THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

Suicidal Ideation Detection in Online Social Content

Shaoxiong Ji

A thesis submitted for the degree of Masters of Philosophy at

The University of Queensland in

School of Information Technology and Electrical Engineering

Abstract

Suicide has become a serious social health issue in the modern society. Suicidal ideation is people's thoughts about committing suicide. Many factors such as long-term exposure to negative feelings or life events can lead to suicidal ideation and suicide attempts. Among all the approaches of suicide prevention, early detection of suicidal ideation is one of the most effective ways. The advances of online communication and social networking services also provide a platform for people to express their sufferings and feelings in the real world, which provides a source for suicidal ideation detection. This thesis investigates the online social content for early detection of suicidal ideation.

User-generated content, especially text posted by users, contains rich information about people's status and reflects their mental states. In this thesis, we firstly have a comprehensive content analysis to discover knowledge from suicide-related text and preforms a benchmarking on binary classification of suicidal ideation including using feature extraction based classifiers and deep neural networks. The reasons of committing suicide are complicated, and suicidal factors vary from individuals. To incorporate suicidal factors for suicidal intention understanding, we consider sentimental clues and topics in people's posts and propose to reason the relations between those factors and posts with attention relation networks for fine-grained suicidal ideation detection. Lastly, we study suicidal ideation detection in another scenario of private chatting. To tackle the challenge of isolated data in private chat rooms, we develop a knowledge transferring framework to train a global model for knowledge sharing with distributed agents.

Overall, early detection suicidal ideation is urgently in demand for suicide prevention. This thesis develops methods with content analysis, feature engineering, and deep learning techniques including deep neural networks, attentive relation networks and federated transfer learning in the hope of using effective suicidal ideation detection to prevent suicide and save people's life.

Keywords

Suicidal ideation detection, online content, feature engineering, relation networks, knowledge transferring

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Mental health issues, such as anxiety and depression, are becoming increasingly concerning in modern society, as they turn out to be especially severe in developed countries and emerging markets. Severe mental disorders without effective treatment can turn to suicidal ideation or even suicide attempts. Some online posts contain a large amount of negative information and generate problematic phenomena such as cyberstalking and cyberbullying. Consequences can be severe and risky since such bad information are often engaged in some form of social cruelty, leading to rumors or even mental damages. Research shows that there is a link between cyberbullying and suicide [1]. Victims overexposed to too much negative messages or events may become depressed and desperate, even worse, some may commit suicide.

1.1 Suicide Statistics

Every year thousands of people around the world fall victims to suicide, making suicide prevention become a critical global public health mission. According to a WHO report[1], 1 in 4 people worldwide are suffering from mental disorders to some extent. Without effective treatment, severe mental disorders but without effective treatment are very likely to turn to suicide. About 900,000 people all over the world commit suicide each year. A research from Suicide Prevention Australia (SPA) in collaboration with the University of New England provided a understanding on the exposure and impact of suicide in Australia based on a survey with 3,000 respondents [2]. It reported that 89% of the respondents were exposed to at least one suicide attempt, 85% were exposed to at least one suicide death, and 80% were exposed to both suicide attempt and death. As for their access to healthcare support six months prior to death, only 36% of respondents were supported, with 19% not supported and 26% don't know, reported based on their knowledge about their own healthcare use. According to a previous report in U.S.A, 2.2

[1] Suicide rates, Global Health Observatory (GHO) data in 2015 from WHO, available in `http://www.who.int/gho/mental_health/suicide_rates/en/`

[2] Suicide Prevention Australia (SPA). Findings from the exposure and impact of suicide in Australia survey, available in `https://www.suicidepreventionaust.org/exposure-and-impact-survey`. Retrieved in Sep 2018

million people estimated nationwide had made suicide plans during the period of 2008-2009 [2]. And a large number of people, especially teenagers, were reported having suicidal ideations.

1.2 Suicide Factors

The reasons people commit suicide are complicated. People with depression are highly likely to commit suicide, but many without depression can also have suicidal thoughts [3]. Nock et al. [4] reported prevalence and suicide factors across 17 countries, and found that risk factors consist of being female, younger, less educated, unmarried, and having mental health issues. According to the American Foundation for Suicide Prevention (AFSP), suicide factors fall under under three categories: health factors, environment factors, and historical factors [5]. Ferrari et al. [6] found that mental health issue and substance use disorders are attributed to the factors of suicide. O'Connor and Nock [7] conducted a thorough review about the psychology of suicide, and summarized psychological risks as personality and individual differences, cognitive factors, social factors, and negative life events.

1.3 Suicide and Internet

Due to the advances of social media and online anonymity, an increasing number of individuals turn to interact with others on the Internet. Online communication channels are becoming a new way for people to express their their feelings, suffering, and suicidal tendencies. Hence, online channels have naturally started to act as a surveillance tool for suicidal ideation, and mining social content can improve suicide prevention [8]. In addition, strange social phenomena are emerging, e.g., online communities reaching an agreement on self-mutilation and copycat suicide. For example, a social network phenomenon call the "Blue Whale Game"[3] in 2016 uses many tasks (such as self-harming) and leads game members to commit suicide in the end. Suicide is a critical social issue and takes thousands of lives every year. Thus, it is necessary to detect suicidality and to prevent suicide before victims end their life. Early detection and treatment are regarded as the most effective ways to prevent potential suicide attempts.

1.4 Suicidal Ideation Detection

Potential victims with suicidal ideation may express their thoughts of committing suicide in the form of fleeting thoughts, suicide plan and role playing. Suicidal ideation detection is to find out these dangerous intentions or behaviors before tragedy strikes. The reasons of suicide are complicated and attributed to a complex interaction of many factors [7]. To detect suicidal ideation, many researchers conducted psychological and clinical studies [9] and classified responses of questionnaires [10]. Based on their social media data, artificial intelligence (AI) and machine learning techniques can predict

[3] `https://thesun.co.uk/news/worldnews/3003805`

people's likelihood of suicide [11], which paves the way for early intervention. Detection on social content focuses on feature engineering [12, 13], sentiment analysis [14, 15], and deep learning [16–18].

Mobile technologies have been studied and applied to suicide prevention, for example, the mobile suicide intervention application iBobbly [19] developed by the Black Dog Institute[4]. Many other suicide prevention tools integrated with social networking services have also been developed including Samaritans Radar[5] and Woebot[6]. The former was a Twitter plugin for monitoring alarming posts which was later discontinued because of privacy issues. The latter is a Facebook chatbot based on cognitive behavioral therapy and natural language processing (NLP) techniques for relieving people's depression and anxiety.

It is inevitable that applying cutting-edge AI technologies for suicidal ideation detection comes with privacy issues [20] and ethical concerns [21]. Linthicum et al. [22] put forward three ethical issues including the influence of bias on machine learning algorithms, the prediction on time of suicide act, and ethical and legal questions raised by false positive and false negative prediction. It is not easy to answer ethical questions for AI as these require algorithms to reach a balance between competing values, issues and interests [20].

One possible approach to preventing suicide effectively is early detection of suicidal ideation for effective intervention. Thus, developing suicidal ideation detection methods becomes an important mission. But there still remains several challenges as

- There is a limited number of benchmarks for training and evaluating suicidal ideation detection.
- Text data is noisy for effective suicidal ideation detection.
- Text with suicidal ideation and text will minor mental disorders share similar language usages, making it difficult to understand suicidal intention.
- In some scenarios such as chatting room, the isolation of data harms the performance of supervised learning models.

This thesis focuses on machine learning techniques especially deep learning models for effective suicidal ideation detection in online social content. It intends to solve three tasks, i.e., benchmarking for suicidal ideation detection, fine-grained suicidal ideation detection considering suicide risk factors, and knowledge transferring to enable effective detection in private communications.

[4] https://blackdoginstitute.org.au/research/digital-dog/programs/ibobbly-app

[5] https://samaritans.org/about-samaritans/research-policy/internet-suicide/samaritans-radar

[6] https://woebot.io

Chapter 2

Literature Review

AI has been applied to solve many challenging social problems. Detection of suicidal ideation with AI techniques is one of potential applications for social good, and should be addressed to meaningfully improve people's wellbeing. The chapter reviews suicidal ideation detection methods from the perspective of AI and machine learning and specific domain applications with social impact. The categorization from these two perspectives is shown in Fig. 2.1. The definition of suicidal ideation detection is described in Definition 1.

Definition 1 (Suicidal Ideation Detection)**.** Given tabular data of a person or textual content written by a person, suicidal ideation detection is to determine whether the person has suicidal ideation or thoughts.

We introduce and discuss both classical content analysis and recent machine learning techniques, plus their application to questionnaires, EHR data, suicide notes and online social content.

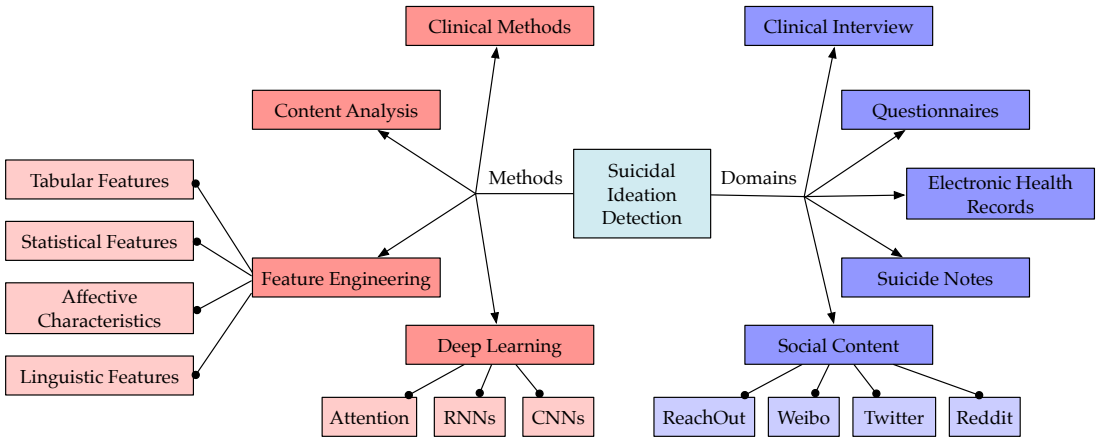


Figure 2.1: The categorization of suicide ideation detection: methods and domains

2.1 Methods and Categorization

Suicide detection has drawn attention of many researchers due to an increasing suicide rate in recent years, and has been studied extensively from many perspectives. The research techniques used to examine suicide also span many fields and methods, for example, clinical methods with patient-clinic interaction [9] and automatic detection from user generated content (mainly text) [12, 17]. Machine learning techniques are widely applied for automatic detection.

Traditional suicide detection relies on clinical methods including self-reports and face-to-face interviews. Venek et al. [9] designed a five-item ubiquitous questionnaire for the assessment of suicidal risks, and applied a hierarchical classifier on the patients' response to determine their suicidal intentions. Through face-to-face interaction, verbal and acoustic information can be utilized. Scherer [23] investigated the prosodic speech characteristics and voice quality in a dyadic interview to identify suicidal and non-suicidal juveniles. Other clinical methods examine resting state heart rate from converted sensing signals [24], classify functional magnetic resonance imaging based neural representations of death- and life-related words [25], and event-related instigators converted from EEG signals [26]. Another aspect of clinical treatment is the understanding of the psychology behind suicidal behavior [7]. This, however, relies heavily on clinician's knowledge and face-to-face interaction. Suicide risk assessment scales with clinical interview can reveal informative cues for predicting suicide [27]. Tan et al. [28] conducted an interview and survey study in Weibo, a Twitter-like service in China, to explore the engagement of suicide attempters with intervention by direct messages.

2.1.1 Content Analysis

Users' post on social websites reveals rich information and their language preferences. Through exploratory data analysis on the user generated content can have an insight to language usage and linguistic clues of suicide attempters. Related analysis includes lexicon-based filtering, statistical linguistic features, and topic modeling within suicide-related posts.

Suicide-related keyword dictionary and lexicon are manually built to enable keyword filtering [29, 30] and phrases filtering [31]. Suicide-related keywords and phrases include "kill", "suicide", "feel alone", "depressed", and "cutting myself". Vioulès et al. [5] built a point-wise mutual information symptom lexicon using annotated Twitter dataset. Gunn and Lester [32] analyzed posts from Twitter in the 24 hours prior to death of a suicide attempter. Coppersmith et al. [33] analyzed the language usage of data from the same platform. Suicidal thoughts may involve strong negative feelings, anxiety, and hopelessness, or other social factors like family and friends. Ji et al. [17] performed word cloud visualization and topics modeling over suicide-related content and found that suicide-related discussion covers both personal and social issues. Colombo et al. [34] analyzed the graphical characteristics of connectivity and communication in the Twitter social network. Coppersmith et al. [35] provided an exploratory analysis on language patterns and emotions in Twitter. Other methods and techniques include Google Trends analysis for monitoring suicide risk [36], detecting social media content and speech patterns analysis [37], assessing the reply bias through linguistic clues [38], human-machine

hybrid method for analyzing the effect of language of social support on suicidal ideation risk [39].

2.1.2 Feature Engineering

The goal of text-based suicide classification is to determine whether candidates, through their posts, have suicidal ideations. Machine learning methods and NLP have also been applied in this field.

Tabular Features

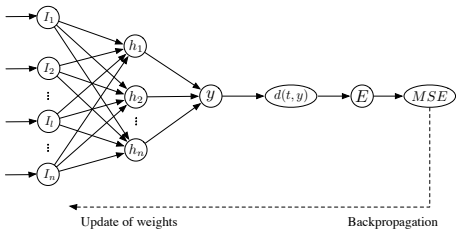
Tabular data for suicidal ideation detection consist of questionnaire responses and structured statistical information extracted from websites. Such structured data can be directly used as features for classification or regression. Masuda et al. [40] applied logistic regression to classify suicide and control groups based on users' characteristics and social behavior variables, and found variables such as community number, local clustering coefficient and homophily have a larger influence on suicidal ideation in a SNS of Japan. Chattopadhyay [41] applied Pierce Suicidal Intent Scale (PSIS) to assess suicide factors and conducted a regression analysis. Questionnaires act as a good source of tabular features. Delgado-Gomez et al. [42] used the international personal disorder examination screening questionnaire and the Holmes-Rahe social readjustment rating scale. Chattopadhyay [43] proposed to apply a multilayer feed forward neural network as shown in Fig. 2.2a to classify suicidal intention indicators according to Beck's suicide intent scale.

General Text Features

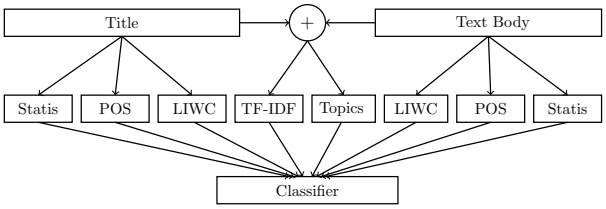
Another direction of feature engineering is to extract features from unstructured text. The main features consist of N-gram features, knowledge-based features, syntactic features, context features and class-specific features [44]. Abboute et al. [45] built a set of keywords for vocabulary feature extraction within nine suicidal topics. Okhapkina et al. [46] built a dictionary of terms pertaining to suicidal content and introduced term frequency-inverse document frequency (TF-IDF) matrices for messages and a singular value decomposition (SVD) for matrices. Mulholland and Quinn [47] extracted vocabulary and syntactic features to build a classifier to predict the likelihood of a lyricist's suicide. Huang et al. [48] built a psychological lexicon dictionary by extending HowNet (a commonsense words collection), and used a support vector machines (SVM) to detect cybersuicide in Chinese microblogs. Topic model [49] is incorporated with other machine learning techniques for identifying suicide in Sina Weibo. Ji et al. [17] extract several informative sets of features, including statistical, syntactic, linguistic inquiry and word count (LIWC), word embedding, and topic features, and then put the extracted features into classifiers as shown in Fig. 2.2b, where four traditional supervised classifiers are compared. Shing et al. [13] extracted several features as bag of words (BoWs), empath, readability, syntactic features, topic model posteriors, word embeddings, linguistic inquiry and word count, emotion features and mental disease lexicon.

Models for suicidal ideation detection with feature engineering include SVM [44], artificial neural networks (ANN) [50] and conditional random field (CRF) [51]. Tai et al. [50] selected several features

including history of suicide ideation and self-harm behavior, religious belief, family status, mental disorder history of candidates and their family. Pestian et al. [52] compared the performance of different multivariate techniques with features of word counts, POS, concepts and readability score. Similarly, Ji et al. [17] compared four classification methods of logistic regression, random forest, gradient boosting decision tree, and XGBoost. Braithwaite et al. [53] validated machine learning algorithms can effectively identify high suicidal risk.



(a) Neural network with feature engineering



(b) Classifier with feature engineering

Figure 2.2: Illustrations of methods with feature engineering

Affective Characteristics

Affective characteristics are one of the most distinct differences between those who attempt suicide and normal individuals, which has drawn great attention of both computer scientists and mental heath researchers. To detect the emotions in suicide notes, Liakata et al. [51] used manual emotion categories including anger, sorrow, hopefulness, happiness/peacefulness, fear, pride, abuse and forgiveness. Wang et al. [44] employed combined characteristics of both factual (2 categories) and sentimental aspects (13 categories) to discover fine-grained sentiment analysis. Similarly, Pestian et al. [52] identified emotions of abuse, anger, blame, fear, guilt, hopelessness, sorrow, forgiveness, happiness, peacefulness, hopefulness, love, pride, thankfulness, instructions, and information. Ren et al. [14] proposed a complex emotion topic model and applied it to analyze accumulated emotional traits in suicide blogs and to detect suicidal intentions from a blog stream. Specifically, the authors studied accumulate emotional traits including emotion accumulation, emotion covariance, and emotion transition among eight basic emotions of joy, love, expectation, surprise, anxiety, sorrow, anger, and hate with a five-level intensity.

2.1.3 Deep Learning

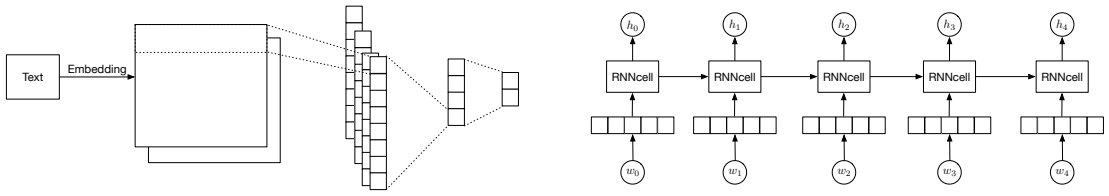
Deep learning has been a great success in many applications including computer vision, NLP, and medical diagnosis. In the field of suicide research, it is also an important method for automatic suicidal ideation detection and suicide prevention. It can effectively learn text features automatically without complex feature engineering techniques. Popular deep neural networks (DNNs) include convolutional neural networks (CNNs) and recurrent neural networks (RNNs) as shown in Fig. 2.3a and 2.3b, respectively. To apply DNNs, natural language text is usually embedded into distributed vector space

with popular word embedding techniques such as word2vec [54] and GloVe [55]. Shing et al. [13] applied user-level CNN with filter size of 3, 4 and 5 to encode users' posts. Long short-term memory (LSTM) network, a popular variant of RNN, is applied to encode textual sequences and then processed for classification with fully connected layers [17].

Recent methods introduce other advanced learning paradigms to integrate with DNNs for suicidal ideation detection. Ji et al. [56] proposed model aggregation methods for updating neural networks, i.e, CNNs and LSTMs, targeting to detect suicidal ideation in private chatting rooms. However, decentralized training relies on coordinators in chatting rooms to label user posts for supervised training, which can only applied to very limited scenarios. One possible better way is to use unsupervised or semi-supervised learning methods. Benton et al. [16] predicted suicide attempt and mental health with neural models under the framework of multi-task learning by predicting the gender of users as auxiliary task. Gaur et al. [57] incorporated external knowledge bases and suicide-related ontology into text representation, and gained an improved performance with a CNN model. Coppersmith et al. [58] developed a deep learning model with GloVe for word embedding, bidirectional LSTM for sequence encoding, and self attention mechanism for capturing the most informative subsequence. Sawhney et al. [59] used LSTM, CNN, and RNN for suicidal ideation detection. Ji et al. [60] proposed attentive relation network with LSTM and topic modeling for encoding text and risk indicators.

