5-2019

Knowledge-aware Assessment of Severity of Suicide Risk for Early Intervention

Manas Gaur
Wright State University

Amanuel Alambo
Wright State University

Joy Prakash Sain
Wright State University

Ugur Kursuncu
Wright State University

Krishnaprasad Thirunarayan
Wright State University

See next page for additional authors

Publication Info

Published in *Companion Proceedings of The 2019 World Wide Web Conference*, 2019, pages 514-525.

Knowledge-aware Assessment of Severity of Suicide Risk for Early Intervention

Manas Gaur
Knoesis Center
Dayton, Ohio
manas@knoesis.org

Amanuel Alambo
Knoesis Center
Dayton, Ohio
amanuel@knoesis.org

Joy Prakash Sain
Knoesis Center
Dayton, Ohio
joy@knoesis.org

Ugur Kursuncu
Knoesis Center
Dayton, Ohio
ugur@knoesis.org

Krishnaprasad Thirunarayan
Knoesis Center
Dayton, Ohio
tkprasad@knoesis.org

Ramakanth Kavuluru
University of Kentucky
Lexington, Kentucky
ramakanth.kavuluru@uky.edu

Amit Sheth
Knoesis Center
Dayton, Ohio
amit@knoesis.org

Randon S. Welton
Department of Psychiatry
Dayton, Ohio
randon.welton@wright.edu

Jyotishman Pathak
Cornell University
New York, NY
jyp2001@med.cornell.edu

ABSTRACT

Mental health illness such as depression is a significant risk factor for suicide ideation, behaviors, and attempts. A report by Substance Abuse and Mental Health Services Administration (SAMHSA) shows that 80% of the patients suffering from Borderline Personality Disorder (BPD) have suicidal behavior, 5-10% of whom commit suicide. While multiple initiatives have been developed and implemented for suicide prevention, a key challenge has been the social stigma associated with mental disorders, which deters patients from seeking help or sharing their experiences directly with others including clinicians. This is particularly true for teenagers and younger adults where suicide is the second highest cause of death in the US. Prior research involving surveys and questionnaires (e.g. PHQ-9) for suicide risk prediction failed to provide a quantitative assessment of risk that informed timely clinical decision-making for intervention. Our interdisciplinary study concerns the use of Reddit as an unobtrusive data source for gleaning information about suicidal tendencies and other related mental health conditions afflicting depressed users. We provide details of our learning framework that incorporates domain-specific knowledge to predict the severity of suicide risk for an individual. Our approach involves developing a suicide risk severity lexicon using medical knowledge bases and suicide ontology to detect cues relevant to suicidal thoughts and actions. We also use language modeling, medical entity recognition and normalization and negation detection to create a dataset of 2181 redditors that have discussed or implied suicidal ideation, behavior, or attempt. Given the importance of clinical knowledge, our gold standard dataset of 500 redditors (out of 2181) was developed by four practicing psychiatrists following the guidelines outlined in Columbia Suicide Severity Rating Scale (C-SSRS), with the pairwise annotator agreement of 0.79 and group-wise agreement of 0.73. Compared to the existing four-label classification scheme (no risk, low risk, moderate risk, and high risk), our proposed C-SSRS-based 5-label classification scheme distinguishes people who are supportive, from those who show different severity of suicidal tendency. Our 5-label classification scheme outperforms the state-of-the-art schemes by improving the graded recall by 4.2% and reducing the perceived risk measure by 12.5%. Convolutional neural network (CNN) provided the best performance in our scheme due to the discriminative features and use of domain-specific knowledge resources, in comparison to SVM-L that has been used in the state-of-the-art tools over similar dataset.

CCS CONCEPTS

• **Human-centered computing** → *HCI design and evaluation methods*; • **Computing methodologies** → **Natural language processing**; **Machine learning**; • **Applied computing** → **Health informatics**.

KEYWORDS

Surveillance and Behavior Monitoring; Reddit; Mental Health; Suicide Risk Assessment; C-SSRS; Medical Knowledge Bases; Perceived Risk Measure; Semantic Social Computing

ACM Reference Format:
Manas Gaur, Amanuel Alambo, Joy Prakash Sain, Ugur Kursuncu, Krishnaprasad Thirunarayan, Ramakanth Kavuluru, Amit Sheth, Randon S. Welton, and Jyotishman Pathak. 2019. Knowledge-aware Assessment of Severity of Suicide Risk for Early Intervention . In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3308558.3313698

1 INTRODUCTION

According to recent data from the US Centers for Disease Control and Prevention (CDC), suicide is the second leading cause of death for people aged between 10-34 [45] and fourth leading cause for people aged 35-64, escalating the suicide rate in the US by 30% since

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313698

1999[1]. Suicide Prevention Resource Center in the US[2] reports that 45% of people who committed suicide had visited a primary care provider one to two months before their death. These visits were often scheduled for something other than complaints of depression or suicide, suicidal patients may be too embarrassed to bring up suicide. Clinicians often have no prior warning that the patient is currently suicidal or will be developing significant signs of suicidality. Hence, novel strategies are necessary to proactively detect, assess, and enable timely intervention to prevent suicide[3].

