual Review of Clinical Psychology*, vol. 12, pp. 307–330, 2016.

[9] Y. Hernández-Díaz, T. B. González-Castro, C. A. Tovilla-Zárate, I. E. Juárez-Rojop, M. L. López-Narváez, N. Pérez-Hernández, J. M. Rodríguez-Pérez, and A. D. Genis-Mendoza, "Association between FKBP5 polymorphisms and depressive disorders or suicidal behavior: A systematic review and meta-analysis study," *Psychiatry Research*, vol. 271, pp. 658–668, Jan. 2019.

[10] "ICD-11 for Mortality and Morbidity Statistics." [Online]. Available: https://icd.who.int/browse11/l-m/en#/http%3a%2f%2fid.who.int%2ficd%2fentity%2f778734771

[11] M. K. Nock, G. Borges, E. J. Bromet, J. Alonso, M. Angermeyer, A. Beautrais, R. Bruffaerts, T. Chiu, G. de Girolamo, S. Gluzman, R. de Graaf, O. Gureje, J. M. Haro, Y. Huang, E. Karam, R. C. Kessler, J. P. Lepine, M. E. Medina-Mora, Y. Ono, J. Posada-Villa, and D. R. Williams, "Cross-National Prevalence and Risk Factors for Suicidal Ideation, Plans, and Attempts," p. 21, 2009.

[12] S. Ji, S. Pan, X. Li, E. Cambria, G. Long, and Z. Huang, "Suicidal Ideation Detection: A Review of Machine Learning Methods and Applications," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 1, pp. 214–226, Feb. 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9199553/

[13] J. Robinson, G. Cox, E. Bailey, S. Hetrick, M. Rodrigues, S. Fisher, and H. Herrman, "Social media and suicide prevention: a systematic review: Suicide prevention and social media," *Early Intervention in Psychiatry*, vol. 10, no. 2, pp. 103–121, Apr. 2016. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/eip.12229

[14] E. Hoffer, "Nature is a self-made machine, more perfectly automated than any automated machine. To create something in the image of nature is to create a machine, and it was by learning the inner working of nature that man became a builder of machines." *Machine Learning*, p. 18, Apr. 2015.

[15] A. L. Samuel, "Some Studies in Machine Learning Using the Game of Checkers," *IBM Journal of Research and Development*, vol. 3, no. 3, pp. 210–229, Jul. 1959.

[16] Q. Bi, K. E. Goodman, J. Kaminsky, and J. Lessler, "What is Machine Learning? A Primer for the Epidemiologist," *American Journal of Epidemiology*, p. kwz189, Oct. 2019. [Online]. Available: https://academic.oup.com/aje/advance-article/doi/10.1093/aje/kwz189/5567515

[17] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification, 2nd Edition.* John Wiley & Sons, Inc., Nov. 2000.

[18] B. Mahesh, "Machine Learning Algorithms - A Review," vol. 9, no. 1, p. 7, 2018.

[19] A. Khan, B. Baharudin, L. H. Lee, and K. Khan, "A Review of Machine Learning Algorithms for Text-Documents Classification," *Journal of Advances in Information Technology*, vol. 1, no. 1, pp. 4–20, Feb. 2010. [Online]. Available: http://www.jait.us/index.php?m=content&c=index&a=show&catid=160&id=859

[20] D. B. Dwyer, P. Falkai, and N. Koutsouleris, "Machine Learning Approaches for Clinical Psychology and Psychiatry," *Annual Review of Clinical Psychology*, vol. 14, no. 1, pp. 91–118, May 2018. [Online]. Available: https://www.annualreviews.org/doi/10.1146/annurev-clinpsy-032816-045037

[21] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective," *Neurocomputing*, vol. 300, pp. 70–79, Jul. 2018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0925231218302911

[22] C. Musto, G. Semeraro, M. De Gemmis, and P. Lops, "Word Embedding techniques for Content-based Recommender Systems: an empirical evaluation," p. 2, 2015.