In the 2019 CLPsych Shared Task [61], many popular DNN architectures were applied. Hevia et al. [62] evaluated the effect of pretraining using different models including GRU-based RNN. Morales et al. [63] studied several popular deep learning models such as CNN, LSTM, and Neural Network Synthesis (NeuNetS). Matero et al. [64] proposed dual-context model using hierarchically attentive RNN, and bidirectional encoder representations form transformers (BERT).

Another sub-direction is the so-called hybrid method which cooperates minor feature engineering with representation learning techniques. Chen et al. [65] proposed a hybrid classification model of behavioral model and suicide language model. Zhao et al. [66] proposed D-CNN model taking word embedding and external tabular features as inputs for classifying suicide attempters with depression.



(a) CNN (b) RNN

Figure 2.3: Deep neural networks for suicidal ideation detection

2.1.4 Summary

The popularization of machine learning has facilitated research on suicidal ideation detection from multi-modal data and provided a promising way for effective early warning. Current research focuses on text-based methods by extracting features and deep learning for automatic feature learning. Many

Table 2.1: Categorization of methods for suicidal ideation detection

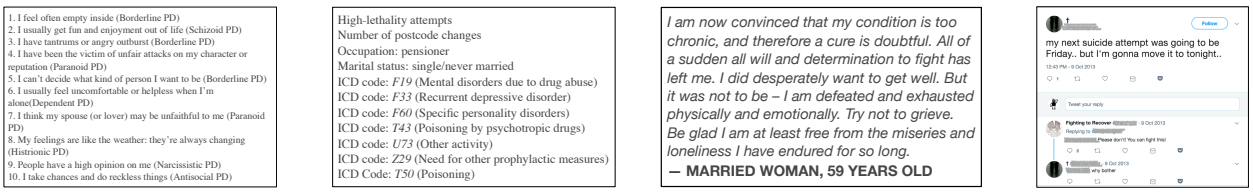
Category	Publications	Methods	Inputs
Feature Engineering	Ji et al. [17]	Word counts, POS, LIWC, TF-IDF + classifiers	Text
	Masuda et al. [40]	Multivariate/univariate logistic regression	Characteristics variables
	Delgado-Gomez et al. [42]	International personal disorder examination screening questionnaire	Questionnaire responses
	Mulholland et al. [47]	Vocabulary features, syntactic features, semantic class features, N-gram	lyrics
	Okhapkina et al. [46]	Dictionary, TF-IDF + SVD	Text
	Huang et al. [48]	Lexicon, syntactic features, POS, tense	Text
	Pestian et al. [52]	Word counts, POS, concepts and readability score	Text
	Tai et al. [50]	self-measurement scale + ANN	Self-measurement forms
	Shing et al. [13]	BoWs, empath, readability, syntactic, topic, LIWC, emotion, lexicon	Text
Deep Learning	Zhao et al. [66]	Word embedding, tabular features, D-CNN	Text+external information
	Shing et al. [13]	Word embedding, CNN, max pooling	Text
	Ji et al. [17]	Word embedding, LSTM, max pooling	Text
	Bento et al. [16]	Multi-task learning, neural networks	Text
	Hevia et al. [62]	Pretrained GRU, word embedding, document embedding	Text
	Morales et al. [63]	CNN, LSTM, NeuNetS, word embedding	Text
	Matero et al. [64]	Dual-context, BERT, GRU, attention, user-factor adaptation	Text
	Gaur et al. [57]	CNN, knowledge base, ConceptNet embedding	Text
	Coppersmith et al. [58]	GloVe, BiLSTM, self attention	Text
	Ji et al. [60]	Relation network, LSTM, attention, lexicon	Text

canonical NLP features such as TF-IDF, topics, syntactics, affective characteristics, and readability, and deep learning models like CNN and LSTM are widely used by researchers. Those methods gained preliminary success on suicidal ideation detection, but some methods may only learn statistical cues and lack of commonsense. The recent work [57] incorporated external knowledge by using knowledge bases and suicide ontology for knowledge-aware suicide risk assessment. It took a remarkable step towards knowledge-aware detection.

2.2 Applications on Domains

Many machine learning techniques have been introduced for suicidal ideation detection. The relevant extant research can also be viewed according to the data source. Specific application covers a wide range of domains including questionnaires, electronic health records (EHRs), suicide notes, and online user content. Fig. 2.4 shows some examples of data source for suicidal ideation detection, where Fig. 2.4a lists selected questions of the "International Personal Disorder Examination Screening Questionnaire" (IPDE-SQ) adapted from [42], Fig. 2.4b are selected patient's records from [67], Fig. 2.4c is a suicide note from a website[1], and Fig. 2.4d is a tweet and its corresponding comments from `Twitter.com`. Some researchers also developed softwares for suicide prevention. Berrouiguet et al. [68] developed a mobile application for health status self report. Meyer et al. [69] developed an e-PASS Suicidal Ideation Detector (eSID) tool for medical practitioners.

[1] `https://paranorms.com/suicide-notes`



(a) Questionnaire (b) EHR (c) Suicide notes (d) Tweets

Figure 2.4: Examples of content for suicidal ideation detection

2.2.1 Questionnaires

Mental disorder scale criteria such as DSM-IV[2] and ICD-10[3], and the IPDE-SQ provides good tool for evaluating an individual's mental status and their potential for suicide. Those criteria and examination metrics can be used to design questionnaires for self-measurement or face-to-face clinician-patient interview.

To study the assessment of suicidal behavior, Delgado-Gomez et al. [10] applied and compared the IPDE-SQ and the "Barrat's Impulsiveness Scale"(version 11, BIS-11) to identify people likely to attempt suicide. The authors also conducted a study on individual items from those two scales. The BIS-11 scale has 30 items with 4-point ratings, while the IPDE-SQ in DSM-IV has 77 true-false screening questions. Further, Delgado-Gomez et al. [42] introduced the "Holmes-Rahe Social Readjustment Rating Scale" (SRRS) and the IPDE-SQ as well to two comparison groups of suicide attempters and non-suicide attempters. The SRRS consists of 43 ranked life events of different levels of severity. Harris et al. [70] conducted a survey on understanding suicidal individuals' online behaviors to assist suicide prevention. Sueki [71] conducted an online panel survey among Internet users to study the association between suicide-related Twitter use and suicidal suicidal behavior. Based on the questionnaire results, they applied several supervised learning methods including linear regression, stepwise linear regression, decision trees, Lars-en and SVMs to classify suicidal behaviors.

2.2.2 Electronic Health Records

The increasing volume of electronic health records (EHRs) has paved the way for machine learning techniques for suicide attempter prediction. Patient records include demographical information and diagnosis-related history like admissions and emergency visits. However, due to the data characteristics such as sparsity, variable length of clinical series and heterogeneity of patient records, many challenges remain in modeling medical data for suicide attempt prediction. In addition, the recording procedures may change because of the change of healthcare policies and the update of diagnosis codes.

There are several works of predicting suicide risk based on EHRs [72,73]. Tran et al. [67] proposed an integrated suicide risk prediction framework with feature extraction scheme, risk classifiers and risk calibration procedure. Specifically, each patient's clinical history is represented as a temporal image. Iliou et al. [74] proposed a data preprocessing method to boost machine learning techniques for

[2] https://psychiatry.org/psychiatrists/practice/dsm
[3] https://apps.who.int/classifications/icd10/browse/2016/en

suicide tendency prediction of patients suffering from mental disorders. Nguyen et al. [75] explored real-world administrative data of mental health patients from hospital for short and medium-term suicide risk assessments. By introducing random forests, gradient boosting machines, and DNNs, the authors managed to deal with high dimensionality and redundancy issues of data. Although, previous method gained preliminary success, Iliou et al. [74] and Nguyen et al. [75] have a limitation on the source of data which focuses on patients with mental disorders in their historical records. Bhat and Goldman-Mellor [76] used an anonymized general EHR dataset to relax the restriction on patient's diagnosis-related history, and applied neural networks as classification model to predict suicide attempters.

2.2.3 Suicide Notes

Suicide notes are the written notes left by people before committing suicide. They are usually written on letters and online blogs, and recorded in audio or video. Suicide notes provide material for NLP research. Previous approaches have examined suicide notes using content analysis [52], sentiment analysis [44,77], and emotion detection [51]. Pestian et al. [52] used transcribed suicide notes with two groups of completers and elicitors from people who have a personality disorder or potential morbid thoughts. White and Mazlack [78] analyzed word frequencies in suicide notes using a fuzzy cognitive map to discern causality. Liakata et al. [51] employed machine learning classifiers to 600 suicide messages with varied length, different readability quality, and multi-class annotations.

Emotion in text provides sentimental cues of suicidal ideation understanding. Desmet et al. [79] conducted a fine-grained emotion detection on suicide notes of 2011 i2b2 task. Wicentowski and Sydes [80] used an ensemble of maximum entropy classification. Wang et al. [44] and Kovačević et al. [81] proposed hybrid machine learning and rule-based method for the i2b2 sentiment classification task in suicide notes.

In the age of cyberspace, more suicide notes are now written in the from of web blogs and can be identified as carrying the potential risk of suicide. Huang et al. [29] monitored online blogs from `MySpace.com` to identify at-risk bloggers. Schoene and Dethlefs [82] extracted linguistic and sentiment features to identity genuine suicide notes and comparison corpus.

2.2.4 Online User Content

The widespread of mobile Internet and social networking services facilitate people's expressing the own life events and feelings freely. As social websites provide an anonymous space for online discussion, an increasing number of people suffering from mental disorders turn to seek for help. There is a concerning tendency that potential suicide victims post their suicidal thoughts in social websites like Facebook, Twitter, Reddit, and MySpace. Social media platforms are becoming a promising tunnel for monitoring suicidal thoughts and preventing suicide attempts [83]. Massive user generated data provide a good source to study online users' language patterns. Using data mining techniques on social

networks and applying machine learning techniques provide an avenue to understand the intent within online posts, provide early warnings, and even relieve a persons suicidal intentions.

Twitter provides a good source for research on suicidality. O'Dea et al. [12] collected tweets using the public API and developed automatic suicide detection by applying logistic regression and SVM on TF-IDF features. Wang et al. [84] further improved the performance with effective feature engineering. Shepherd et al. [85] conducted psychology-based data analysis for contents that suggests suicidal tendencies in Twitter social networks. The authors used the data from an online conversation called #dearmentalhealthprofessionals.

Another famous platform Reddit is a online forum with topic-specific discussions, has also attracted much research interest for studying mental health issues [86] and suicidal ideation [38]. A community on Reddit called SuicideWatch is intensively used for studying suicidal intention [17, 87]. De Choudhury et al. [87] applied a statistical methodology to discover the transition from mental health issues to suicidality. Kumar et al. [88] examined the posting activity following the celebrity suicides, studied the effect of celebrity suicides on suicide related contents, and proposed a method to prevent the high-profile suicides.

Many researches [48, 49] work on detecting suicidal ideation in Chinese microblogs. Guan et al. [89] studied user profile and linguistic features for estimating suicide probability in Chinese microblogs. There also remains some work using other platforms for suicidal ideation detection. For example, Cash et al. [90] conducted a study on adolescents' comments and content analysis on MySpace. Steaming data provides a good source for user pattern analysis. Vioulès et al. [5] conducted user-centric and post-centric behavior analysis and applied a martingale framework to detect sudden emotional changes in Twitter data stream for monitoring suicide warning signs. Blog stream collected from public blog articles written by suicide victims is used by Ren et al. [14] to study the accumulated emotional information.

2.2.5 Summary

Applications of suicidal ideation detection mainly consist of four domains, i.e., questionnaires, electronic health records, suicide notes, and online user content. Table 2.2 gives a brief summary of categories, data sources, and methods. Among these four main domains, questionnaires and EHRs require self-report measurement or patient-clinician interactions, and rely highly on social workers or mental health professions. Suicide notes have a limitation on immediate prevention, as many suicide attempters commit suicide in a short time after they write suicide notes. However, they provide a good source for content analysis and the study of suicide factors. The last domain of online user content is one of the most promising way of early warning and suicide prevention when empowered with machine learning techniques. With the rapid development of digital technology, user generated content will play a more important role in suicidal ideation detection, and other forms of data such as health data generated by wearable devices can be very likely to help with suicide risk monitoring in the near future.

Table 2.2: Summary of Studies on Suicidal Ideation Detection

categories	self-report examination face-to-face suicide prevention automatic suicidal ideation detection
data	questionnaires suicide notes suicide blogs electronic health records online social texts
methods	clinical methods content analysis feature engineering deep learning

2.3 Summary

Suicide prevention remains an important task in our modern society. Early detection of suicidal ideation is an important and effective way to prevent suicide. This survey investigates existing methods for suicidal ideation detection from a broad perspective which covers clinical methods like patient-clinician interaction and medical signal sensing; textual content analysis such as lexicon-based filtering and word cloud visualization; feature engineering including tabular, textual, and affective features; and deep learning based representation learning like CNN- and LSTM-based text encoders. Four main domain-specific applications on questionnaires, EHRs, suicide notes and online user content are introduced.

Most work in this field has been conducted by psychological experts with statistical analysis, and computer scientists with feature engineering based machine learning and deep learning based representation learning. Based on current research, we summarized existing tasks and further propose new possible tasks. Last but not least, we discuss some limitations of current research and propose a series of future directions including utilizing emerging learning techniques, interpretable intention understanding, temporal detection, and proactive conversational intervention.

Chapter 3

Benchmarking for Suicidal Ideation Detection

3.1 Introduction

Suicide might be considered as one of the most serious social health problems in the modern society. Many factors can lead to suicide, e.g, personal issues, such as hopelessness, severe anxiety, schizophrenia, alcoholism or impulsivity; social factors, like social isolation, overexposure to deaths; or negative life events, including traumatic events, physical illness, affective disorders, and previous suicide attempts. Thousands of people around the world fall victims to suicide every year, making suicide prevention become a critical global public health mission.

Suicidal ideation or suicidal thoughts are people's thoughts of committing suicide. It can be regarded as a risk indicator of suicide. Suicidal thoughts include fleeting thoughts, extensive thoughts, detailed planning, role playing, incomplete attempts, and so forth. According to a WHO report [91], 788,000 people estimated worldwide committed suicide in 2015. And a large number of people, especially teenagers, were reported having suicidal ideation. Thus, one possible approach to preventing suicide effectively is early detection of suicidal ideation.

With the widespread emergence of mobile Internet technologies and online social networks, there is a growing tendency for people to talk about their suicide intentions in online communities. This online content could be helpful for detecting individuals' intentions and their suicidal ideation. Some people, especially adolescents, choose to post their suicidal thoughts in social networks, ask about how to commit suicide in online communities and enter into online suicide pacts. The anonymity of online communication also allows people to freely express the pressures and anxiety they suffer in the real world. This online user-generated content provides another possible angle for early suicide detection and prevention.

Previous research on suicide understanding and prevention mainly concentrates on its psychological and clinical aspects [9]. Recently, many studies have turned to natural language processing methods and classifying questionnaire results via supervised learning, which learns a mapping function from

labelled training data [92]. Some of these researches have used the "International Personal Examination Screening Questionnaire", and analysed suicide blogs and posts from social networking websites. However, these studies have their limitations. (1) from both a psychological and a clinical perspective, collecting data and/or patients is typically expensive, and some online data may help in understanding thoughts and behaviours (2) simple feature sets and classification models are not predictive enough to detect suicidal tendencies.

In this section, we investigate the problem of suicidal ideation detection in online social websites, with a focus on understanding and detecting the suicidal thoughts in online user content. We perform a thorough analysis of the content, the language preferences, and the topic descriptions to understand the suicidal thoughts from a data mining perspective. Six different sets of informative features were extracted and six supervised learning algorithms were compared to detect suicidal ideation within the data. It is a novel application of automatic suicide intention detection on social content with the combination of our proposed effective feature engineering and classification models.

This section makes notable contributions and novelties to the literature in the following respects:

1. **Knowledge Discovery.** This is a novel application of knowledge discovery and data mining to detect suicidal ideation in online users content. Previous work in this field has been conducted by psychological experts with statistical analysis; this approach reveals knowledge on suicidal ideation from a data analytics perspective. Insights from our analysis reveal that suicidal individuals often use personal pronouns to show their ego. They are more likely to use words expressing negativity, anxiety, and sadness in their dialogue. They are also more likely to choose the present tense to describe their suffering and the future tense to describe their hopelessness and plans for suicide.

2. **Dataset and Platform:** This section introduces the Reddit platform and collects a new dataset for suicidal ideation detection. Reddit's SuicideWatch BBS is a new online channel for people with suicidal ideation to express their anxiety and pressures. Social volunteers respond in positive, supportive ways to relieve the depression and hopefully prevent potential suicides. This data source is not only useful for suicide detection but also for studying how to effectively prevent suicide through effective online communication.

3. **Features, Models, and Bench-marking:** Rather than using basic models with simple features for suicidal ideation detection, this approach (1) identifies informative features from a number of perspectives, including statistical, syntactic, linguistic, word embedding features, and topic features; (2) compares with different classifiers from both traditional and deep learning perspectives, such as support vector machine [93], random forest [94], gradient boost classification tree (GBDT) [95], XGBoost [96], MLFFNN [43] and Long Short Term Memory (LSTM) [97]; and (3) provides benchmarks for suicidal ideation detection on SuicideWatch on Reddit, one active online forum for communication about suicide.

This section is organised as follows: We introduce the datasets in Section 3.2 along with data exploration and knowledge discovery. Section 3.3 describes classification and features extraction method. Section 3.4 is the experimental study. We conclude this section in 3.5.

3.2 Data and Knowledge

We collect the suicidal ideation texts from Reddit and Twitter, and manually check all the posts to ensure they were correctly labelled. Our annotation rules and examples of posts appear in Table 3.1.

Table 3.1: Annotation rules and examples of social texts

Categories	Rules	Examples
Suicide Text	• Expressing suicidal thoughts • Including potential suicidal actions	*I want to end my life tonight.* *Yesterday, I tried to cut my wrist, but failed.*
Non-suicide Text	• Formally discussing suicide • Referring to other's suicide • Not relevant to suicide	*The global suicide rate is increasing.* *I am so sad to hear that Robin Williams ended his life.* *I love this TV show and watch every week.*

3.2.1 Reddit Dataset

Reddit is a registered online community that aggregates social news and online discussions. It consists of many topic categories, and each area of interest within a topic is called a subreddit.

In this dataset, online user content include a title and a body of text. To preserve privacy, we replace personal information with a unique ID to identify each user. We collected posts with potential suicide intentions from a subreddit called "Suicide Watch"(SW)[1]. Posts without suicidal content were sourced from other popular subreddits [2] [3]. The collection of non-suicidal data are totally user generated content, and the posts of news aggregation and administrator are excluded. To facilitate the study and demonstration, we will study the balanced dataset in Reddit, and study imbalanced dataset in Twitter as following subsection.

The Reddit data set includes 3,549 suicidal ideation samples and a number of non-suicide texts. In particular, we construct two datasets for Reddit shown in Table 3.2. The first dataset includes two subreddits in which one is from suicideWatch and another is from popular posts in Reddit. The second dataset is composed of six subreddits include SuicdeWatch and another five hot topics: Gaming [4],

[1] https://www.reddit.com/r/SuicideWatch/
[2] https://www.reddit.com/r/all/
[3] https://www.reddit.com/r/popular/
[4] https://www.reddit.com/r/gaming/

Jokes [5], Books [6], Movies [7] and AskReddit [8]. In the second dataset, the combination of SuicideWatch with any other subreddit will be a new balanced sub-dataset, for example, Suicide vs. Gaming, and Suicide vs. Jokes. These two datasets will be studied on subsection 3.4.1 and 3.4.2 separately.

Table 3.2: Two Balanced Reddit Datasets.

Dataset	Subreddits
1	SuicideWatch vs. Others (Non-suicide)
2	SuicideWatch vs. Gaming SuicideWatch vs. Jokes SuicideWatch vs. Books SuicideWatch vs. Movies SuicideWatch vs. AskReddit

3.2.2 Twitter Dataset

Many online users also want to talk about the suicidal ideation in social networks. However, Twitter is quiet different with Reddit as 1) each tweet's length is limited in 140 characters [9], 2) Tweet users may have some social networks friends from real world while Reddit users are fully anonymous, 3) the communication and interaction type are totally different between social networking websites and online forums.

The Twitter Dataset is collected using a keyword filtering technique. Suicidal words and phrases include "suicide", "die", "end my life", and so forth. Many of collected Tweets has the suicidal related words, but they possibly talk about a suicide movie or advertisement which does not contain suicidal ideation. Therefore, we manually checked and labeled collected Tweets according to the annotation rules in Table 3.1. Finally, the Twitter dataset has totally 10,288 tweets with 594 tweets (around 6%) with suicidal ideation. This dataset is an imbalanced dataset, and will be studied in Section 3.4.3.

3.2.3 Data Exploration and Knowledge Discovering

To understand suicidal individuals, we analysed the words, languages, and topics in online user content.

Word-cloud. Word-clouds were used to provide a visual understanding of the data. The users' posts in Reddit and tweets in Twitter with potential suicide risk are showed separately in Figs. 3.1a and 3.1b. As we can see, suicidal posts frequently use words such as "life", "suicide", and "kill", providing a direct indication of the users' suicidal thoughts. Words expressing feelings or intentions are also frequently used, such as "feel", "want", and "know". For example, some suicidal posts wrote, "I feel like I have no one left and I want to end it", "I want to end my life", and "I don't know how much of it was psychological trauma".

[5] `https://www.reddit.com/r/Jokes/`
[6] `https://www.reddit.com/r/books/`
[7] `https://www.reddit.com/r/movies/`
[8] `https://www.reddit.com/r/AskReddit/`
[9] This limit is now 280 characters.

In addition, the dominant words in these two social platforms have different styles due to the posting rules of the platforms. The Reddit users are willing to compose their posts in a specific way. For instance, they describe their life events and their stories about their friends. While the content in Twitter is much more straightforward with expressions like “want kill”, “going kill” and “wanna kill”. The details are usually not included in their tweets.

Figure 3.1: Word cloud visualisation of suicidal texts in Reddit and Twitter

(a) Reddit word cloud

(b) Twitter word cloud

Language Preferences. Language preferences provide an overview of the statistical linguistic information of the data. The listed variables shown in Table 3.3 were extracted using LIWC 2015 [98]. All these categories are features based on word counts. We calculated the average value of each variable in both suicide-related texts and suicide-free posts. As shown in the table, content with or without suicidality quite differs in many items.

- Users with suicidal ideation use many personal pronouns to show their ego. For example, “I want to end my life.”
- They express more negative emotions, like anxiety and sadness. For example, “I was drowning in guilt and depression for several years after.”
- As for the tense, texts with suicidal ideation tend to use the present and future tense. They tend to use the present tense to describe their suffering pain and depression. For example, “I’m feeling so bad.” The future tense is used to describe their hopeless feelings about the future and their suicide intentions. For example, “I’m eventually going to kill myself.”
- Both types of posts discuss family, friends and make female or male references.
- Unsurprisingly, more words related to death appear in texts about suicide. For example, “kill”, “die”, “end life”, and “suicide”.
- Both types of posts contain a similar number of swear words.

One of findings from Table 3.3 and Fig. 3.1 is that people with suicidal thoughts tend to directly show their intentions in anonymous online communities when faced with some kinds of problem in the real world. Their posts often show negative feelings with strong ego and intention.

Table 3.3: Linguistic statistical information extracted by LIWC

Average Word Count	Suicide	Non-suicide
personal nouns	30.01	14.6
quantifiers	3.78	3.37
positive emotion	5.61	7.84
negative emotion	11.12	4.89
anxiety	1.46	0.55
sadness	3.86	0.63
past focus	6.78	6.27
present focus	34.81	17.86
future focus	4.06	1.76
family	1.07	0.82
friend	1.02	0.78
female references	0.95	1.35
male references	1.03	2.40
work	2.50	3.92
money	0.60	1.38
death	4.81	0.61
swear words	1.47	1.62

Table 3.4: Topics words extracted from posts containing suicidal thoughts

No.	Top 10 words for each suicide related topics in SuicideWatch
1	money, working, suicide, gun, fucked, come, yet, failed, erase, thats
2	said, got, went, started, friend, back, father, told, mother, girl
3	im, school, go, year, time, know, one, ive, day, got
4	im, dont, its, ive, cant, get, know, around, time, pain
5	im, feel, like, want, know, friend, would, life, get, time
6	imagine, cellophane, abandoned, anyone, medical, cheated, mr, surgery, yelling, letter
7	im, want, life, like, get, feel, ive, know, year, even
8	fucking, very, tomorrow, bottom, accept, sharp, n't, went, wife, attacked
9	condition, suicide, also, hope, tx, california, chronic, jumping, crisis, age
10	please, find, mother, car, social, live, need, accident, debt, month

Topic Description. We extracted 10 topics from posts containing suicidal ideation using the Latent Dirichlet allocation (LDA) [99] topic modelling method, as shown in Table 3.4. There are some Internet slangs such as "tx"(thanks) and abbreviations like "im"(I'm) and "n't" ("negatory"). In the field of standard natural language processing, personal words like "I", "me" and "you" are stop words and should be removed, but we kept them in this exploration because they contain important information. Thus, there are many personal pronouns included in these topics words, which are identical to the results in Table 3.3.

Interestingly, we observed that posts containing suicidal themes could be summarised into three categories: internal factors, external social factors and mixed internal/external factors. Specifically, internal factors, including words like "know" (Topics 3, 4, 5 and 7), "want", "feel", and "like"(Topics 5 and 7), and "hope"(Topic 9) express people's feelings, intentions, and desires. While other words such as "money" and "working"(Topic 1), "friend"(Topics 2 and 5), "school"(Topic 3), "surgery"(Topic 6),

"crisis"(Topic 9), and "accident"(Topic 10) indicate that posts are linked to social factors. In Topic 3, 5, 9 and 10 both factors are represented.

3.3 Methods and Technical Solutions

3.3.1 Features Processing

By preprocessing and cleaning the data in advance, we extracted several features including statistics, words-based features (e.g., suicidal words, pronouns and etc.), TF-IDF, semantics and syntactics. Additionally, we used distributed features by training neural networks to embed word into vector representations, along with topic features extracted by LDA [99] as unsupervised features.

Statistical Features. User-generated posts are varied in length, and some statistical features can be extracted from texts. Some posts use short and simple sentences, while others use complex sentences and long paragraphs.

After segmentation and tokenization, we captured statistical features as follows:

- the number of words, tokens, and characters in the title
- the number of words, tokens, characters, sentences, and paragraphs in the text body

Syntactic Features: POS. Syntactic features are useful information in natural language processing tasks. We extracted parts of speech (POS) [100] as features for our suicidal ideation detection model to capture the similar grammatical properties in users' posts.

Common POS tags include nouns, verbs, participles, articles, pronouns, adverbs and conjunctions. POS subgroups were also identified to provide more detail about the grammatical properties of the posts. Each post was parsed and tagged, and the number of each category in the title and text body were simply counted.

Linguistic Features: LIWC. Online users' posts usually contain emotions, relativity and harassment words. Lexicons are widely applied for extracting these features. To analyse the linguistic and emotional features in the data, we used Linguistic Inquiry and Word Count [98] (LIWC 2015 [10]) which was proposed and developed by the University of Texas at Austin. This approach was used in previous study [38]. The tool contains a powerful internally-built dictionary for matching the target words in posts when parsing data. About 90 variables were output. In addition to word count based features, it could extract features based on emotional tone, cognitive processes, perceptual processes and many types of abusive words. Specific categories include word count, summary language, general descriptors, linguistic dimensions, psychological constructs, personal concern, informal language markers, and punctuation.

Words Frequency Features: TF-IDF. Many kinds of expression are related to suicide. We used TF-IDF to extract these features and measure the importance of various words from both suicidal posts

[10] `http://liwc.wpengine.com/`

and non-suicidal posts. TF-IDF measures the number of times that each word occurs in the documents, and adds a penalty depending on the frequency of the word in the entire corpus. The processing to calculate TF-IDF as follow: Given term and document denoted as t and d respectively, First, calculate the term frequency by:

$$TF(t,d) = \frac{n_{t,d}}{\sum_{t' \in d} n_{t',d}}$$

where $n_{t,d}$ denotes the count of a term in a document. Second, calculate the inverse document

$$IDF(t,D) = \log \frac{N}{f_t}$$

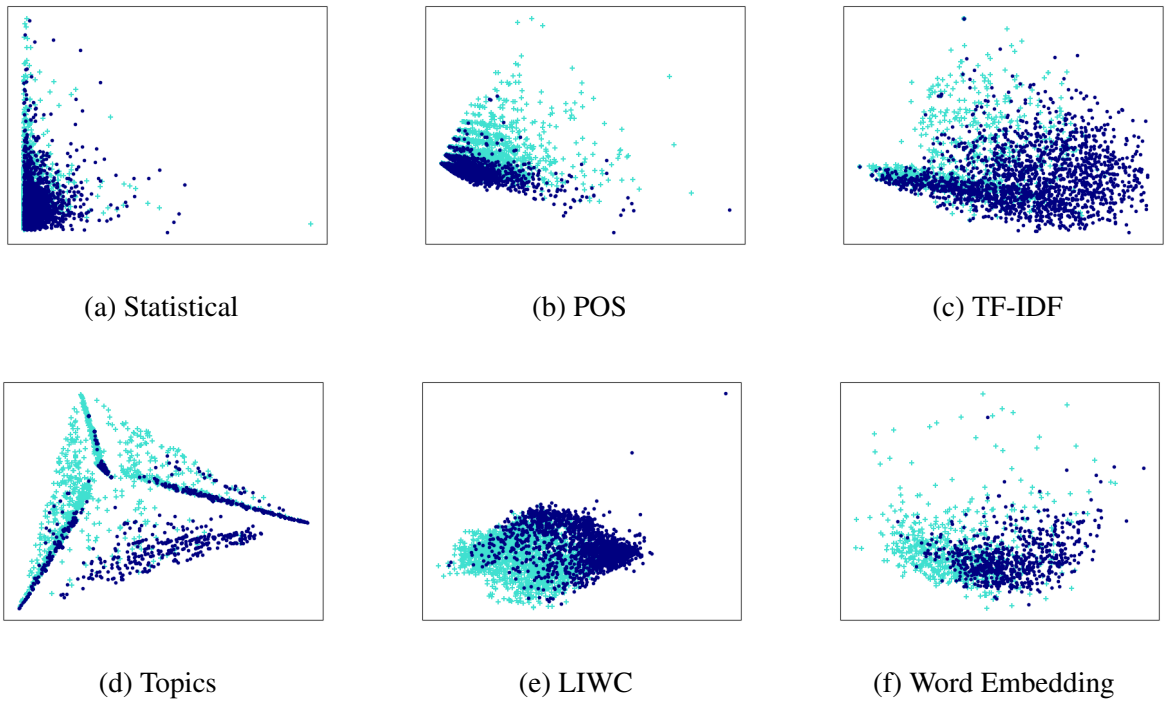
where D is the corpus, N is the total number of documents, and f_t is the number of documents which Third, we get the TF-IDF as:

$$TF\text{-}IDF(t,d,D) = TF(t,d) \times IDF(t,D)$$

Word Embedding Features. The distributed representation, which is able to preserve the semantic information in texts, is popular and useful for many natural language processing tasks. It embeds words into a vector space. There are several techniques for word embedding. We employed the *word2vec* [101] [11] to derive a distributed semantic representation of the words.

There are two architectures for *word2vec* word embedding, i.e., CBOW and Skip-gram. CBOW predicts the present word based on the context, Skip-gram predicts the closest words to the current word provided.

Figure 3.2: Visualisation of extracted features using PCA



(a) Statistical (b) POS (c) TF-IDF

(d) Topics (e) LIWC (f) Word Embedding

[11] https://code.google.com/archive/p/word2vec/

Topics Features. Suicidal posts and non-suicidal posts talk about different topics which can provide good understanding for two categories. We applied the Latent Dirichlet Allocation (LDA) [99] to reveal latent topics in user posts. Each topic is a mixtures probability of words occurrence in the topic, and each posts is a mixtures probability of topics.

Given the set of documents and the number of topics, we used LDA to extract the topics from each posts, then calculate the probability that each post belonged to every generated topics. Hence, the posts are represented by their thematic properties as probability vectors at the length of the number of topics.

Feature Visualisation. To understand the informativeness of these feature sets, we visualise the features on the Reddit dataset in a 2-dimensional space by using Principle Component Analysis (PCA) [102] in Fig. 3.2. The results demonstrate that we indeed extract features that can largely separate the points in different classes. We will further validate the effectiveness of our feature sets in Section 5.

3.3.2 Classification Models

Suicidality detection in social content is a typical classification problem of supervised learning as in Definition 2.

Definition 2 (Suicidal Ideation Detection on Text)**.** Given a training set of m documents with a specific class $C = \{c_1, c_2, \ldots, c_n\}$,, suicidal ideation detection on text trains a classifier using the training set, and learns a learned classifier for prediction on testing set.

Given a dataset $\{x_i, y_i\}_i^n$ consisting a set of texts $\{x_i\}_i^n$ with labels $\{y_i\}_i^n$, we trained a supervised classification model to learn the function from the training data pairs of input objects and supervisory signals:

$$y_i = F(x_i)$$

where $y_i = 1$ means that the expression x_i is "suicide text" (ST), otherwise $y_i = 0$ means "not suicide text (non-ST)". The training or learning of the classification model is to minimise the prediction error in the given training data. The prediction error is to be presented as a loss function $L(y, F(x))$ in where y is the real label and $F(x)$ is the predicted label by using classification model. In summary, the goal of training algorithm is to obtain an optimal prediction model $F(x)$ by solving below optimisation task:

$$\hat{F} = \arg\min_F \mathbb{E}_{x,y}[L(y, F(x))]$$

Different classification method may have different definition of loss function and pre-defined structure of model. We employed both classical supervised learning classification methods and deep learning methods to solve the suicidal ideation classification task.

The structure of our feature extraction method is shown in Fig. 3.3. As mentioned in Section 3.3.1, features comprised statistics, POS counts, LIWC features, TF-IDF vectors and topics probability features. Among these features, we applied POS features and LIWC features to both the title and text body of user posts. We combined the title and the body into one pieces of text to extract topics probability vectors and TF-IDF vectors. All extracted features were input to the classifiers.

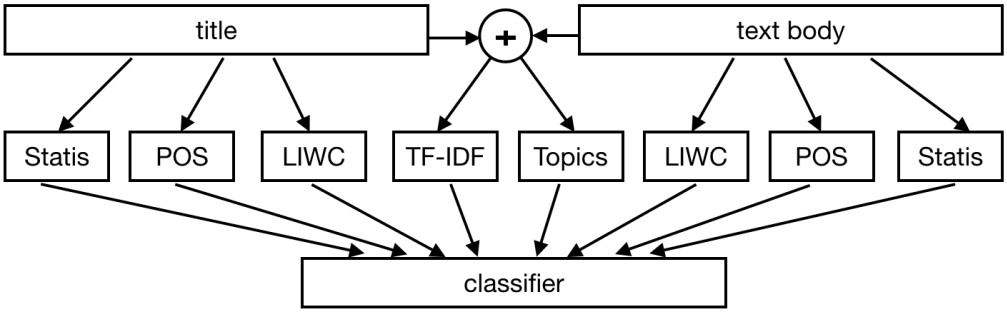


Figure 3.3: The model's structure for Reddit dataset

3.4 Empirical Evaluation

3.4.1 Comparison and Analysis on Suicide vs. Non-suicide

Table 3.5: Comparison of different methods using different features

Methods	Features	Acc.	Prec.	Recall	F1-score	AUC
SVM	Statis	0.8064	0.8045	0.8189	0.8116	0.8061
	Statis+Topic	0.8609	0.881	0.8406	0.8603	0.8613
	Statis+Topic+TF-IDF	0.8571	0.8414	0.8865	0.8634	0.8565
	Statis+Topic+TF-IDF+POS	0.8674	0.8545	0.8916	0.8727	0.8670
	Statis+Topic+TF-IDF+POS+LIWC	0.9123	0.9144	0.9133	0.9138	0.9123
Random Forest	Statis	0.7732	0.8094	0.7258	0.7653	0.7741
	Statis+Topic	0.8973	0.8922	0.9082	0.9001	0.8971
	Statis+Topic+TF-IDF	0.8915	0.8795	0.912	0.8954	0.8911
	Statis+Topic+TF-IDF+POS	0.8986	0.8801	0.9273	0.9031	0.8981
	Statis+Topic+TF-IDF+POS+LIWC	0.9357	0.9213	0.9554	0.938	0.9353
GBDT	Statis	0.7505	0.7632	0.7398	0.7513	0.7507
	Statis+Topic	0.898	0.8856	0.9184	0.9017	0.8976
	Stati+Topics+TF-IDF	0.896	0.89	0.9082	0.899	0.8958
	Statis+Topic+TF-IDF+POS	0.8928	0.8893	0.9018	0.8955	0.8926
	Statis+Topic+TF-IDF+POS+LIWC	0.9461	0.9354	0.9605	0.9478	0.9458
XGBoost	Statis	0.7667	0.7822	0.7513	0.7664	0.7670
	Statis+Topic	0.8999	0.8938	0.912	0.9028	0.8997
	Statis+Topic+TF-IDF	0.9019	0.8941	0.9158	0.9049	0.9016
	Statis+Topic+TF-IDF+POS	0.9103	0.8998	0.9273	0.9133	0.9100
	Statis+Topic+TF-IDF+POS+LIWC	**0.9571**	0.9499	**0.9668**	**0.9583**	**0.9569**
MLFFNN	Statis	0.7647	0.7742	0.7742	0.7742	0.7731
	Statis+Topic	0.8821	0.8740	0.8525	0.8631	0.8961
	Statis+Topic+TF-IDF	0.8606	0.8369	0.8401	0.8385	0.8855
	Statis+Topic+TF-IDF+POS	0.9068	0.9038	0.8868	0.8952	0.9369
	Statis+Topic+TF-IDF+POS+LIWC	0.9283	0.9391	0.9205	0.9295	0.9403
LSTM	*word2vec* word embedding	0.9266	**0.9786**	0.8750	0.9239	0.9276

This section compares various classification methods using different combinations of features with 10-fold cross validation [12]. The specific classification models include Support Vector Machine [93], Random Forest [94], Gradient Boost Classification Tree (GBDT) [95], XGBoost [96] and Multilayer

[12]Our codes are available in `https://github.com/shaoxiongji/sw-detection`

Feed Forward Neural Net (MLFFNN) [43]. SVM is able to solve problems that are not linearly separable in lower space by constructing a hyperplane in high-dimensional space. It can be adapted to many kinds of classification tasks. Random Forest, GBDT, and XGBoost are tree ensemble methods that use decision trees as base classifiers and producing a form of committee to gain better performance than any single base classifier. MLFFNN takes the different features as input and learns the combination of them with nonlinearity.

For comparison and to solve the problem of understanding the semantic meaning and syntactic structure of sentences, deep learning provides powerful performance. We used Long Short Term Memory (LSTM) [97] network, one state-of-the-art deep neural network. LSTM takes title and text body of user posts with word embedding as its inputs, and uses memory cell to preserve the state over long periods, capturing the long-term dependencies in long conversations detection.

As shown in Table 3.5, all methods' performance increase by combining more features on the whole. This observation validates the effectiveness and informativeness of our extracted features. However, the contribution each feature makes varies, which leads to fluctuations in the results of individual methods. The XGBoost had the best performance of the these six methods when taking all groups of features as inputs. Although LSTM does not require features processing and is renowned for its state-of-the-art performance in many other natural language processing tasks, it did not perform as well as some of the other ensemble learning methods with sufficient features in this case. Random Forest, GBDT, XGBoost, and MLFFNN with proper features produced better accuracy and F1-scores than LSTM on our Reddit dataset. Admittedly, deep learning with word embedding is rather convenient and typically achieves adequate results, even without complicated features engineering.

The AUC performance measurement in each classification is the area under the receiver operating characteristic curve with all extracted features. In the last column of Table 3.5, the AUC has an increasing tendency with more combined features. The XGBoost method gains the highest AUC of 0.9569 while other methods has very similar AUC value above 0.9.

3.4.2 Suicide Vs. Single Subreddits Topics

To evaluate the classification on suicide with any other specific online communities, we extended our datasets and experiments to other specific subreddits, including “gaming”, “Jokes”, “books”, “movies” and “AskReddit”.

The results are shown in Figure 3.4. Using the features extracted with our approach was very effective way of classifying the suicidal ideation posts from another subreddits domains. In fact, the classification results on suicidal dataset vs the subreddit data set were better than suicidal vs non-suicidal data set in where the non-suicidal samples are composed of multiple popular subreddit domains. In these experiments, XGBoost produced the best results on “movies” and “AskReddit” in terms of accuracy and F1-scores. LSTM and Random Forest outperformed the other models in “gaming” and “books” respectively.