[23] A. E. Aladağ, S. Muderrisoglu, N. B. Akbas, O. Zahmacioglu, and H. O. Bingol, "Detecting Suicidal Ideation on Forums: Proof-of-Concept Study," *Journal of Medical Internet Research*, vol. 20, no. 6, p. e9840, Jun. 2018, company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada. [Online]. Available: https://www.jmir.org/2018/6/e215

[24] T. Pickard, "Comparing word2vec and GloVe for Automatic Measurement of MWE Compositionality," in *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*. online: Association for Computational Linguistics, Dec. 2020, pp. 95–100. [Online]. Available: https://aclanthology.org/2020.mwe-1.12

[25] K. W. Church, "Word2Vec," *Natural Language Engineering*, vol. 23, no. 1, pp. 155–162, Jan. 2017, publisher: Cambridge University Press. [Online]. Available: https://www.cambridge.org/core/journals/natural-language-engineering/article/word2vec/B84AE4446BD47F48847B4904F0B36E0B

[26] J. Pennington, R. Socher, and C. Manning, "GloVe: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. [Online]. Available: https://aclanthology.org/D14-1162

[27] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The Development and Psychometric Properties of LIWC2015," p. 26, 2015.

[28] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*,

vol. 374, no. 2065, p. 20150202, Apr. 2016. [Online]. Available: https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0202

[29] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, Jul. 2015. [Online]. Available: https://www.science.org/doi/10.1126/science.aaa8415

[30] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao, "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey," *Multimedia Tools and Applications*, vol. 78, no. 11, pp. 15 169–15 211, Jun. 2019. [Online]. Available: http://link.springer.com/10.1007/s11042-018-6894-4

[31] X. Huang, X. Li, T. Liu, D. Chiu, T. Zhu, and L. Zhang, "Topic Model for Identifying Suicidal Ideation in Chinese Microblog," in *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, Shanghai, China, Oct. 2015, pp. 553–562. [Online]. Available: https://aclanthology.org/Y15-1064

[32] Y. Zhang, H. Zhang, J. Cai, and B. Yang, "A Weighted Voting Classifier Based on Differential Evolution," *Abstract and Applied Analysis*, vol. 2014, pp. 1–6, 2014. [Online]. Available: http://www.hindawi.com/journals/aaa/2014/376950/

[33] R. K. Dewang and A. K. Singh, "State-of-art approaches for review spammer detection: a survey," *Journal of Intelligent Information Systems*, vol. 50, no. 2, pp. 231–264, Apr. 2018. [Online]. Available: http://link.springer.com/10.1007/s10844-017-0454-7

[34] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001. [Online]. Available: https://doi.org/10.1023/A:1010933404324

[35] E. Russek, R. A. Kronmal, and L. D. Fisher, "The effect of assuming independence in applying Bayes' Theorem to risk estimation and classification in diagnosis," *Computers and Biomedical Research*, vol. 16, no. 6, pp. 537–552, Dec. 1983. [Online]. Available: https://www.sciencedirect.com/science/article/pii/001048098390040X

[36] O. I. Abiodun, A. Jantan, A. E. Omolara, K. V. Dada, N. A. Mohamed, and H. Arshad, "State-of-the-art in artificial neural network applications: A survey," *Heliyon*, vol. 4, no. 11, p. e00938, Nov. 2018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2405844018332067

[37] N. Seliya, T. M. Khoshgoftaar, and J. Van Hulse, "A Study on the Relationships of Classifier Performance Metrics," in *2009 21st IEEE International Conference on Tools with Artificial Intelligence*. Newark, NJ: IEEE, Nov. 2009, pp. 59–66. [Online]. Available: https://ieeexplore.ieee.org/document/5364367/

[38] G. Coppersmith, R. Leary, P. Crutchley, and A. Fine, "Natural Language Processing of Social Media as Screening for Suicide Risk," *Biomedical Informatics Insights*, vol. 10, p. 1178222618792860, Jan. 2018, publisher: SAGE Publications Ltd STM. [Online]. Available: https://doi.org/10.1177/1178222618792860

[39] J. M. Lobo, A. Jiménez-Valverde, and R. Real, "AUC: a misleading measure of the performance of predictive distribution models," *Global Ecology and Biogeography*, vol. 17, no. 2, pp. 145–151, Mar. 2008. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/j.1466-8238.2007.00358.x

[40] W. Geyser, "What is a Microblog? [And Why Do You Need One]," Feb. 2020. [Online]. Available: https://influencermarketinghub.com/what-is-a-microblog/