3.4.3 Experiments on Twitter Dataset

To evaluate the performance of our proceeded features and the classification models, we do another experiment on our Twitter dataset. Tweets text without long text body is different with Reddit text. Thus, for the experimental setting, there is a slight difference between them. We exclude the number of paragraph in statistical features, POS and LIWC features of text bodies. The rest settings are similar to our previous experiment. Considering the class imbalance in Twitter data, we adopt under-sampling techniques. The results are the average metrics of each under-sampled data shown in Table 3.6. The receiver operating characteristic curves of these methods are showed in Fig. 3.5. In these dataset, random forest gains better performance than most models except for the metric of precision in which the MLFFNN gains slightly better result.

Table 3.6: Comparison of different models using all processed features on Twitter data

Model	Acc.	Prec.	Recall	F1	AUC
Random Forest	**0.9638**	0.9638	**0.9917**	**0.9646**	**0.9862**
GBDT	0.9500	0.9413	0.9603	0.9503	0.9825
XGBoost	0.9591	0.9425	0.9782	0.9597	0.9843
SVM	0.9485	0.9261	0.9755	0.9497	0.9813
MLFFNN	0.9412	**0.9661**	0.9194	0.9421	0.9823
LSTM	0.9108	0.9399	0.8802	0.9059	0.9747

3.5 Summary

The amount of text keeps growing with the popularisation of social networking services. And suicide prevention remains an important task in our modern society. It is therefore essential to develop new methods to detect online texts containing suicidal ideation in the hope that suicide can be prevented.

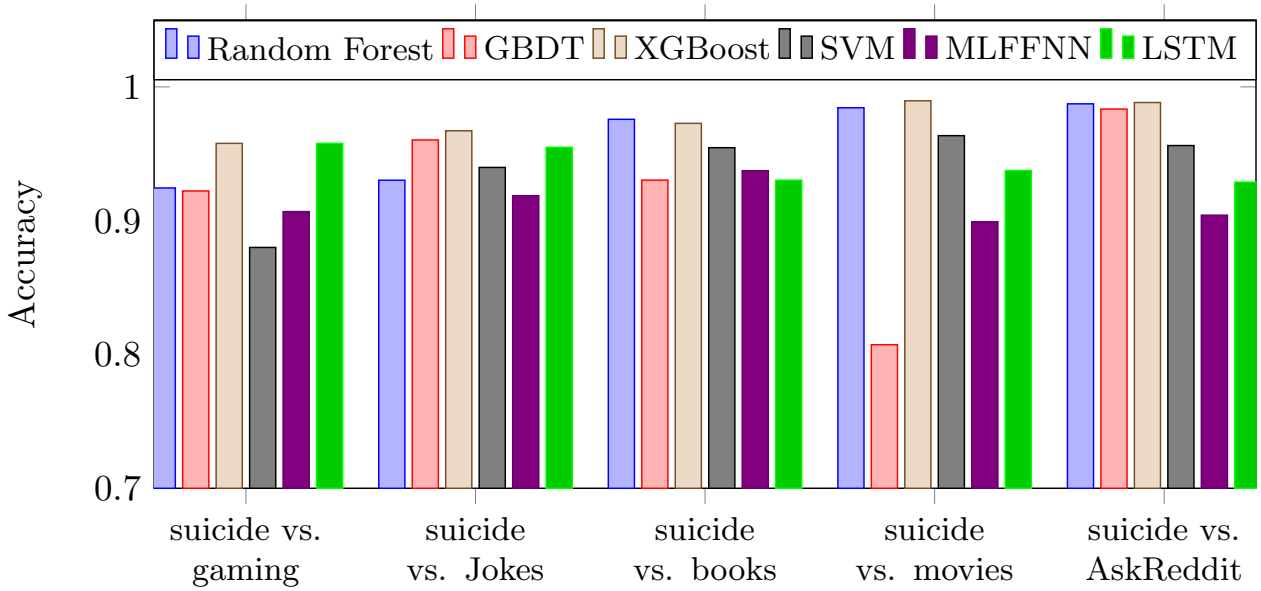


Figure 3.4: Classification for suicidal ideation of SuicideWatch vs. other six subreddits

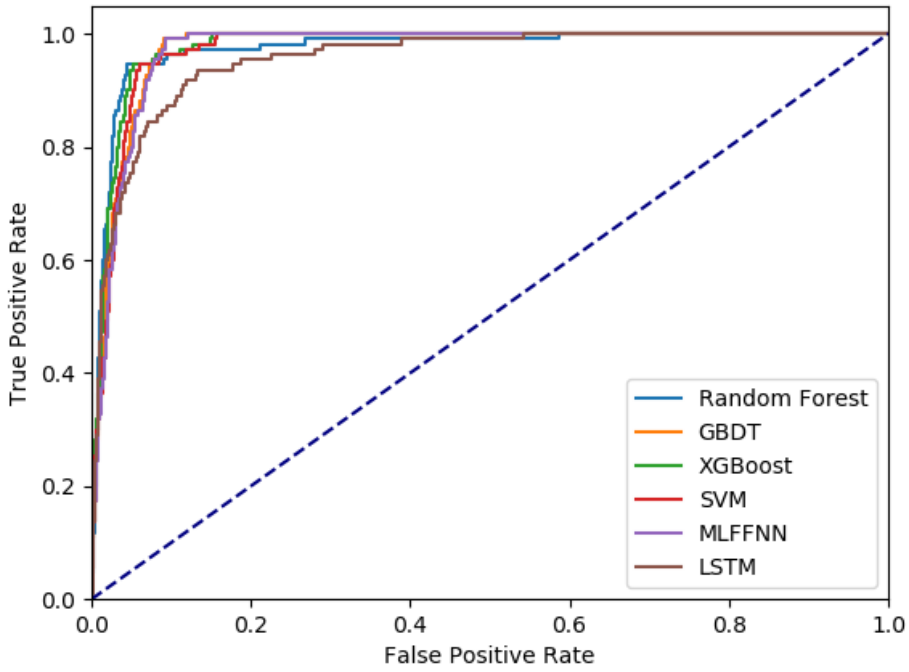


Figure 3.5: The receiver operating characteristic curve of six methods with all processed features

In this section, we investigated the problem of suicidality detection in online user-generated content. We argue that most work in this field has conducted by psychological experts with statistical analysis, which is limited by the cost and privacy issue in obtaining data. By collecting and analysing the anonymous online data from an active Reddit platform and Twitter, we provide rich knowledge that can complement the understanding of suicidal ideation and behaviour. Though applying feature processing and classification methods to our carefully built datasets - Reddit and Twitter, we evaluated, analysed and demonstrated that the developed framework can achieve high performance (accuracy) in distinguishing suicidal thoughts out of normal posts in online user content.

While exploiting more effective feature sets, complex models, or other factors such as temporal information may improve the detection of suicidal ideation - these will be our future directions, the contribution and impact of this section are threefold: (1) delivering rich knowledge in understanding suicidal ideation, (2) introducing datasets for the research community to study this significant problem, and (3) proposing informative features and effective models for suicidal ideation detection.

Chapter 4

Attentive Relation Network

4.1 Introduction

Mental health is a global issue, especially severe in most developed countries and many emerging markets. According to a WHO report[1], 1 in 4 people worldwide suffer from mental disorder to some extent. And 3 out of 4 people with severe mental disorders do not receive treatment, which makes the problem worse. Figure 4.1 shows the prevalence of mental and substance use disorders in 2016[2]. Partly due to severe mental disorders, 900,000 persons commit suicide each year all over the world, making suicide the second most common cause of death among the young. Suicide attempters are also reported as suffering from mental disorders. The US National Alliance on Mental Illness reported that 46% of suicide victims have experienced mental health conditions[3].

With the advance of social network services, people begin to express their feelings in the forums and seek for online support. Regular ways of prevention include oral conversation based consultation and psychological intervention. However, due to the scarcity and inequality of public resources in health services [103], many victims could not get effective treatments even through some of them are suffering from severe mental disorders. Transferring from suicidal ideation to action is a long-term process. Gilat et al. [104] scaled suicide risks into four levels, i.e., non-suicidal, suicidal thoughts or wishes, suicidal intentions, and suicidal act or plan. Before suicidal ideation, victims may suffer from different kinds of other mental disorders. According to meta-analyses underlying mental disorders can lead to suicide, especially in high-income countries with a figure of 90%[4]. The social networking service has become one of most useful tools to provide support and feedback for people with mental heath issues [85]. To provide effective suicide early prevention given limited support resources, it is necessary to triage the risk levels automatically and provide conversational support accordingly to

[1]Mental health action plan 2013 - 2020, available in http://www.who.int/mental_health/action_plan_2013/mhap_brochure.pdf?ua=1

[2]Published by Seattle, United States: Institute for Health Metrics and Evaluation (IHME), 2017. Available at http://ghdx.healthdata.org/gbd-results-tool. Downloaded from https://ourworldindata.org/mental-health

[3]NAMI report on Risk Of Suicide, available at https://www.nami.org/Learn-More/Mental-Health-Conditions/Related-Conditions/Suicide

[4]Report by Hannah Ritchie and Max Roser published online at OurWorldInData.org. Retrieved from https://ourworldindata.org/mental-health

relieve victims' issues. Our motivation is to use deep learning techniques to enable early detection and identify people's risk levels, which can help the social workers or experts to have a prior understanding of people's situation when trying to relieve their mental health issues. The automatic detection technique can be applied to mental health monitoring and help to facilitate online support.

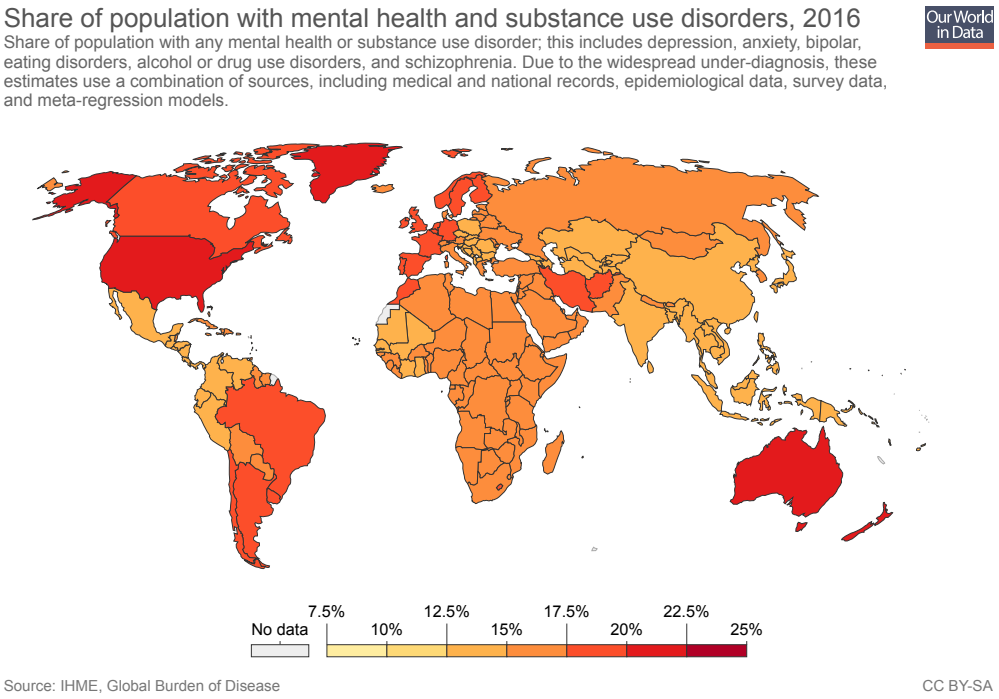


Figure 4.1: Global Burden of Disease Collaborative Network. Global Burden of Disease Study 2016 (GBD 2016) Results

To identify the risk levels of suicidal ideation and the types of mental disorders that the online users or their relations may suffer from is typically a text classification problem. There are many types of mental disorders according to two main diagnostic schemes for identifying mental disorders, i.e., Diagnostic and Statistical Manual of Mental Disorders (DSM-5)[5] and Chapter V Mental and Behavioral Disorders of International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10)[6]. Suicidal ideation and mental disorders like depression, anxiety and bipolar in online social content share quite similar patterns including the language usage, topic distribution, and sentimental polarity. Most of them contains a lot of negative expression. The popular topic among those posts are quite common, including job stress, family issues, and personal crisis. Thus, classifying suicidal ideation and other mental health issues requires attention to have a good understanding on the subtle differences among those characteristics.

Suicide and mental health issues could be categorized as different levels which could be taken as multi-class classification problem. There is an explosion of recent works on text classification using deep neural networks. But classifying mental health and suicidal ideation is a more specific task which requires to focus on potential victims' language usage. Noticing that people's posting

[5] https://www.psychiatry.org/psychiatrists/practice/dsm
[6] http://apps.who.int/classifications/icd10/browse/2016/en#/V

showing feelings or expressing sufferings contains their sentiment to some extend, we propose to capture these useful and important information to learn richer sentence representation and reason over risk actions and people mental or social state. In this section, we develop several popular deep learning models for text classification on some existing datasets and self collected datasets in real-world social networking websites, and propose a new enhanced relation reasoning model to provide a more accurate classification on suicide risk levels and suicidal ideation vs. other mental disorders such depression and anxiety.

Our contributions could be summarized as:

- This section focuses on identifying suicidal ideation and different kinds of mental disorders for early warning. Specifically, we consider user-level and post-level detection.

- To improve the performance of risk identification, we propose the relation network model with attention for reasoning over text representation and two sets of risk indicators, i.e., lexicon-based sentimental state and latent topics within posts.

- Experiments on public datasets and our own collected dataset show that our proposed method can improve the predictive performance.

This section is organized as follows. Related work on mental disorders, suicidal ideation, text classification, and relational reasoning are reviewed in Section 4.2. In Section 4.3, we introduce the proposed method that introduces sentimental lexicon and topic model into reasoning with attention-based relation network. Datasets are introduced in Section 4.4, together with a simple exploratory analysis. Experimental settings and results are presented in Section 4.5. In Section 4.6, we make a conclusion and have a brief outlook for future work.

4.2 Related Work

4.2.1 Text Classification

Text classification has experienced a rapid development with the development of deep neural networks. Distributed word representation techniques such as word2vec [54] and GloVe [55] provide powerful tools for text representation. Kim [105] proposed convolutional neural networks for sentence classification. To capture the long term dependency in sentences, the long short-term memory (LSTM) [106] were applied. Lai et. al [107] proposed recurrent convolutional neural networks combining two popular neural network architectures for text classification. Attention mechanism [108] is also widely used in text classification. Yang et al. [109] proposed hierarchical attention networks using attention mechanism in word and sentence level. Lin et al. [110] proposed self-attention to learn structured sentence embedding.

4.2.2 Relational Reasoning

Relational reasoning with relation networks (RNs) is originally utilized for scene object discovery by exploiting relations among objects [111]. RNs are further introduced to relational reasoning for visual question answering by calculating the relation score of the feature maps of object pairs and question representation [112]. In the community of knowledge base representation learning, relational reasoning between subjects and objects in knowledge bases is also studied [113]. As for our application scenario of suicidal ideation detection, it is critical to understand the relation between suicidality and risk indicators such as individual's sentiment and life events.

4.3 Methods

4.3.1 Problem Definition

Detecting suicidal ideation and mental disorders in social content is technically a domain-specific task of text classification. In this section, we conduct fine-grained suicide risk assessment and classification of multiple mental health issues, which are naturally regarded as multi-class classification. For fine-grained suicide risk, the risk levels include none, low, moderate, and severe risk, while for mental health classification, specific mental disorders are depression, anxiety, bipolar, and so on. And there are two subtasks for specific settings of data in social content, i.e., post-level classification and user-level classification. The former one takes single post p as input, while the latter one detects the suicide attempter with multiple posts $P = \{p_1, p_2, \ldots, p_n\}$.

4.3.2 Model Architecture

The proposed model consists of two steps, i.e., post representation and relation reasoning module as illustrated in Fig. 4.2. The post representation includes two parts of extraction of risk-related state indicators and LSTM text encoder. The relation module as shown in the dashed box of Fig. 4.2 utilizes the vanilla relation network for reasoning the connection between state indicators and user's posts, and attention mechanism for prioritizing more important relation scores of relational reasoning.

4.3.3 Text Encoding and Risk Indicators

User post sequence is embedded into word vectors of $p = \{w_1, w_2, \ldots, w_n\} \in \mathbb{R}^{l \times d}$. We apply bidirectional LSTM in Eq. 4.1 for text encoding to capture the adjacent dependency of words.

$$\begin{aligned} \overrightarrow{h_t} &= \overrightarrow{\text{LSTMcell}}\left(w_t, \overrightarrow{h_{t-1}}\right) \\ \overleftarrow{h_t} &= \overleftarrow{\text{LSTMcell}}\left(w_t, \overleftarrow{h_{t+1}}\right) \end{aligned} \tag{4.1}$$

The hidden state is obtained by concatenating each direction as $h_t = \text{concat}(\overrightarrow{h_t}, \overleftarrow{h_t})$, where $h_t \in \mathbb{R}^{l \times 2n}$ given n as the number of hidden units.

Sentimental information plays an important role when people expressing their sufferings and feelings in online social networks. To measure the sentiment, we take sentiment lexicons as additional information, specifically, domain-specific sentiment lexicons [114] from communities in Reddit are used. The sentiment lexicons are induced by seed words with domain-specific word embedding and a label propagation framework. The extracted sentiment information of post denoted as $s \in \mathbb{R}^l$ acts as a state indicator representing post authors internal sentimental state. Correspondingly, external indicators such as people's life events reveal another dimension as the risk indicator. To capture external factors of suicidal ideation or mental disorders, we introduce topic model to learn unsupervised topical features. Specifically, Latent Dirichlet Allocation (LDA) [115] is applied to extract latent topics in social posts to represent people's sufferings such as life events, social exposure and other experience in real world. The probability score vectors of posts belonging to all extracted topics are represented as $v \in \mathbb{R}^m$, where m is the number of topics.

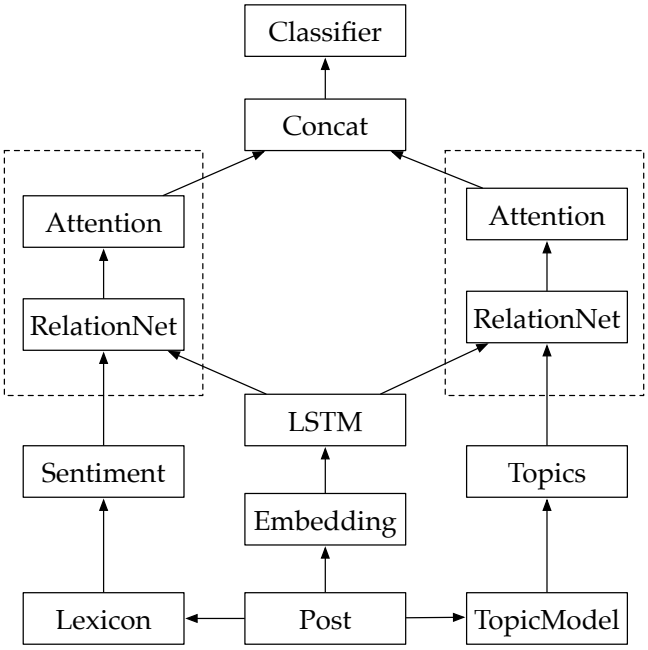


Figure 4.2: The architecture of the proposed model

4.3.4 Relation Network with Attention

The relation network [112] is a neural module for relational reasoning. It is originally proposed to capture the relation between objects. Given objects of $\mathcal{O} = \{o_1, o_2, \ldots, o_n\}$ and functions of f_ϕ and g_θ, a relation network is defined in Eq. 4.2. The output of g_θ is called the learned "relation", while the f_ϕ function acts as the classifier.

$$\mathrm{RN}(\mathcal{O}) = f_\phi \left(\sum_{i,j} g_\theta \left(o_i, o_j \right) \right) \tag{4.2}$$

Our aim is to reason risk factors of suicidal ideation and mental disorders. Thus, we take text encoding and state indicators as the input of relation networks to calculate relation scores between each token in posts and state indicators modeled by sentiment and topic features. To enhance the

relational reasoning, the attention mechanism is incorporated with relation module by assigning attention weights to the learnt relations. The idea of attentive relation network is shown in Fig. 4.3. The text representation is encoded by an LSTM network which captures the sequential independency, and then concatenated with the expanded state indicators. Here, we consider two indicators of sentiment and topic features, with the expanded representations denoted as $S = [s, s, \ldots, s] \in \mathbb{R}^{l \times l}$ and $V = [v, v, \ldots, v] \in \mathbb{R}^{l \times m}$ respectively. Then, they are inputed into relation networks to calculate relation vector $r_i \in \mathbb{R}^k$ with a multiple layer perception (MLP) as in Eq. 4.3 for the sentiment indicator.

$$r_i = \text{MLP}(h_i, s_i) \tag{4.3}$$

The attention is calculated as follows:

$$\alpha = \text{softmax}\left([r_1, r_2, \ldots, r_l] W^T + b\right), \tag{4.4}$$

where $W \in \mathbb{R}^{1 \times k}$, $b \in \mathbb{R}^l$ and $\alpha \in \mathbb{R}^l$. By element-wise product, the attentive representation of learnt relations can be calculated as

$$\tilde{r} = \alpha \otimes [r_1, r_2, \ldots, r_i] \tag{4.5}$$

where $\tilde{r} \in \mathbb{R}^{l \times k}$. By applying element-wise sum over $\tilde{r}$, we get the final attentive relational representation.

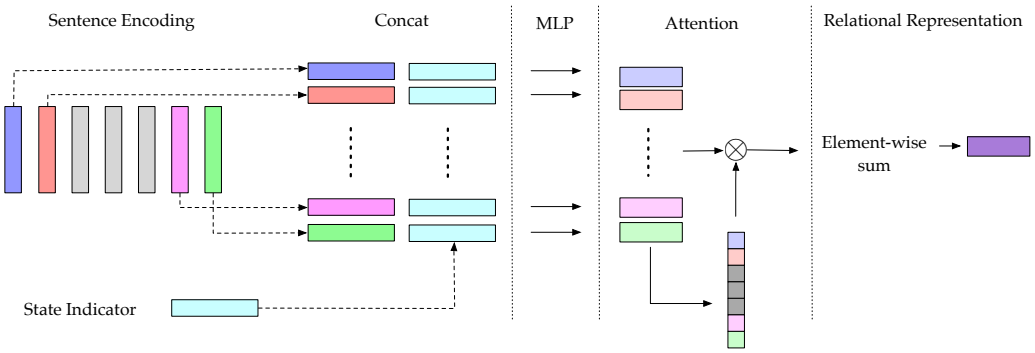


Figure 4.3: Relation network with attention

4.3.5 Classification

The last step is to use the learnt representation, which contain sequential information and relation of risk indicators, for classification. Specifically, we concatenate relational representations of $e = [\tilde{r}_s, \tilde{r}_v]$ from two channels as shown in Fig. 4.2, and use the fully connected layer with non-linear activation function of $f(\cdot)$ to produce the logits for prediction as follows.

$$\begin{aligned} l &= \text{f}\left(W_l e + b_l\right) \\ \mathcal{P} &= \text{softmax}\left(W_o l + b_o\right) \end{aligned} \tag{4.6}$$

where $W_l \in \mathbb{R}^{d_l \times d_e}, b_l \in \mathbb{R}^{d_l}, W_o \in \mathbb{R}^{c \times d_l}, b_o \in \mathbb{R}^c, \mathcal{P} \in \mathbb{R}^c$ For multi-class classification, the predicted label is produced by

$$\hat{y} = \underset{i}{\text{argmax}}\left(\mathcal{P}_i\right) \tag{4.7}$$

4.3.6 Training

There are two training phases of our proposed model, i.e., training the LDA topic model and the classification model. For the topic model, LDA assumes a generative process of documents as random mixtures over latent topics, and a topic can be inferred as a distribution over the words, where the Bayesian inference is used for learning various distributions. In practice, we use the Gensim library[7] to build the topic model during implementation.

For the ultimate target of suicide ideation and mental health detection, we use the cross entropy loss with L2 regularization as follows.

$$L = -\frac{1}{\sum_{s=1}^{N} c(s)} \sum_{i=1}^{N} \sum_{j=1}^{c(i)} \log \mathcal{P}_{i,j}\left[y_{i,j}\right] + \lambda \|\theta\|_2 \tag{4.8}$$

where $c(s)$ is the set of labels, θ represents all the trainable parameters, and λ is the regularization coefficient or the so-called weight decay rate. We apply the Adam algorithm [116] to optimize the objective function.

4.4 Data

We use data from two popular social networking websites, i.e, Reddit and Twitter, with totally three datasets derived. Two of them are from Reddit with one public dataset and one firstly collected in this section. People's posts from an active subreddit for online support in Reddit, called "SuicideWatch"(SW)[8], are intensively used in these two datasets. The last one is collected from Twitter by combining several existing data sources. These datasets cover suicide and other mental health issues, with specific categories reported in ICD-10 as listed in Table 4.1.

Table 4.1: Description of mental disorders in ICD-10

Category	Descriptions mentioned in ICD-10
Suicide	Intentional self-harm, suicidal ideation (tendencies)
Depression	In typical mild, moderate, or severe depressive episodes, the patient suffers from lowering of mood, reduction of energy, and decrease in activity.
Anxiety	Phobic anxiety and other anxiety disoders
Bipolar	A disorder characterized by two or more episodes in which the patient's mood and activity levels are significantly disturbed
PTSD	Arises as a delayed or protracted response to a stressful event or situation (of either brief or long duration) of an exceptionally threatening or catastrophic nature

[7] https://radimrehurek.com/gensim/

[8] https://www.reddit.com/r/SuicideWatch

4.4.1 UMD Reddit Suicidality Dataset

The UMD Reddit Suicidality Dataset [13] was collect from anonymous discussion forums in `Reddit.com`. It contains posts of 620 users in the training set and 245 users in the testing set sampled from 11,128 users in the subreddit "SuicideWatch" and 11,129 users in other subreddits. It is annotated by crowdsourcing workers and human experts via crowdsourced platform. The suicide risk is scaled to four levels, i.e., no risk(a), low(b), moderate(c), and severe risk(d). It also provides coarse labels where no risk and low risk are given the label of 0, moderated and severe risk are labelled as 1, together with the control group as the label of -1.

This dataset was released as the CLPsych 2018 Shared Task [13] and then a new version of it acted as the CLPsych 2019 Shared Task [61]. In this section, we use to dataset derived from the UMD dataset in user level four categories of suicide risk. The statistical information of the dataset is illustrated in Tab. 4.2. In addition, we include control users (labeled as "None") into this annotation set.

Table 4.2: Statistical information of UMD Reddit Suicidality Dataset

Annotation	Numbers	% of a/b/c/d/ levels
crowd	621	26%/10%/24%/40%
expert	245	29%/9%/25%/37%

We use the transformed labels from the raw label according to the original description of this dataset. Specifically, raw labels of "c" or "d" are transformed into 1, raw labels of "a" or "b" are transformed into 0, and the label of a control user is -1 by definition. We split the whole dataset into training, validation, and testing sets as listed in Table 4.3.

Table 4.3: Statistical information of UMD dataset with train/validation/test split

Label	#/% of train	#/% of valid.	#/% of test
-1	495/49.8489%	126/50.6024%	245/50.000%
0	188/31.2185%	89/35.7430%	86/17.551%
1	310/18.9325%	34/13.6546%	159/32.449%

4.4.2 Reddit SWMH Dataset

As severe mental health issues are very likely to lead to suicidal ideation, we also collect another dataset from some mental health related subreddits in `Reddit.com` to further the study of mental disorders and suicidal ideation. We name this dataset as Reddit SuicideWatch and Mental Health Collection, or SWMH for short, where discussions comprise suicide-related intention and mental disorders like depression, anxiety, and bipolar. We use the Reddit official API [9] and develop a web

[9] `https://www.reddit.com/dev/api/`

spider to collect the targeted forums. This collection contains totally 54,412 posts. Specific subreddits are listed in Table 4.4, as well as the number and the percentage of posts collected in train-val-test split.

In those communities or so-called subreddits, people discussed about their own or their relative's mental disorders and sought for advice or help. We perform experimental analysis on this dataset to identify discussions about suicidality and mental disorders.

Table 4.4: Statistical information of SuicideWatch and mental health related subreddits, i.e., SWMH dataset

Subreddit	#/% of train	#/% of valid.	#/% of test
depression	11,940/34.29	3,032/34.83	3,774/34.68
SuicideWatch	6,550/18.81	1,614/18.54	2,018/18.54
Anxiety	6,136/17.62	1,508/17.32	1,911/17.56
offmychest	5,265/15.12	1,332/15.30	1,687/15.50
bipolar	4,932/14.16	1,220/14.01	1,493/13.72

4.4.3 Twitter Datasets Collection

The third dataset is a collection of different subsets from Twitter. Sampled instances from two datasets consist of most samples of this dataset. First, 594 instances of tweets containing suicidal ideation are from Ji et al. [17], with additional 606 tweets manually collected by this work. Second, the same number of depression and post traumatic stress disorder (PTSD) posts are sampled from CLPsych 2015 shared task dataset [117]. This dataset is available upon request [10]. Last, control group where Twitter users not identified as having a mental condition or suicidal ideation is comprised by sampling normal tweets from previous mentioned datasets [17, 117]. Finally, this Twitter dataset collection contains totally 4,800 tweets with four classes of suicidality, depression, PTSD, and control.

4.4.4 Linguistic Clues and Emotion Polarity

We have a brief exploratory analysis on the data. Some selected linguistic statistical information of UMD dataset extracted by Linguistic Inquiry and Word Count software (LIWC)[11] is shown in Fig. 4.5. The risk of suicide increases among labels of -1, 0, and 1. The linguistic inquiry results show that negative emotion, anxiety, and sadness are expressed more in posts with high-level suicide risk. The same trends exist in family issues, death-related mentions and swear words. Naturally, positive emotions are less presented in posts with high suicide risk.

[10] Request for data access via `http://www.cs.jhu.edu/~mdredze/datasets/clpsych_shared_task_2015/`.

[11] `http://liwc.wpengine.com`

Table 4.5: Selected linguistic statistical information of UMD dataset extracted by LIWC

Linguistic clues	Label -1	Label 0	Label 1
positive emotion	3.30	3.12	2.96
negative emotion	1.56	2.30	2.74
anxiety	0.17	0.33	0.41
sadness	0.28	0.50	0.68
family	0.29	0.39	0.47
friend	0.43	0.56	0.54
work	2.54	1.92	1.80
money	1.13	0.71	0.61
death	0.22	0.29	0.36
swear words	0.23	0.33	0.40

4.5 Experiments

To evaluate the performance of our proposed model, we compare it with several text classification models on three real-world datasets. Baselines and empirical settings are introduced, and results are reported and discussed in this section.

4.5.1 Baseline and Settings

We compared five popular classification models with our proposed method. These baseline models are described as follow:

- **fastText** [118]: an efficient text classification model with bag of words sentence representation and a linear classifier.
- **CNN** [105]: it applies convolutional neural networks over the word embedding of sentence to produce feature maps, and then uses max-pooling over the features.
- **LSTM** [106]: it takes sequential word vectors as input to the recurrent LSTM cells and applies pooling over the output to obtain final representation. By combining the forward and backward direction, it becomes bidirectional LSTM (BiLSTM).
- **RCNN** [107]: this model at first applies LSTM model [106] to capture sequential information, and then applies CNN [105] to further extract features. It has bidirectional version using BiLSTM.
- **SSA** [110]: it proposed a structured self attention mechanism with multiple hops by introducing a 2D matrix for embedding representation. The self attention is applied to the sequential hidden states of the LSTM network.

All the baseline models and our proposed method are implemented by PyTorch[12] and run in a single GPU (Nvidia GeForce GTX 1080 Ti). We train the models for 50 epochs by default, setting batch

[12] `https://www.pytorch.org/`

size to be 128 and 16 according to the size of datasets. Specifically, the batch size of UMD dataset is 16, and for SWMH and Twitter data collection, the batch size is 128. For the word embedding, we use pretrained GloVe [55] word representation, with either static or dynamic embedding utilized. Our proposed method enumerates all the 250 subreddit lexicons of Reddit and the number of topics from 5 to 20. We select the best validation performance in multiple trails and report the testing performance as experimental results.

The goal of automatic detection is dedicated to producing effective diagnoses (i.e., true positive) and decreasing the incorrect diagnoses (i.e., false positive) to avoid patients' stress and anxiety caused by false detection. Thus, during the evaluation process, we only only focus on the prediction accuracy, but also report the weighted average F-score metric. For unbalanced datasets, we apply weight penalty to the objective function and report the weighted average results.

4.5.2 Results

We evaluate the experimental performance on three datasets collected from Reddit and Twitter. For the UMD Suicidality dataset and the SWMH dataset, the reported results are weighted average.

UMD Suicidality

We firstly implement our method and baselines on the UMD suicidality dataset for user-level classification. To process a set of posts from users, all posts of users are concatenated as user-level representation. The results of four metrics of accuracy, precision, recall, and F1-score are very close in this dataset. The BiLSTM model gains the highest accuracy of 56.94%, and our model follows at the second place of 56.73%. But our model has a higher F1-score than all the baselines. Noticing these very close results, we then go further analysis on results of each class in the next section.

Table 4.6: Comparison of different models on UMD dataset for user-level classification, where precision, recall, and F1 score are weighted average.

Model	Accuracy	Precision	Recall	F1
fastText	0.5327	0.5300	0.5327	0.5202
CNN	0.5531	0.4498	0.5531	0.4935
LSTM	0.5612	0.4625	0.5612	0.5071
BiLSTM	**0.5694**	0.5029	**0.5694**	0.5233
RCNN	0.5592	0.4953	0.5592	0.5111
SSA	0.5633	0.4711	0.5633	0.4839
RN	0.5673	**0.5405**	0.5673	**0.5453**

Reddit SWMH

Then, we perform experiments on the Reddit SWMH dataset, which contains both suicidal ideation and mental health issues to study the predictive performance of our model. It is a larger dataset with more instances when compared with the UMD dataset. Experiments on this dataset help us to have

an insight into the reasoning power of relation network for mental health related text with similar characteristics. As shown in Table 4.7, our model beats all baseline models in terms of all the four metrics.

Table 4.7: Comparison of different models on Reddit SWMH collection, where precision, recall, and F1 score are weighted average.

Model	Accuracy	Precision	Recall	F1
fastText	0.5722	0.5760	0.5722	0.5721
CNN	0.5657	0.5925	0.5657	0.5556
LSTM	0.5934	0.6032	0.5934	0.5917
BiLSTM	0.6196	0.6204	0.6196	0.6190
RCNN	0.6096	0.6161	0.6096	0.6063
SSA	0.6214	0.6249	0.6214	0.6226
RN	**0.6474**	**0.6510**	**0.6474**	**0.6478**

Twitter Collection

Lastly, we conduct experiments on the Twitter dataset with similar settings of previous experiments. Unlike posts in Reddit, tweets in this dataset are short sequences due to the tweet's length limit of 280 characters. The results of all baseline methods and our proposed method are shown in Table 4.8. Among these competitive methods, our model gains the best performance on these four metrics, with 1.77% and 1.82% improvement than the second best BiLSTM model in terms of accuracy and F1-score respectively. Our proposed method introduces auxiliary information of lexicon-based sentiment and topics features learnt from corpus, and utilizes relational reasoning for modeling the interaction between LSTM-based text encodings and risk indicators. Encoding richer information and efficient reasoning help our model boost performance in short tweet classification.

4.5.3 Performance on Each Class

This section studies the performance on each class of UMD dataset. We select two baselines with better performance for comparison. The results are shown in Fig. 4.9. The proposed RN-based model is poor on predicting post without suicidality, but good at predicting posts with high suicide risk.

Table 4.8: Performance comparison on Twitter dataset, where precision, recall, and F1 score are weighted average.

Model	Accuracy	Precision	Recall	F1
fastText	0.7927	0.7924	0.7927	0.7918
CNN	0.7885	0.7896	0.7885	0.7887
LSTM	0.8021	0.8094	0.8021	0.8039
BiLSTM	0.8208	0.8207	0.8208	0.8195
RCNN	0.8094	0.8089	0.8094	0.8090
SSA	0.8156	0.8149	0.8156	0.8152
RN	**0.8385**	**0.8381**	**0.8385**	**0.8377**

Unfortunately, all these three models have very poor capacity on predicting posts with low suicide risk. In the UMD dataset with a small volume of instances, these models tend to predict posts as classes with more instances, even though we apply penalty on the objective function.

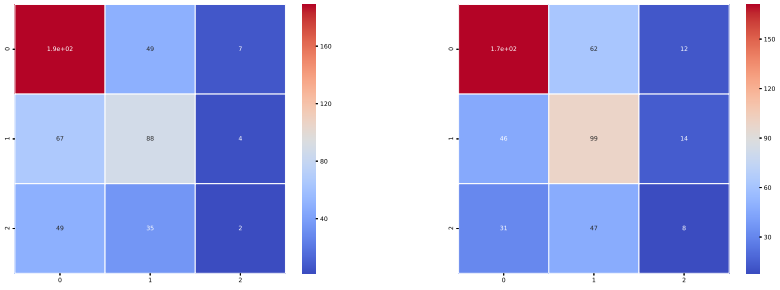
Table 4.9: Performance on each class of UMD suicidality dataset

Label	Metrics	BiLSTM	SSA	RN
-1	Precision	0.62	0.57	0.69
	Recall	0.77	**0.92**	0.70
	F1-score	0.69	0.70	0.69
1	Precision	0.51	0.57	0.48
	Recall	0.55	0.31	**0.62**
	F1-score	0.53	0.41	0.54
0	Precision	0.15	0.00	0.24
	Recall	0.02	0.00	0.09
	F1-score	0.04	0.00	0.13

4.5.4 Error Analysis

This section conducts error analysis, taking UMD dataset as an example. As mentioned before in last section, most methods suffer from poor performance on predicting low risk posts. Fig. 4.4 shows the heat maps of confusion matrix of BiLSTM and our RN-based model, where axis 0, 1, and 2 represent label of -1, 1 and 0. These two methods tend to predict more instances as none or high risk. And our proposed method has a slightly better result than its counterpart. We also notice that our proposed model can achieve a higher accuracy of 59.18% on the UMD dataset. But it fails in predicting low risk suicidal ideation, with similar performance of other baselines.

Figure 4.4: Confusion matrix on UMD dataset



(a) BiLSTM (b) RN

We use sentiment lexicons and topic model to exact sentiment- and topic-related risk indicators for relational reasoning with text encoding. This preprocessing procedure can cause error propagation. Sentiment varies in different social communities. Using existing lexicons in popular communities may have a limitation. In the future work, we will consider build lexicons from mental health related communities.

4.6 Summary

Although text classification on mental disorders can not be treated as medical diagnosis of professional practitioners, it can provide early warnings for the online users automatically when integrated into the online websites. It can also help the social workers and volunteers to identify the type of mental disorders, relieve online users' mental health issues through conversations, and suggest them with proper consultations or treatments.

This section attempts to reasoning suicidal ideation with sentimental indicators and life event related topical indicators, and proposes relation networks with attention mechanism for relational reasoning. Experiments show the effectiveness of our proposed model. We argue that it is a noteworthy step to combine canonical feature extraction with relation networks for reasoning.

Chapter 5

Federated Knowledge Transferring

5.1 Introduction

Proactive care is a kind of public service for healthcare and community assistance by connecting health organizations, social workers, and targeted patients. Traditional care service is based on face-to-face interaction in a certain place between general practitioners or social workers and people in need. Recently, with the help of online communication such as social networking services and private chatting, a new form of proactive online social service for mental health care has become available to online communities. Proactive social care provides people with early warning and support information to detect and relieve their mental disorders and social-related issues before their condition worsens.

Proactive social care for patients with mental disorders, especially depression and suicidality, is one of the most crucial services of social care in the modern society and has attracted worldwide attention. Mental health plays an important role in an individual's state of well-being. Mental health issues, such as depression, anxiety and post-traumatic stress disorder, have an adverse impact on people's daily life and health status. Untreated severe mental disorders could lead to suicidal ideation. According to WHO reports, around 300 million people suffer from depression[1], and about 900,000 people commit suicide worldwide every year[2]. Moreover, these figures continue to increase in every country across the world.

The traditional way to treat a mental health condition is psychological treatment, such as cognitive behavior therapy and interpersonal psychotherapy. This treatment relies heavily on health professionals such as general practitioners and psychiatrists. But current health services are not adequate to ensue effective treatment for such a huge number of potential sufferers. Furthermore, it is difficult to identify mental health issues at an early stage and take preventative action.