[41] D. Ramírez-Cifuentes, A. Freire, R. Baeza-Yates, J. Puntí, P. Medina-Bravo, D. A. Velazquez, J. M. Gonfaus, and J. Gonzàlez, "Detection of Suicidal Ideation on Social Media: Multimodal, Relational, and Behavioral Analysis," *Journal of Medical Internet Research*, vol. 22, no. 7, p. e17758, Jul. 2020, company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada. [Online]. Available: https://www.jmir.org/2020/7/e17758

[42] M. De Choudhury, E. Kiciman, M. Dredze, G. Coppersmith, and M. Kumar, "Discovering Shifts to Suicidal Ideation from Mental Health Content in Social Media," *Proceedings of the SIGCHI conference on human factors in computing systems . CHI Conference*, vol. 2016, pp. 2098–2110, May 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5659860/

[43] S. Ji, C. P. Yu, S.-f. Fung, S. Pan, and G. Long, "Supervised Learning for Suicidal Ideation Detection in Online User Content," *Complexity*, vol. 2018, p. e6157249, Sep. 2018, publisher: Hindawi. [Online]. Available: https://www.hindawi.com/journals/complexity/2018/6157249/

[44] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Detection of Suicide Ideation in Social Media Forums Using Deep Learning," *Algorithms*, vol. 13, no. 1, p. 7, Jan. 2020, number: 1 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/1999-4893/13/1/7

[45] M. Efron, "Information search and retrieval in microblogs," *Journal of the American Society for Information Science and Technology*, vol. 62, no. 6, pp. 996–1008, 2011, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.21512. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.21512

[46] Q. Gao, F. Abel, G.-J. Houben, and Y. Yu, "A Comparative Study of Users' Microblogging Behavior on Sina Weibo and Twitter," in *User Modeling, Adaptation, and Personalization*, D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, J. Masthoff, B. Mobasher, M. C. Desmarais, and R. Nkambou, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, vol. 7379, pp. 88–101, series Title: Lecture Notes in Computer Science. [Online]. Available: http://link.springer.com/10.1007/978-3-642-31454-4_8

[47] "_us200 million Tweets per day." [Online]. Available: https://blog.twitter.com/en_us/a/2011/200-million-tweets-per-day

[48] F. Yang, Y. Liu, X. Yu, and M. Yang, "Automatic detection of rumor on Sina Weibo," in *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics - MDS '12*. Beijing, China: ACM Press, 2012, pp. 1–7. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2350190.2350203

[49] L. Cao, H. Zhang, L. Feng, Z. Wei, X. Wang, N. Li, and X. He, "Latent Suicide Risk Detection on Microblog via Suicide-Oriented Word Embeddings and Layered Attention," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 1718–1728. [Online]. Available: https://aclanthology.org/D19-1181

[50] X. Liu, X. Liu, J. Sun, N. X. Yu, B. Sun, Q. Li, and T. Zhu, "Proactive Suicide Prevention Online (PSPO): Machine Identification and Crisis Management for Chinese Social Media Users With Suicidal Thoughts and Behaviors," *Journal of Medical Internet Research*, vol. 21, no. 5, p. e11705, May 2019, company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada. [Online]. Available: https://www.jmir.org/2019/5/e11705

[51] P. Burnap, G. Colombo, R. Amery, A. Hodorog, and J. Scourfield, "Multi-class machine classification of suicide-related communication on Twitter," *Online Social Networks and Media*, vol. 2, pp. 32–44, Aug. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2468696417300605

[52] F. Chiroma, H. Liu, and M. Cocea, "Suiciderelated Text Classification With Prism Algorithm," in *2018 International Conference on Machine Learning and Cybernetics (ICMLC)*. Chengdu: IEEE, Jul. 2018, pp. 575–580. [Online]. Available: https://ieeexplore.ieee.org/document/8527032/

[53] J. Du, Y. Zhang, J. Luo, Y. Jia, Q. Wei, C. Tao, and H. Xu, "Extracting psychiatric stressors for suicide from social media using deep learning," *BMC Medical Informatics and Decision Making*, vol. 18, no. 2, p. 43, Jul. 2018. [Online]. Available: https://doi.org/10.1186/s12911-018-0632-8