With the advances in the Web and mobile technology, mental health services are now using mobile devices to monitor a patient's health status and provide a platform for private communication

[1]WHO fact sheets about mental disorders, available at `https://www.who.int/news-room/fact-sheets/detail/mental-disorders`

[2]Suicide rates, Global Heath Observatory (GHO) data, available at `http://www.who.int/gho/mental_health/suicide_rates/en/`

within online communities to express mental stress. Conversation is one of the simplest and most effective ways to relieve an individual's mental disorders and even suicidal ideation. Online text-based communication helps people to express their feelings and sufferings in their daily life and work, providing psycholinguistic clues for early detection. It also provides a possible channel for volunteers and social workers to respond to risky social posts and address a sufferer's mental health issues through supportive comments.

There are many online platforms, forums and applications for chatting, peer support and early prevention, for instance, HealthfulChat, an online peer health support community, containing several chat rooms for mental health such as anxiety, bipolar and depression[3]; ReachOut Forums[4], an anonymous space for 14 to 25 year old Australians to share stories and receive support online discussion; Turn2Me[5], a lifeline web space for sharing and discussing personal issues; and Ibobbly, a mobile health intervention application [19]. Some services offered on these websites are delivered by mental health professionals. These services help people reach out and engage in conversations and consultation through the online communities.

Text-based chatting services such as SMS services, mobile APPs, Web applications, and social networking services could also be helpful in promoting mental wellbeing when integrated with mental health care services. Several works have studied the text-based synchronous conversations for mental health intervention [119]. Other works, especially those on early detection, focus on social networks for recognizing depression [120], detecting stress [121], and social network mental disorders [122]. These early detection strategies are preliminary for proactive online social care services for social support.

However, user generated messages are isolated in some private communication platforms like SMS and private chatting room services, which leads to challenge for effective detection. Deep learning techniques benefit from large-scale training data. But in such isolated small datasets, the detection performance could be severely dampened. Human beings have the ability of learning knowledge by transferring within two different domains. Federated transfer learning (FTL) [123] provides an approach for knowledge transferring among federated agents in a decentralized manner. Following the principle of FTL, we developed a knowledge transferring framework for online social care by aggregating learned knowledge on distributed clients. Specifically, clients send their local models to a global server and the global server ensembles models by aggregation algorithms. The aggregated model on a global server acts as the shared knowledge of all the clients, and it can be redistributed to clients for knowledge transferring.

This section proposes a cooperative framework with knowledge transferring and a novel model aggregation method for online social care, together with several components for proactive care services. This solution comprises four key features, i.e., language representation, on-device training, mental health detection, and effectiveness stratification of supportive responses, to empower intelligent proactive social care for mental health. We develop deep neural networks to learn the representation of

[3]Available at `http://www.healthfulchat.org/mental-health-chat-rooms.html`

[4]`https://au.reachout.com`

[5]`https://turn2me.org/`

text for language understanding, and study two tasks to enable proactive online social care under the framework of data protection.

This section contributes to the literature in the following three ways:

- We propose a cooperative knowledge ensemble framework with model aggregation for proactive social care by introducing a third party model server for knowledge ensemble.
- To improve knowledge ensemble for effective knowledge sharing and transferring, we proposed a two-step optimization and average difference descent for model aggregation.
- To evaluate our proposed algorithm, a case study on suicidal ideation detection and effectiveness stratification as services of online social care is conducted, resulting in better performance than the baselines.

The structure of this section is as follows. Related work are reviewed in Section 5.2. Our proposed framework is introduced in Section 5.3 together with an improved optimization algorithm. In Section 5.4, an experimental evaluation is conducted under the settings of proposed framework for online social care. A conclusion is drawn in Section 5.5 together with a brief discussion.

5.2 Related Work

This section is related to mental health care, such as the detection of depression or suicidality and conversation treatment, and federated learning and its variant federated transfer learning.

5.2.1 Mental Health Care

A large body of research focuses on mental health to provide proactive care for those who need it, especially the detection of mental health issues such as stressor events [124], depression [125], and suicidality [126]. Shuai et al. used a machine learning based model to perform multi-source learning for mental disorder detection in social media [122]. Tsugawa et al. extracted features from a user's Twitter activities and detected that the user was suffering from depression [120]. Nguyen et al. performed affective and content analysis through a comparison between depression communities as the clinical group and normal communities as the control group [127]. Lin et al. proposed a hybrid method of factor graph and convolutional neural network to detect psychological stress through tweet content and user interaction [121].

Severe mental disorders could turn to suicidality. Suicidal risk has been studied from the perspective of interaction between clinicians and patients [9], and knowledge discovery and detection using online social content [17]. De Choudhury et al. investigated the transition of mental health to suicidality in online social communities [87]. Ren et al. proposed a complex emotion model for suicidal intention detection in blogs [14]. Ji et al. proposed an improved model aggregation method to detect suicidal ideation in a distributed manner [18].

5.2.2 Federated Transfer Learning

Federated learning [128] is an on-device solution to decouple the training procedures from data collection. The method uses an iterative averaging model that can perform distributed training and learn efficiently from decentralized data to achieve the goal of preserving privacy. To improve communication efficiency, Konečnỳ et al. proposed structured updates and sketched updates to reduce uplink communication costs [129]. Geyer et al. proposed differential privacy preserving techniques on the client side to balance performance and privacy [130].

Another scenario is federated transfer learning [123] was proposed to transfer knowledge in a federation with the combination of transfer learning [131] where decentralized agents have different samples and features. The learning principle is quite similar to fast adaptive meta-learning [132, 133] and zero-data learning [134], that is, it learns a well-generalized global model in the data-free model server by aggregating the information learned from distributed clients.

5.3 Method

In this section, we develop a novel model aggregation over a federation of decentralized clients that enables knowledge transferring for proactive social care. It is designed with a third-party model server for knowledge ensemble and a two-step optimization strategy to decouple model training and data collecting. In particular, the datasets are located on decentralized clients, e.g. a physical electronic devices or an isolated software container, and the locally trained models serve as client's learned knowledge which are transported to global server for aggregation. Then, the global server aggregates the models into a global one as shared knowledge for clients.

5.3.1 Knowledge Ensemble and Transferring

Our method is under a decentralized learning framework by aggregation-based knowledge sharing and transferring for on-device training on local client devices. It is powered by a communication server with a chatting service for user data transmission and a third-party service provider, i.e., a mental health care service provider in this section, for the communication of model parameters. The framework is illustrated in Figure 5.1, where the communication service is separated from third-party proactive social care, providing third-party applications with an approach to making inferences without accessing the raw data. Knowledge transferring is enabled via model aggregation and redistribution through client-server communication.

The workflow of the knowledge transferring framework is illustrated in Figure 5.2. First, the model server chooses a learning model as the client model for each client to perform specific tasks on devices. In this section, we take two proactive social care tasks into consideration, i.e., text-based suicidality detection and social comment categorization. The first task aims to provide an early detection and warning system. The second task makes it easier for target users to access more effective responses. These two tasks are typically regarded as binary classification and multi-class classification problems,

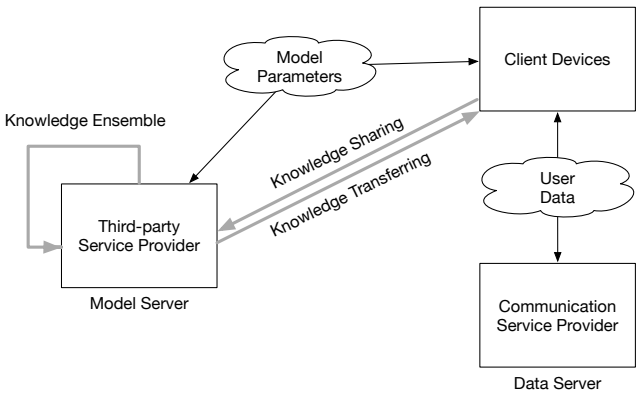


Figure 5.1: The illustration of knowledge ensemble framework with sharing and transferring

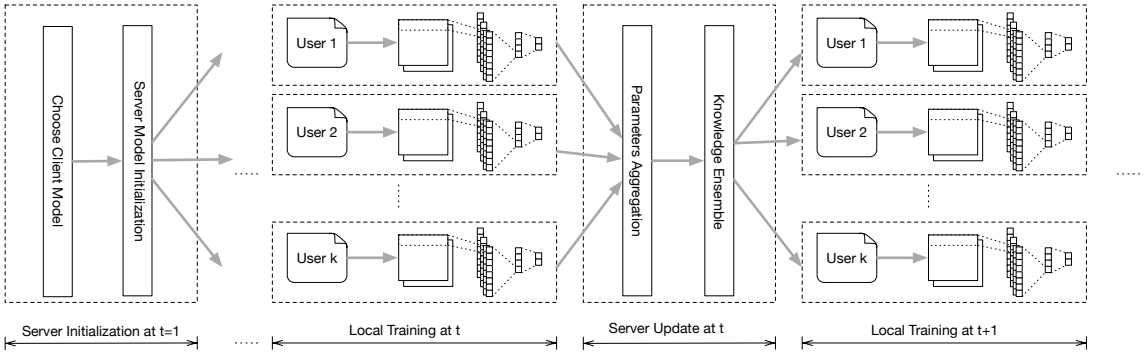


Figure 5.2: The workflow of knowledge transferring framework with model aggregation when taking CNN as the learning model

respectively. Deep neural networks, such as convolutional neural networks (CNNs) [105] for text and long short-term memory networks (LSTM) [106] are chosen as the classification model for clients to learn the language features in user generated content. In this illustration, CNN is used as an example. After parameter initialization, the model parameters as a form of shared knowledge are sent to the selected online clients to perform local training using the data of each user. Then, the locally trained models are sent back to the model server for model aggregation and knowledge updating. The intuitive way to undertake model aggregation is model averaging which takes local models equally. We propose a novel approach to aggregating a client model to optimize the global model for better knowledge ensemble, called **Av**era**g**e **Diff**erence **Agg**regation or AvgDiffAgg for short. The objective function and our proposed two-step optimization is introduced in detail in Section 5.3.2 and 5.3.3. A round of training consists of local training on devices, parameter sending, and model aggregation on the model server. The learning framework works through client and server communication by an iterative update.

5.3.2 Objective Function

In the private chatting scenario, deep learning methods are trained on user own data on their devices. However, isolated data is inadequate for training a deep learning model. To solve this problem, knowledge ensemble and transfeering in our proposed framework aim to provide a good initialization

to every user so that they can fine-tune a personalized deep learning model accordingly.

Training a deep learning model is a non-convex optimization task with many local optimal solutions or optimal points in the solution manifold. To address the local optimal problem in deep learning model training, there is an empirical assumption that if the initialization point of the model parameters is close to the global optimal point, arriving at the global optimal point or gaining a "better" local optimal point is more likely if the model is fine-tuned. Here, "better" is compared to the average results with randomly selected initialization points. Therefore, the optimal initialization point θ should be

$$\arg\min_{\theta} \mathbb{L} = \arg\min_{\theta} \sum_{k=1}^{n} \frac{1}{n} L(\theta, \theta^k) \tag{5.1}$$

where θ^k is the optimal parameter solution for the k-th user, and L is the loss function for measuring the distance between initialization point θ and each user's optimal point θ_k.

The procedure of finding the optimal global parameters is illustrated in Fig. 5.3. The central body of this illustration shows how the server weights are updated to optimal weights. The subfigure in the upper left corner shows how the local weights are composed as the gradient. The brown arrow in the form of the average difference between the model weights acts as the gradient. The optimization objective on the server side minimizes the average or expectation of the Euclidean distance between the server weight and the user weights. To facilitate this calculation, the loss function can be re-written as

$$\mathbb{L} = \sum_{k=1}^{n} \frac{1}{2n} L(\theta, \theta^k)^2 \tag{5.2}$$

where $L(\cdot, \cdot)$ is specified to Euclidean distance between two sets of weights, and m is the number of users or local devices.

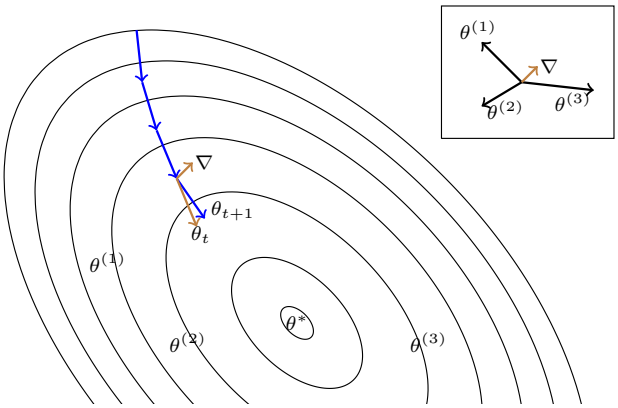


Figure 5.3: Finding the nearest global weight to all the optimal local weights. The blue arrows show the model updating towards the optimum parameter θ^*. The brown arrow ∇ acts as the "gradient".

5.3.3 Two-step Optimization

In Equation 5.2, the global model parameter θ and the k-th client model parameter θ_k are two correlated parameters that both need to be optimized. To solve this optimization problem, we propose a two-step optimization algorithm that uses gradient descent to simultaneously approach optimal θ and θ_k.

Specifically, optimization is an iterative procedure and each iteration t includes two steps that aim to separately update θ and θ_k.

In the first step of each iteration, we clip the value of each user's parameters θ_k, and update the global initialization point θ with the gradient derived from Equation 5.2 as:

$$\frac{\partial \mathbb{L}}{\partial \theta} = \frac{1}{n}\sum_{k=1}^{n}(\theta_t - \theta_t^k), \tag{5.3}$$

where $(\theta_t - \theta_t^k)$ is the difference between the initialization point and the optimal point for the k-th user. In each iteration, we update the global initialization point with the average difference for all users, corresponding to Algorithm 1 in this section.

$$\theta_{t+1} \leftarrow \theta_t - \varepsilon\frac{1}{n}\sum_{k=1}^{n}(\theta_t - \theta_{t+1}^k) \tag{5.4}$$

In practice, we can randomly sample part of the users in each iteration to estimate the "average difference" so that we can reduce computation and avoid overfitting. As the estimation of θ only requires part of the users' parameters, the proposed optimization framework is robust for the scenario in which some users are disconnected during the training procedure.

In the second step of any iteration, the global initialization parameters θ are fixed as transferred knowledge to selected clients, and then we can fine-tune each user's θ_k with gradient descent by using the user's own data D_k:

$$\theta^k = O(\theta, D_k, F_k) \tag{5.5}$$

where O is an operator that iteratively updates the θ for a certain number of epochs, and F_k is an arbitrary deep learning function applied on the local model of the k-th client.

Once the local model's optimal parameters θ_t^k have been learned from the local model, they are sent to the central server to estimate the average difference which contributes to updating the global parameters θ_{t+1} for next-round knowledge tranferring.

Algorithm 1 Average Difference Descent for the Optimization of Knowledge Ensemble

K is the total number of users; C is the fraction of users; ε is the stepsize of server execution.
Input: server parameters θ_t at t, client parameters $\theta_{t+1}^1, ..., \theta_{t+1}^m$ at $t+1$.
Output: aggregated server parameters θ_{t+1}.
procedure SERVER OPTIMIZATION
    initialize θ_0
    for each round t=1, 2, ... **do**
        $m \leftarrow max(C \cdot K, 1)$
        $S_t \leftarrow (random\ set\ of\ m\ users)$
        for each user k $\in S_t$ on local device **do**
            $\theta_{t+1}^k \leftarrow \text{LocalTraining}(k, \theta_t)$
        end for
        $\theta_{t+1} \leftarrow \theta_t - \varepsilon\frac{1}{m}\sum_{k=1}^{m}(\theta_t - \theta_{t+1}^k)$
    end for
end procedure

5.4 Experimental Evaluation

In this section, we introduce the architecture of proactive social care in online communities. Two tasks for proactive service, i.e., suicidal ideation detection and social response categorization are studied. Datasets and baselines are introduced as well as a series of comparative experiments.

5.4.1 Online Social Care

Online social care provides many kinds of care services for targeted users. In this section, we focus on mental health care in online communities for people who have a wide range of mental health issues. Under the learning framework, we produce two types of services, i.e., mental health detection and effectiveness stratification of social comments. We use suicidal ideation detection as the case study to demonstrate the application of proactive mental health detection. Suicide gestures and attempts are included in F60.3 – Emotionally unstable personality disorder of ICD-10 code from WHO[6]. Suicide is the most severe consequence of mental disorders. Post-schizophrenic depressive states may increase the risk of suicide[7]. For effectiveness stratification, it provides an evaluation and ranking of people's comments and can be used for easy access to more persuasive comments. The architecture of proactive social care for mental health is illustrated in Figure 5.4. We focus on the content from mental health discussion including the user's original post and the other user's comment on it. The proactive mental health care service is empowered by deep neural networks to learn language representation for early detection on posts and effectiveness stratification on comments.

5.4.2 Datasets

We collected data from two social websites – Reddit and Twitter. Table 5.1 lists the basic information of three datasets containing user posts derived from these platforms. For the task of effectiveness stratification, a dataset containing comment text is collected from Reddit.

Table 5.1: Summary of datasets

Datasets	# of users	# of posts/tweets
Reddit I	99	39,600
Reddit II	260	9,052
Twitter	102	10,200

Reddit Dataset. We obtained two datasets from the website Reddit, which is ranked No. 6 on the list of top websites worldwide by Alexa[8] world wide as of June 2018. As a social website, Reddit aggregates a variety of topics for online discussions and each discussion community with an interest in a particular discussion is called a "subreddit". There are a wide range of topics for online discussion, including social events and personal experience.

[6] `http://apps.who.int/classifications/apps/icd/icd10online2003/fr-icd.htm?gx60.htm+`

[7] According to the ICD-10 code F20.4 – Post-schizophrenic depression, available at `http://apps.who.int/classifications/apps/icd/icd10online2003/fr-icd.htm?gx60.htm+`.

[8] https://www.alexa.com/topsites

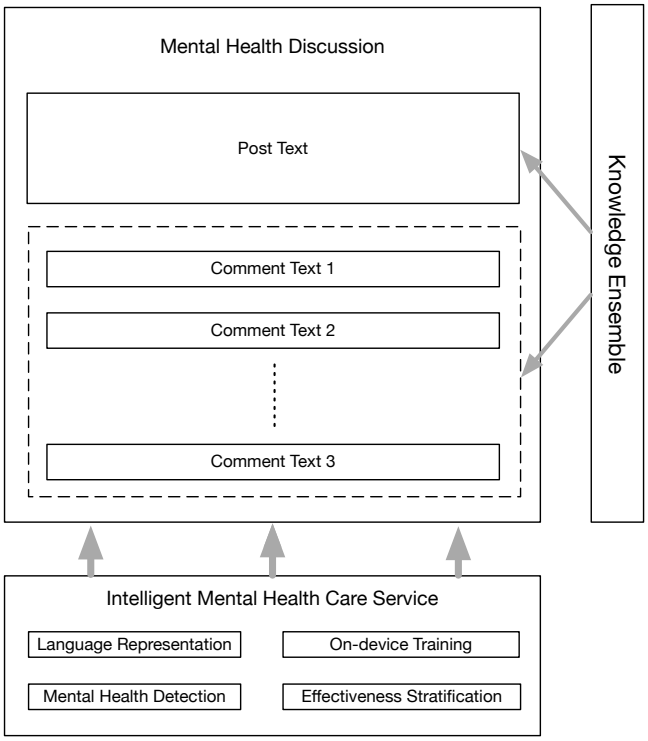


Figure 5.4: The architecture of proactive social care for mental health

The first aim of this work is to detect an individual's intent from social texts that involve suicidal ideation for early warning in proactive social care. A suicide-related subreddit called "SuicideWatch"[9], and two other subreddits not related to suicide, "popular"[10] and "AskReddit"[11] are taken as the source of content with a total of 39,600 posts collected. Of these posts, there are 48.16% of them containing suicidal ideation. We call this dataset as Reddit I.

Another dataset from Reddit, referred as Reddit II, contains a total of 9,052 posts from a total of 260 selected users in the Reddit community.

Twitter Dataset. The third dataset was collected from the social website Twitter. A keyword filtering technique was applied to collect the original tweets. The filtering terms included words such as "suicide", "die", and "death", and suicide-related phrases, such as "end my life" and "kill myself". Then, we manually checked and labeled the posts. Tweets containing keywords but without suicidal ideation were put them in the control group. This Twitter dataset contains a total of 10,200 tweets, of which 5.8% of tweets contained suicidal intention in the text.

Reddit Comments. We collected comments from all the users in the Reddit II dataset for effectiveness stratification. Each comment had a score given by the users who had viewed the comment and clicked the "like" button in the online forum. We scaled the scores of the comments into five classes according to the score distribution. The number of comments on different posts varies. Most posts contain less than 40 comments.