[54] S. R. Braithwaite, C. Giraud-Carrier, J. West, M. D. Barnes, and C. L. Hanson, "Validating Machine Learning Algorithms for Twitter Data Against Established Measures of Suicidality," *JMIR Mental Health*, vol. 3, no. 2, p. e4822, May 2016, company: JMIR Mental Health Distributor: JMIR Mental Health Institution: JMIR Mental Health Label: JMIR Mental Health Publisher: JMIR Publications Inc., Toronto, Canada. [Online]. Available: https://mental.jmir.org/2016/2/e21

[55] B. O'Dea, S. Wan, P. J. Batterham, A. L. Calear, C. Paris, and H. Christensen, "Detecting suicidality on Twitter," *Internet Interventions*, vol. 2, no. 2, pp. 183–188, May 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2214782915000160

[56] K. Kroenke and R. L. Spitzer, "The PHQ-9: A New Depression Diagnostic and Severity Measure," *Psychiatric Annals*, vol. 32, no. 9, pp. 509–515, Sep. 2002, publisher: SLACK Incorporated. [Online]. Available: https://journals.healio.com/doi/10.3928/0048-5713-20020901-06

[57] R. O'Connor and M. Nock, "The psychology of suicidal behaviour," *The Lancet Psychiatry*, vol. 1, pp. 73–85, Jun. 2014.

[58] A. Benton, G. Coppersmith, and M. Dredze, "Ethical Research Protocols for Social Media Health Research," in *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Valencia, Spain: Association for Computational Linguistics, 2017, pp. 94–102. [Online]. Available: http://aclweb.org/anthology/W17-1612

[59] M. A. Moreno, N. Goniu, P. S. Moreno, and D. Diekema, "Ethics of Social Media Research: Common Concerns and Practical Considerations," *Cyberpsychology, Behavior, and Social Networking*, vol. 16, no. 9, pp. 708–713, Sep. 2013. [Online]. Available: http://www.liebertpub.com/doi/10.1089/cyber.2012.0334

[60] L. Townsend and C. Wallace, "Social media research: A guide to ethics," *University of Aberdeen*, vol. 1, p. 16, 2016.

[61] E. d. June 1, . w. a. e. June 1, 2010, J. 1, and . C. . A. P. A. A. r. reserved, "Ethical principles of psychologists and code of conduct." [Online]. Available: https://www.apa.org/ethics/code

[62] D. O'Connor, "The apomediated world: regulating research when social media has changed research," *The Journal of Law, Medicine & Ethics: A Jour-*

nal of the American Society of Law, Medicine & Ethics, vol. 41, no. 2, pp. 470–483, 2013.

[63] "Previous Privacy Policy." [Online]. Available: https://twitter.com/en/privacy/previous/version_12

[64] B. K. Ahmedani, C. Stewart, G. E. Simon, F. Lynch, C. Y. Lu, B. E. Waitzfelder, L. I. Solberg, A. A. Owen-Smith, A. Beck, L. A. Copeland, E. M. Hunkeler, R. C. Rossom, and L. K. Williams, "Racial/ethnic differences in healthcare visits made prior to suicide attempt across the United States," *Medical care*, vol. 53, no. 5, pp. 430–435, May 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4397662/

[65] A. T. Beck, G. K. Brown, and R. A. Steer, "Psychometric characteristics of the Scale for Suicide Ideation with psychiatric outpatients," *Behaviour Research and Therapy*, vol. 35, no. 11, pp. 1039–1046, Nov. 1997. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0005796797000739

[66] J. J. Mann, A. Apter, J. Bertolote, A. Beautrais, D. Currier, A. Haas, U. Hegerl, J. Lonnqvist, K. Malone, A. Marusic, L. Mehlum, G. Patton, M. Phillips, W. Rutz, Z. Rihmer, A. Schmidtke, D. Shaffer, M. Silverman, Y. Takahashi, A. Varnik, D. Wasserman, P. Yip, and H. Hendin, "Suicide Prevention StrategiesA Systematic Review," *JAMA*, vol. 294, no. 16, pp. 2064–2074, 10 2005. [Online]. Available: https://doi.org/10.1001/jama.294.16.2064