Dataset Partitioning. To mimic the real scenario of private chatting and decentralized training on

[9] `http://reddit.com/r/SuicideWatch`
[10] `http://reddit.com/r/popular`
[11] `http://reddit.com/r/AskReddit`

the client, we partitioned the data using independently identical distribution (I.I.D.). First, a shuffle is applied to the entire dataset and it is partitioned into several users with a certain number of examples. There are 99 users and 102 users in the Reddit dataset and the Twitter dataset, respectively. Users of Reddit and Twitter had 400 posts and 100 tweets, respectively.

5.4.3 Settings and Baselines

To evaluate our model, two baselines with model aggregation, i.e., FullbatchAgg and AverageAgg (where Agg stands for **Agg**regation), are used for comparative experiments.

These two baselines are described as follows:

1. FullbatchAgg: assembles an overall aggregation on the full batch of all users for only a single gradient descent step on each local device.

2. AverageAgg: samples a fraction of users for model aggregation applying weights over knowledge ensemble during model averaging.

The FullbatchAgg is a special case of AverageAgg where the epoch of local training equals 1 and the fraction of users equals 1.

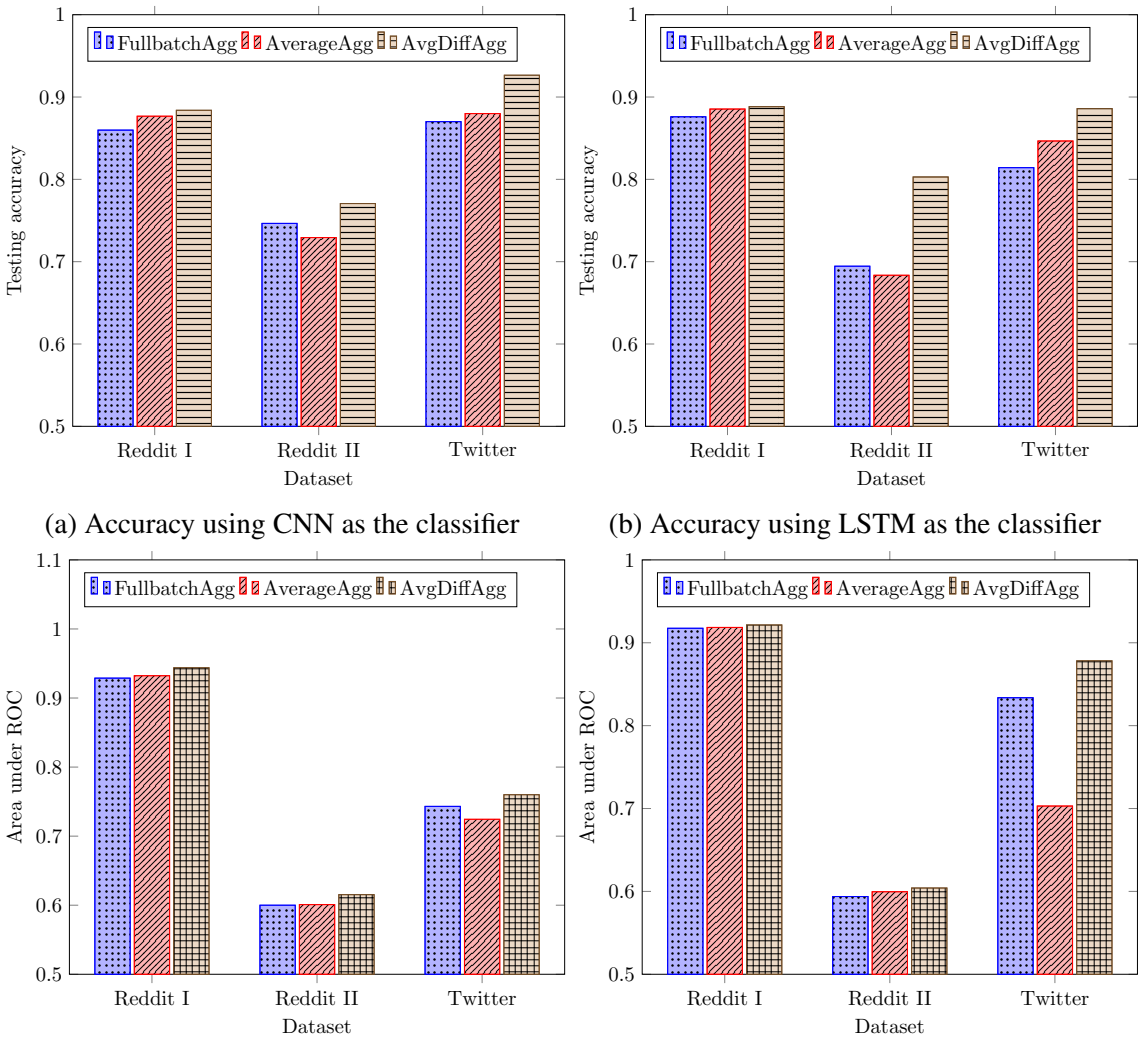
For the learning models of clients, two popular deep neural models, i.e., CNN [105] and LSTM [106], were used. First, we embedded the input sentence into a 100-dimension word vector to get the distributed representation of text. The word embedding was then placed into three convolutional layers. The learned features of the convolution layers are concatenated together to get the final representation of the text. Lastly, a fully connected layer was used as a classifier in the last layer to produce the prediction. For the LSTM model, we used the same settings for the word embedding and a 64-dimension LSTM hidden unit was used in the recurrent network.

5.4.4 Suicidal Ideation Detection

We firstly conduct experiments on suicidal ideation detection. To test the performance of our proposed learning framework and two-step knowledge transferring, we performed an empirical evaluation by comparing our method with those baselines.

Results. We compared all methods in terms of average testing accuracy and the average of area under the receiver operation curve (AUROC). The results are shown in Figure 5.5. We used the same hyperparameter settings using the same number of training rounds of 10. The local batchsize was 10, and the local training epochs were 5. For AverageAgg and our AvgDiffAgg, the fraction of users was both set to 0.1. As we can see from these figures, our proposed method achieves the best scores when using both CNN and LSTM as the classifier.

Learning curve. We drew the learning curve to visualize the performance of FullbatchAgg, AverageAgg, and our AvgDiffAgg as shown in Fig. 5.6. The training loss for our method decreased more rapidly than AverageAgg. The test accuracy of our method was higher than AverageAgg during the



(a) Accuracy using CNN as the classifier

(b) Accuracy using LSTM as the classifier

(c) AUROC using CNN as the classifier

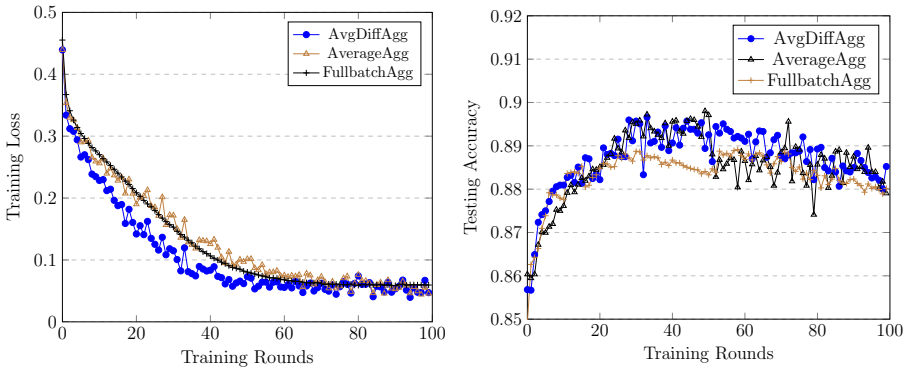
(d) AUROC using LSTM as the classifier

Figure 5.5: Classification Accuracy and AUROC using CNN and LSTM as the classifier

first 20 rounds of training and better still, between rounds 50 and 60. In the other rounds, the testing accuracy was similar. The training curve was smoother for the FullbatchAgg because it uses the full batch of users during aggregation at each iteration, while the other two methods use a random selection of users for model aggregation.

5.4.5 Effectiveness Stratification of Supporting Words

Conversation is one of the most effective ways to provide supportive words to the vulnerable people with mental health issues and even suicidal ideation. Gilat et al. [135], compared the responses to suicidal messages from trained volunteers and lay individuals, and found that trained volunteers employ more emotion-based strategies and more therapeutic-like cognitive-focused strategies than lay individuals who rely more on self-disclosure. The effectiveness stratification of supportive words evaluates social workers' responses in a given social care case, and it can help social workers to improve their conversational skill and compose better supportive words to persuade potential victims to relieve their mental health issue or give up a suicide attempt.



(a) Training loss (b) Testing accuracy

Figure 5.6: Training loss and testing accuracy on Reddit

In this section, we applied our AvgDiffAgg and the same baseline methods to evaluate the potential effectiveness of social comments according to the score of each comment received from other users in a supervised way.

We performed the experiments by training a CNN model and an LSTM model on the entire dataset with 10-fold cross validation. The average testing accuracy for the CNN and LSTM was 36.27% and 35.33%, respectively. These levels of accuracy are treated as the upper bound of the methods with data protection. The experiments using FullbatchAgg, AverageAgg and AvgDiffAgg were then performed 10 trials. The performance of different methods based on CNN and LSTM were then compared in terms of average testing accuracy. The experiment settings for these three methods were the same as the previous experiments, except for the fraction of users. For FullbatchAgg, this was always 1, and for the other two methods, it was set to 0.1. The results are shown in Table 5.2. Methods without data protection had higher accuracy than methods with data protection. Of the methods with data protection, our proposed method using CNN as the classifier was slightly better than AverageAgg. When using an LSTM as the classifier, the testing accuracy of our method was more than 2% higher than AverageAgg.

Table 5.2: Comparison of accuracy on predicting comment scores as effectiveness stratification

Methods	Avg. Acc.	
	CNN	LSTM
FullbatchAgg	30.49%	31.67%
AverageAgg	31.16%	32.80%
AvgDiffAgg	**31.35%**	**35.08%**

5.5 Summary

Third-party intelligent web information systems could pave the way for effective social support and improve proactive online social care services from broad perspectives. To solve the scarcity of training data in decentralized settings, especially in the private chatting, this section develops knowledge transferring method for proactive social care by using a decentralized learning framework

with on-device model training, knowledge transferring, and novel model aggregation. In particular, the proposed model aggregation strategy for knowledge ensemble updates the model parameters with the average difference descent that is inferred from a newly developed loss function customized for the proactive social service application scenario. The experiment evaluation on two tasks of suicidal ideation detection and effective stratification of social comments shows the effectiveness of the learning framework and the model aggregation algorithm.

Due to the highly sensitive nature of collecting real-world private data, this work mimics the real-world private chatting scenarios by using the public online data. The contents in the mimic dataset and the real-world private chatting dataset share similar characteristics and patterns that enable the proposed method to be a very promising solution to the development of new mental health care services in decentralized application.

Chapter 6

Conclusion

Mental health issues, such as anxiety, depression and suicidal ideation, are becoming increasingly concerning in modern society. Without effective treatment, severe mental disorders but without effective treatment are very likely to turn to suicide. The reasons people commit suicide are complicated, including social factors like social isolation; personal issues, such as alcoholism or career failure; or the influence of negative life events. Thus, developing effective suicide prevention techniques is urgently needed. Detecting suicidal ideation early is one of the most effective methods to prevent suicide. Research on suicidal intention understanding and suicide prevention mainly concentrates on its psychological and clinical aspects and classifying questionnaire results via supervised learning. However, collecting data and/or patients is typically expensive, from both a psychological and a clinical perspective.

With the advances in social media, more and more individuals are expressing their feelings and suffering on the internet. Anonymous online websites provide a comfortable place for people to interact with others using asynchronous communication. The social content in online communities for depression provides topic features and psycholinguistic clues for the automatic detection and prediction of suicidal ideation. Using data mining techniques on social networks and applying neural networks provide an avenue to understand the intention within online posts and even relieve a person's suicidal intentions.

This thesis firstly conducts a comprehensive literature review on current suicidal ideation detection methods including clinical methods based on the interaction between social workers or experts and the targeted individuals, and machine learning techniques with feature engineering or deep learning for automatic detection based on online social contents. Domain-specific applications of suicidal ideation detection are also reviewed according to their data sources, i.e., questionnaires, electronic health records, suicide notes, and online user content.

Then, to understand suicidal ideation through online user-generated content with the goal of early detection via supervised learning, we analyze users' language preferences and topic descriptions that reveals rich knowledge for detecting suicidal tendencies. Suicidal individuals express strong negative feelings, anxiety, and hopelessness. Suicidal thoughts may involve family and friends. And topics

they discuss cover both personal and social issues. To detect suicidal ideation, we extract several informative sets of features, including statistical, syntactic, linguistic, word embedding, and topic features, and we compare six classifiers, including four traditional supervised classifiers and two neural network models. An experimental study demonstrates the feasibility and practicability of the approach and provides benchmarks for the suicidal ideation detection on the active online platforms: Reddit SuicideWatch and Twitter.

Classifying suicidal ideation and other mental disorders is a challenging task as they share quite similar patterns in language usage and sentimental polarity. In this thesis, we enhance text representation with lexicon-based sentiment scores and latent topics, and propose to use relation networks for detecting suicidal ideation and mental disorders with related risk indicators. The relation module is further equipped with the attention mechanism to prioritize more important relational features. Through experiments on three real-world datasets, our model outperforms most of its counterparts.

In the last, we study a distributed setting and develop a knowledge transferring framework via model aggregation. Under this framework, distributed clients perform on-device training, and a third-party server integrates multiple clients' models and redistributes to clients for knowledge transferring among users. To improve the generalizability of the knowledge sharing, we further propose a novel model aggregation algorithm, namely the average difference descent aggregation. In particular, to evaluate the effectiveness of the learning algorithm, we use a case study on the early detection and prevention of suicidal ideation, and the experiment results on four datasets derived from social communities demonstrate the effectiveness of the proposed learning method.

The advances of deep learning techniques has boosted research on suicidal ideation detection. In the future work, more emerging learning techniques such as attention mechanism and graph neural networks will be studied for suicide text representation learning. Other learning paradigms such as transfer learning, adversarial training and reinforcement learning can also be utilized. For example, knowledge on mental health detection domain can be transferred for suicidal ideation detection, and generative adversarial networks can be used to generated adversarial samples for data augmentation. In social networking services, posts with suicidal ideation are in the long tail of the distribution of different post categories. In order to achieve effective detection in the ill-balanced distribution of real-world scenario, few-shot learning can be utilized to train on a few labeled posts with suicidal ideation among the large social corpus.

Bibliography

[1] S. Hinduja, J. W. Patchin, Bullying, cyberbullying, and suicide, Archives of suicide research 14 (3) (2010) 206–221.

[2] A. E. Crosby, B. Han, L. A. G. Ortega, S. E. Parks, J. Gfroerer, Suicidal thoughts and behaviors among adults aged $\geq$ 18 years–united states, 2008-2009., Morbidity And Mortality Weekly Report 60 (13) (2011) 1 – 22.

[3] J. Joo, S. Hwang, J. J. Gallo, Death ideation and suicidal ideation in a community sample who do not meet criteria for major depression, Crisis (2016) 161–165.

[4] M. K. Nock, G. Borges, E. J. Bromet, J. Alonso, M. Angermeyer, A. Beautrais, R. Bruffaerts, W. T. Chiu, G. De Girolamo, S. Gluzman, et al., Cross-national prevalence and risk factors for suicidal ideation, plans and attempts, The British Journal of Psychiatry 192 (2) (2008) 98–105.

[5] M. J. Vioulès, B. Moulahi, J. Azé, S. Bringay, Detection of suicide-related posts in twitter data streams, IBM Journal of Research and Development 62 (1) (2018) 1–21. doi:10.1147/JRD.2017.2768678.

[6] A. J. Ferrari, R. E. Norman, G. Freedman, A. J. Baxter, J. E. Pirkis, M. G. Harris, A. Page, E. Carnahan, L. Degenhardt, T. Vos, et al., The burden attributable to mental and substance use disorders as risk factors for suicide: findings from the global burden of disease study 2010, PloS one 9 (4) (2014) e91936.

[7] R. C. O'Connor, M. K. Nock, The psychology of suicidal behaviour, The Lancet Psychiatry 1 (1) (2014) 73–85.

[8] J. Lopez-Castroman, B. Moulahi, J. Azé, S. Bringay, J. Deninotti, S. Guillaume, E. Baca-Garcia, Mining social networks to improve suicide prevention: A scoping review, Journal of neuroscience research (2019) 1–10.

[9] V. Venek, S. Scherer, L.-P. Morency, J. Pestian, et al., Adolescent suicidal risk assessment in clinician-patient interaction, IEEE Transactions on Affective Computing 8 (2) (2017) 204–215.

[10] D. Delgado-Gomez, H. Blasco-Fontecilla, A. A. Alegria, T. Legido-Gil, A. Artes-Rodriguez, E. Baca-Garcia, Improving the accuracy of suicide attempter classification, Artificial intelligence in medicine 52 (3) (2011) 165–168.

[11] G. Liu, C. Wang, K. Peng, H. Huang, Y. Li, W. Cheng, SocInf: Membership inference attacks on social media health data with machine learning, IEEE Transactions on Computational Social Systems (2019) 907 – 921.

[12] B. O'Dea, S. Wan, P. J. Batterham, A. L. Calear, C. Paris, H. Christensen, Detecting suicidality on twitter, Internet Interventions 2 (2) (2015) 183–188.

[13] H.-C. Shing, S. Nair, A. Zirikly, M. Friedenberg, H. Daumé III, P. Resnik, Expert, crowdsourced, and machine assessment of suicide risk via online postings, in: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, 2018, pp. 25–36.

[14] F. Ren, X. Kang, C. Quan, Examining accumulated emotional traits in suicide blogs with an emotion topic model, IEEE Journal of Biomedical and Health Informatics 20 (5) (2016) 1384–1396.

[15] L. Yue, W. Chen, X. Li, W. Zuo, M. Yin, A survey of sentiment analysis in social media, Knowledge and Information Systems (2018) 1–47.

[16] A. Benton, M. Mitchell, D. Hovy, Multi-task learning for mental health using social media text, arXiv preprint arXiv:1712.03538 (2017).

[17] S. Ji, C. P. Yu, S.-f. Fung, S. Pan, G. Long, Supervised learning for suicidal ideation detection in online user content, Complexity 2018 (2018) 1–11.

[18] S. Ji, G. Long, S. Pan, T. Zhu, J. Jiang, S. Wang, Detecting suicidal ideation with data protection in online communities, in: International Conference on Database Systems for Advanced Applications, Springer, Cham, 2019, pp. 225–229.

[19] J. Tighe, F. Shand, R. Ridani, A. Mackinnon, N. De La Mata, H. Christensen, Ibobbly mobile health intervention for suicide prevention in australian indigenous youth: a pilot randomised controlled trial, BMJ open 7 (1) (2017) e013518.

[20] N. N. G. de Andrade, D. Pawson, D. Muriello, L. Donahue, J. Guadagno, Ethics and artificial intelligence: suicide prevention on facebook, Philosophy & Technology 31 (4) (2018) 669–684.

[21] L. C. McKernan, E. W. Clayton, C. G. Walsh, Protecting life while preserving liberty: Ethical recommendations for suicide prevention with artificial intelligence, Frontiers in psychiatry 9 (2018) 650.

[22] K. P. Linthicum, K. M. Schafer, J. D. Ribeiro, Machine learning in suicide science: Applications and ethics, Behavioral sciences & the law 37 (3) (2019) 214–222.

[23] S. Scherer, J. Pestian, L.-P. Morency, Investigating the speech characteristics of suicidal adolescents, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2013, pp. 709–713.

[24] D. Sikander, M. Arvaneh, F. Amico, G. Healy, T. Ward, D. Kearney, E. Mohedano, J. Fagan, J. Yek, A. F. Smeaton, et al., Predicting risk of suicide using resting state heart rate, in: Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016 Asia-Pacific, IEEE, 2016, pp. 1–4.

[25] M. A. Just, L. Pan, V. L. Cherkassky, D. L. McMakin, C. Cha, M. K. Nock, D. Brent, Machine learning of neural representations of suicide and emotion concepts identifies suicidal youth, Nature human behaviour 1 (12) (2017) 911.

[26] N. Jiang, Y. Wang, L. Sun, Y. Song, H. Sun, An erp study of implicit emotion processing in depressed suicide attempters, in: Information Technology in Medicine and Education (ITME), 2015 7th International Conference on, IEEE, 2015, pp. 37–40.

[27] M. Lotito, E. Cook, A review of suicide risk assessment instruments and approaches, Mental Health Clinician 5 (5) (2015) 216–223.

[28] Z. Tan, X. Liu, X. Liu, Q. Cheng, T. Zhu, Designing microblog direct messages to engage social media users with suicide ideation: interview and survey study on weibo, Journal of medical Internet research 19 (12) (2017) e381.

[29] Y.-P. Huang, T. Goh, C. L. Liew, Hunting suicide notes in web 2.0-preliminary findings, in: Multimedia Workshops, 2007. ISMW'07. Ninth IEEE International Symposium on, IEEE, 2007, pp. 517–521.

[30] K. D. Varathan, N. Talib, Suicide detection system based on twitter, in: Science and Information Conference (SAI), 2014, IEEE, 2014, pp. 785–788.

[31] J. Jashinsky, S. Burton, C. Hanson, J. West, C. Giraud-Carrier, M. Barnes, T. Argyle, Tracking suicide risk factors through twitter in the US, Crisis: The Journal of Crisis Intervention and Suicide Prevention 35 (1) (2014) 51–59.

[32] J. F. Gunn, D. Lester, Twitter postings and suicide: An analysis of the postings of a fatal suicide in the 24 hours prior to death, Suicidologi 17 (3) (2015) 28–30.

[33] G. Coppersmith, R. Leary, E. Whyne, T. Wood, Quantifying suicidal ideation via language usage on social media, in: Joint Statistics Meetings Proceedings, Statistical Computing Section, JSM, 2015, pp. 1–15.

[34] G. B. Colombo, P. Burnap, A. Hodorog, J. Scourfield, Analysing the connectivity and communication of suicidal users on twitter, Computer communications 73 (2016) 291–300.

[35] G. Coppersmith, K. Ngo, R. Leary, A. Wood, Exploratory analysis of social media prior to a suicide attempt, in: Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology, 2016, pp. 106–117.

[36] P. Solano, M. Ustulin, E. Pizzorno, M. Vichi, M. Pompili, G. Serafini, M. Amore, A google-based approach for monitoring suicide risk, Psychiatry research 246 (2016) 581–586.

[37] M. E. Larsen, N. Cummins, T. W. Boonstra, B. O'Dea, J. Tighe, J. Nicholas, F. Shand, J. Epps, H. Christensen, The use of technology in suicide prevention, in: Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE, IEEE, 2015, pp. 7316–7319.

[38] H. Y. Huang, M. Bashir, Online community and suicide prevention: Investigating the linguistic cues and reply bias, in: CHI'16, 2016.

[39] M. De Choudhury, E. Kiciman, The language of social support in social media and its effect on suicidal ideation risk, in: Eleventh International AAAI Conference on Web and Social Media, AAAI, 2017, pp. 1–10.

[40] N. Masuda, I. Kurahashi, H. Onari, Suicide ideation of individuals in online social networks, PloS one 8 (4) (2013).

[41] S. Chattopadhyay, A study on suicidal risk analysis, in: e-Health Networking, Application and Services, 2007 9th International Conference on, IEEE, 2007, pp. 74–78.

[42] D. Delgado-Gomez, H. Blasco-Fontecilla, F. Sukno, M. S. Ramos-Plasencia, E. Baca-Garcia, Suicide attempters classification: Toward predictive models of suicidal behavior, Neurocomputing 92 (2012) 3–8.

[43] S. Chattopadhyay, A mathematical model of suicidal-intent-estimation in adults, American Journal of Biomedical Engineering 2 (6) (2012) 251–262.

[44] W. Wang, L. Chen, M. Tan, S. Wang, A. P. Sheth, Discovering fine-grained sentiment in suicide notes, Biomedical informatics insights 5 (Suppl 1) (2012) 137.

[45] A. Abboute, Y. Boudjeriou, G. Entringer, J. Azé, S. Bringay, P. Poncelet, Mining twitter for suicide prevention, in: International Conference on Applications of Natural Language to Data Bases/Information Systems, Springer, 2014, pp. 250–253.

[46] E. Okhapkina, V. Okhapkin, O. Kazarin, Adaptation of information retrieval methods for identifying of destructive informational influence in social networks, in: Advanced Information Networking and Applications Workshops (WAINA), 2017 31st International Conference on, IEEE, 2017, pp. 87–92.

[47] M. Mulholland, J. Quinn, Suicidal tendencies: The automatic classification of suicidal and non-suicidal lyricists using nlp., in: IJCNLP, 2013, pp. 680–684.

[48] X. Huang, L. Zhang, D. Chiu, T. Liu, X. Li, T. Zhu, Detecting suicidal ideation in chinese microblogs with psychological lexicons, in: Ubiquitous Intelligence and Computing, 2014 IEEE 11th Intl Conf on and IEEE 11th Intl Conf on and Autonomic and Trusted Computing, and IEEE 14th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UTC-ATC-ScalCom), IEEE, 2014, pp. 844–849.

[49] X. Huang, X. Li, T. Liu, D. Chiu, T. Zhu, L. Zhang, Topic model for identifying suicidal ideation in chinese microblog, in: Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation, 2015, pp. 553–562.

[50] Y.-M. Tai, H.-W. Chiu, Artificial neural network analysis on suicide and self-harm history of taiwanese soldiers, in: Innovative Computing, Information and Control, 2007. ICICIC'07. Second International Conference on, IEEE, 2007, pp. 363–363.

[51] M. Liakata, J. H. Kim, S. Saha, J. Hastings, D. Rebholzschuhmann, Three hybrid classifiers for the detection of emotions in suicide notes, Biomedical Informatics Insights,2012,Suppl. 1(2012-01-30) 2012 ((Suppl. 1)) (2012) 175–184.

[52] J. Pestian, H. Nasrallah, P. Matykiewicz, A. Bennett, A. Leenaars, Suicide note classification using natural language processing: A content analysis, Biomedical informatics insights 2010 (3) (2010) 19.

[53] S. R. Braithwaite, C. Giraud-Carrier, J. West, M. D. Barnes, C. L. Hanson, Validating machine learning algorithms for twitter data against established measures of suicidality, JMIR mental health 3 (2) (2016) e21.

[54] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781 (2013).

[55] J. Pennington, R. Socher, C. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[56] S. Ji, G. Long, S. Pan, T. Zhu, J. Jiang, S. Wang, X. Li, Knowledge transferring via model aggregation for online social care, arXiv preprint arXiv:1905.07665 (2019).

[57] M. Gaur, A. Alambo, J. P. Sain, U. Kursuncu, K. Thirunarayan, R. Kavuluru, A. Sheth, R. Welton, J. Pathak, Knowledge-aware assessment of severity of suicide risk for early intervention, in: The World Wide Web Conference, ACM, 2019, pp. 514–525.

[58] G. Coppersmith, R. Leary, P. Crutchley, A. Fine, Natural language processing of social media as screening for suicide risk, Biomedical Informatics Insights 10 (2018) 1–11.

[59] R. Sawhney, P. Manchanda, P. Mathur, R. Shah, R. Singh, Exploring and learning suicidal ideation connotations on social media with deep learning, in: Proceedings of the 9th Workshop

on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 2018, pp. 167–175.

[60] S. Ji, X. Li, Z. Huang, E. Cambria, Suicidal ideation and mental disorder detection with attentive relation networks, arXiv preprint arXiv:2004.07601 (2020).

[61] A. Zirikly, P. Resnik, O. Uzuner, K. Hollingshead, Clpsych 2019 shared task: Predicting the degree of suicide risk in reddit posts, in: Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology, 2019, pp. 24–33.

[62] A. G. Hevia, R. C. Menéndez, D. Gayo-Avello, Analyzing the use of existing systems for the clpsych 2019 shared task, in: Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology, 2019, pp. 148–151.

[63] M. Morales, P. Dey, T. Theisen, D. Belitz, N. Chernova, An investigation of deep learning systems for suicide risk assessment, in: Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology, 2019, pp. 177–181.

[64] M. Matero, A. Idnani, Y. Son, S. Giorgi, H. Vu, M. Zamani, P. Limbachiya, S. C. Guntuku, H. A. Schwartz, Suicide risk assessment with multi-level dual-context language and bert, in: Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology, 2019, pp. 39–44.

[65] L. Chen, A. Aldayel, N. Bogoychev, T. Gong, Similar minds post alike: Assessment of suicide risk using a hybrid model, in: Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology, 2019, pp. 152–157.

[66] X. Zhao, S. Lin, Z. Huang, Text classification of micro-blog's tree hole based on convolutional neural network, in: Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence, ACM, 2018, p. 61.

[67] T. Tran, D. Phung, W. Luo, R. Harvey, M. Berk, S. Venkatesh, An integrated framework for suicide risk prediction, in: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2013, pp. 1410–1418.

[68] S. Berrouiguet, R. Billot, P. Lenca, P. Tanguy, E. Baca-Garcia, M. Simonnet, B. Gourvennec, Toward e-health applications for suicide prevention, in: 2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE), IEEE, 2016, pp. 346–347.

[69] D. Meyer, J.-A. Abbott, I. Rehm, S. Bhar, A. Barak, G. Deng, K. Wallace, E. Ogden, B. Klein, Development of a suicidal ideation detection tool for primary healthcare settings: using open access online psychosocial data, Telemedicine and e-Health 23 (4) (2017) 273–281.

[70] K. M. Harris, J. P. McLean, J. Sheffield, Suicidal and online: How do online behaviors inform us of this high-risk population?, Death studies 38 (6) (2014) 387–394.

[71] H. Sueki, The association of suicide-related twitter use with suicidal behaviour: a cross-sectional study of young internet users in japan, Journal of affective disorders 170 (2015) 155–160.

[72] K. W. Hammond, R. J. Laundry, T. M. OLeary, W. P. Jones, Use of text search to effectively identify lifetime prevalence of suicide attempts among veterans, in: 2013 46th Hawaii International Conference on System Sciences, IEEE, 2013, pp. 2676–2683.

[73] C. G. Walsh, J. D. Ribeiro, J. C. Franklin, Predicting risk of suicide attempts over time through machine learning, Clinical Psychological Science 5 (3) (2017) 457–469.

[74] T. Iliou, G. Konstantopoulou, M. Ntekouli, D. Lymberopoulos, K. Assimakopoulos, D. Galiatsatos, G. Anastassopoulos, Machine learning preprocessing method for suicide prediction, in: L. Iliadis, I. Maglogiannis (Eds.), Artificial Intelligence Applications and Innovations, Springer International Publishing, Cham, 2016, pp. 53–60.

[75] T. Nguyen, T. Tran, S. Gopakumar, D. Phung, S. Venkatesh, An evaluation of randomized machine learning methods for redundant data: Predicting short and medium-term suicide risk from administrative records and risk assessments, arXiv preprint arXiv:1605.01116 (2016).

[76] H. S. Bhat, S. J. Goldman-Mellor, Predicting adolescent suicide attempts with neural networks, in: NIPS 2017 Workshop on Machine Learning for Health, 2017, pp. 1–8.

[77] J. P. Pestian, P. Matykiewicz, M. Linn-Gust, B. South, O. Uzuner, J. Wiebe, K. B. Cohen, J. Hurdle, C. Brew, Sentiment analysis of suicide notes: A shared task, Biomedical informatics insights 5 (Suppl. 1) (2012) 3.

[78] E. White, L. J. Mazlack, Discerning suicide notes causality using fuzzy cognitive maps, in: Fuzzy Systems (FUZZ), 2011 IEEE International Conference on, IEEE, 2011, pp. 2940–2947.

[79] B. Desmet, V. Hoste, Emotion detection in suicide notes, Expert Systems with Applications 40 (16) (2013) 6351–6358.

[80] R. Wicentowski, M. R. Sydes, Emotion detection in suicide notes using maximum entropy classification, Biomedical informatics insights 5 (2012) BII–S8972.

[81] A. Kovačević, A. Dehghan, J. A. Keane, G. Nenadic, Topic categorisation of statements in suicide notes with integrated rules and machine learning, Biomedical informatics insights 5 (2012) BII–S8978.

[82] A. M. Schoene, N. Dethlefs, Automatic identification of suicide notes from linguistic and sentiment features, in: Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, 2016, pp. 128–133.

[83] J. Robinson, G. Cox, E. Bailey, S. Hetrick, M. Rodrigues, S. Fisher, H. Herrman, Social media and suicide prevention: a systematic review, Early intervention in psychiatry 10 (2) (2016) 103–121.

[84] Y. Wang, S. Wan, C. Paris, The role of features and context on suicide ideation detection, in: Proceedings of the Australasian Language Technology Association Workshop 2016, 2016, pp. 94–102.

[85] A. Shepherd, C. Sanders, M. Doyle, J. Shaw, Using social media for support and feedback by mental health service users: thematic analysis of a twitter conversation, BMC psychiatry 15 (1) (2015) 29.

[86] M. De Choudhury, S. De, Mental health discourse on reddit: Self-disclosure, social support, and anonymity., in: ICWSM, 2014, pp. 1–10.

[87] M. De Choudhury, E. Kiciman, M. Dredze, G. Coppersmith, M. Kumar, Discovering shifts to suicidal ideation from mental health content in social media, in: CHI, ACM, 2016, pp. 2098–2110.

[88] M. Kumar, M. Dredze, G. Coppersmith, M. De Choudhury, Detecting changes in suicide content manifested in social media following celebrity suicides, in: Proceedings of the 26th ACM Conference on Hypertext & Social Media, ACM, 2015, pp. 85–94.

[89] L. Guan, B. Hao, Q. Cheng, P. S. Yip, T. Zhu, Identifying chinese microblog users with high suicide probability using internet-based profile and linguistic features: classification model, JMIR mental health 2 (2) (2015) e17.

[90] S. J. Cash, M. Thelwall, S. N. Peck, J. Z. Ferrell, J. A. Bridge, Adolescent suicide statements on myspace, Cyberpsychology, Behavior, and Social Networking 16 (3) (2013) 166–174.

[91] Suicide rates, Global Health Observatory (GHO) data, available in `http://www.who.int/gho/mental_health/suicide_rates/en/` (2015).

[92] M. Mohri, A. Rostamizadeh, A. Talwalkar, Foundations of machine learning, MIT press, 2012.

[93] C. Cortes, V. Vapnik, Support vector machine, Machine learning 20 (3) (1995) 273–297.

[94] L. Breiman, Random forests, Machine learning 45 (1) (2001) 5–32.

[95] J. H. Friedman, Greedy function approximation: a gradient boosting machine, Annals of statistics (2001) 1189–1232.

[96] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, ACM, 2016, pp. 785–794.

[97] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (8) (1997) 1735–1780.

[98] J. W. Pennebaker, R. L. Boyd, K. Jordan, K. Blackburn, The development and psychometric properties of LIWC2015, Tech. rep. (2015).

[99] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, Journal of machine Learning research 3 (Jan) (2003) 993–1022.

[100] A. Voutilainen, Part-of-speech tagging, The Oxford handbook of computational linguistics (2003) 219–232.

[101] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781 (2013).

[102] I. T. Jolliffe, Principal component analysis and factor analysis, in: Principal component analysis, Springer, 1986, pp. 115–128.

[103] K. Jacob, V. Patel, Classification of mental disorders: a global mental health perspective, The Lancet 383 (9926) (2014) 1433–1435.

[104] I. Gilat, Y. Tobin, G. Shahar, Offering support to suicidal individuals in an online support group, Archives of Suicide Research 15 (3) (2011) 195–206.

[105] Y. Kim, Convolutional neural networks for sentence classification, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1746–1751.

[106] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (8) (1997) 1735–1780.

[107] S. Lai, L. Xu, K. Liu, J. Zhao, Recurrent convolutional neural networks for text classification., in: AAAI, Vol. 333, 2015, pp. 2267–2273.

[108] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473 (2014).

[109] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, E. Hovy, Hierarchical attention networks for document classification, in: NAACL, 2016, pp. 1480–1489.

[110] Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, Y. Bengio, A structured self-attentive sentence embedding, arXiv preprint arXiv:1703.03130 (2017).

[111] D. Raposo, A. Santoro, D. Barrett, R. Pascanu, T. Lillicrap, P. Battaglia, Discovering objects and their relations from entangled scene representations, arXiv preprint arXiv:1702.05068 (2017).

[112] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, T. Lillicrap, A simple neural network module for relational reasoning, in: NIPS, 2017, pp. 4967–4976.

[113] R. Socher, D. Chen, C. D. Manning, A. Ng, Reasoning with neural tensor networks for knowledge base completion, in: NIPS, 2013, pp. 926–934.

[114] W. L. Hamilton, K. Clark, J. Leskovec, D. Jurafsky, Inducing domain-specific sentiment lexicons from unlabeled corpora, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Vol. 2016, 2016, p. 595.

[115] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, Journal of Machine Learning Research 3 (Jan) (2003) 993–1022.

[116] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[117] G. Coppersmith, M. Dredze, C. Harman, K. Hollingshead, M. Mitchell, Clpsych 2015 shared task: Depression and ptsd on twitter, in: Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, 2015, pp. 31–39.

[118] A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, Bag of tricks for efficient text classification, arXiv preprint arXiv:1607.01759 (2016).

[119] S. Hoermann, K. L. McCabe, D. N. Milne, R. A. Calvo, Application of synchronous text-based dialogue systems in mental health interventions: systematic review, Journal of Medical Internet Research 19 (8) (2017) e267.

[120] S. Tsugawa, Y. Kikuchi, F. Kishino, K. Nakajima, Y. Itoh, H. Ohsaki, Recognizing depression from twitter activity, in: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, ACM, 2015, pp. 3187–3196.

[121] H. Lin, J. Jia, J. Qiu, Y. Zhang, G. Shen, L. Xie, J. Tang, L. Feng, T.-S. Chua, Detecting stress based on social interactions in social networks, IEEE Transactions on Knowledge and Data Engineering 29 (9) (2017) 1820–1833.

[122] H.-H. Shuai, C.-Y. Shen, D.-N. Yang, Y.-F. C. Lan, W.-C. Lee, S. Y. Philip, M.-S. Chen, A comprehensive study on social network mental disorders detection via online social media mining, IEEE Transactions on Knowledge and Data Engineering 30 (7) (2018) 1212–1225.

[123] Q. Yang, Y. Liu, T. Chen, Y. Tong, Federated machine learning: Concept and applications, ACM Transactions on Intelligent Systems and Technology (TIST) 10 (2) (2019) 1–19.

[124] Q. Li, Y. Xue, L. Zhao, J. Jia, L. Feng, Analyzing and identifying teens' stressful periods and stressor events from a microblog, IEEE Journal of Biomedical and Health Informatics 21 (5) (2017) 1434–1448.

[125] M. De Choudhury, M. Gamon, S. Counts, E. Horvitz, Predicting depression via social media, in: AAAI ICWSM, 2013, pp. 1–10.

[126] B. O'Dea, S. Wan, P. J. Batterham, A. L. Calear, C. Paris, H. Christensen, Detecting suicidality on twitter, Internet Interventions 2 (2) (2015) 183–188.

[127] T. Nguyen, D. Phung, B. Dao, S. Venkatesh, M. Berk, Affective and content analysis of online depression communities, IEEE Transactions on Affective Computing 5 (3) (2014) 217–226.

[128] B. McMahan, E. Moore, D. Ramage, S. Hampson, B. A. y Arcas, Communication-efficient learning of deep networks from decentralized data, in: Artificial Intelligence and Statistics, 2017, pp. 1273–1282.

[129] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, D. Bacon, Federated learning: Strategies for improving communication efficiency, arXiv preprint arXiv:1610.05492 (2016).

[130] R. C. Geyer, T. Klein, M. Nabi, Differentially private federated learning: A client level perspective, arXiv preprint arXiv:1712.07557 (2017).

[131] S. J. Pan, Q. Yang, A survey on transfer learning, IEEE Transactions on Knowledge and Data Engineering 22 (10) (2010) 1345–1359.

[132] C. Finn, P. Abbeel, S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks, arXiv preprint arXiv:1703.03400 (2017).

[133] A. Nichol, J. Achiam, J. Schulman, On first-order meta-learning algorithms, arXiv preprint arXiv:1803.02999 (2018).

[134] H. Larochelle, D. Erhan, Y. Bengio, Zero-data learning of new tasks., in: AAAI, 2008, pp. 646–651.

[135] I. Gilat, Y. Tobin, G. Shahar, Responses to suicidal messages in an online support group: comparison between trained volunteers and lay individuals, Social psychiatry and psychiatric epidemiology 47 (12) (2012) 1929–1935.