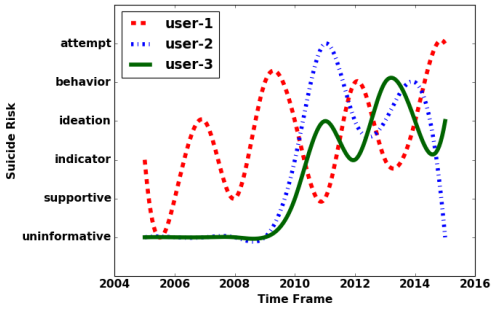

Figure 1: Changing Suicide Risk of 3 Redditors over a period of 11 years

Mental health conditions have been closely linked to suicide [17]. Depression, bipolar and other mood disorders are known to be the main risk factors for suicide, while substance abuse and addiction have been closely linked to suicidal thoughts[4]. SAMHSA[5] reports that people with BPD, Alcoholism, and Drug Addiction are more prone to having suicidal behaviors (e.g., holding gun to the head, driving sharp knife through nerves) and committing suicide. Apart from mental health conditions, there are various other factors exacerbating an individual's urge to commit suicide such as workplace/sexual harassment, religious scripts encouraging self-sacrifice, and heroic portrayal of death in movies. Moreover, popular celebrities who commit suicide can lead to "copycat" suicides or Werther effect [39]. It refers to the contagious influence that a popular figure's suicide can have on an individual, encouraging them to commit suicide. There are several resources for patients to seek help from such as CrisisTextLine, teen line, 7cups.com, imalive.org, and The Trevor Project for LGBTQ. Additional measures are necessary to improve timely intervention [5]. Unobtrusive collection and analysis of social media data can provide a means for gathering insights about an individual's emotions, and suicidal ideation and behavior [33]. A system capable of gleaning digital markers of suicide risk assessment from social media conversations of a patient (see Figure 1) can help a mental health professional (MHP) for making informed decisions as the patients may be reluctant to directly share all the relevant information due to the social stigma associated with mental illness and suicide[22].

There is a significant body of work addressing issues concerning suicide and mental health using social media content. TeenLine, Tumblr, Instagram, Twitter, and Reddit have been common sources of data for research in computational social science [7, 8, 56]. Among these, Reddit has emerged as the most promising one due to the anonymity it affords, its popularity as measured by its content size, and its variety as evident from the diverse subreddits being used for posting that reflects a user's state of mind and mental health disorder, e.g., r/Depression, r/SuicideWatch, r/BipolarSoS. Analysis of the content on Reddit can be leveraged to help an MHP develop an insight into the current situation of an individual, to improve the quality of the diagnosis and intervention strategies if necessary. Shing et al.[54] analyzed the postings of users in SuicideWatch and other related subreddits (e.g., r/bipolarreddit, r/EatingDisorder, r/getting_over_it, and r/socialanxiety) for assessment of suicide risk. The critical opportunity to improve upon these efforts is to utilize reliable domain-specific knowledge sources for understanding the content from a clinical perspective. Specifically, this strategy can augment raw Reddit content to normalize it into a standard medical context and improve the decision-making process of the MHP.

Prior research on suicide risk assessment employs four-label (no risk, low risk, moderate risk, and high risk) classification scheme for categorization of suicidal users [54]. In this research, we provide a C-SSRS-based five-label (supportive, indicator, ideation, behavior, and attempt) classification scheme guided by clinical psychiatrists, which allows the MHP to determine an actionable measure of an individual's suicidality and appropriate care [65]. We compared our 5-label scheme with two other variants: 4-label (indicator, ideation, behavior, and attempt) and (3+1)-label (supportive + indicator, ideation, behavior, and attempt) for monitoring progression and for alerting an MHP as necessary.

Apart from identifying the risk factors of suicide, we can develop approaches to generate answers to the questions from the content in C-SSRS[6], such as (1) Have you wished you were dead or wished you could go to sleep and not wake up? and (2) Have you actually had any thoughts of killing yourself? Our study aims to develop mapping and learning approaches for estimating the suicide risk severity level of an individual, based on his/her posted content [1].

Key Contributions: (1) We develop an annotated gold standard dataset of 500 Reddit users, out of 2181 potentially suicidal users, using their content from mental health-related subreddits. (2) Using domain-specific resources- SNOMED-CT, DataMed, Drug Abuse Ontology (which incorporates DSM-5 [60]) and ICD-10, we created *suicide risk severity lexicon*, curated by MHPs. This enabled us to create a competitive baseline for evaluating our approach. (3) Using four evaluation metrics (graded recall, confusion matrix, ordinal error, and perceived risk measure), we show that the C-SSRS based 5-label classification scheme improves upon the state-of-the-art scheme to characterize suicidality of a user. (4) Our evaluation shows that CNN emerges as a superior model for suicide risk prediction task outperforming the two competing baselines: rule-based and SVM-linear. Technological advancements over the last decade have transformed the health care system with a trend towards real-time monitoring, personal data analysis, and evidence-based diagnosis. Specifically, with the anticipated inclusion of individual's social data and the rapidly growing patient-generated health data [52], MHPs will be better informed about the patient's conditions including their suicidality to enable timely intervention.

[1]https://www.cdc.gov/mmwr/preview/mmwrhtml/mm6217a1.htm
[2]https://www.integration.samhsa.gov/about-us/esolutions-newsletter/suicide-prevention-in-primary-care
[3]https://bit.ly/2QiYqbo
[4]https://www.psychologytoday.com/us/blog/real-healing/201402/suicide-one-addiction-s-hidden-risks
[5]https://www.samhsa.gov/suicide-prevention
[6]https://www.integration.samhsa.gov/clinical-practice/screening-tools#suicide

In Section 2, we review related research. In Section 3, we discuss the resources we use. In Section 4, the critical components of the approach are developed. In Section 5 we give details of experimental design and in Section 6 we discuss our results.

2 RELATED WORK

In this section, we describe prior research related to our study.

2.1 Suicide and Social Media

Jashinsky et al. [28], and Christensen et al. [9] predicted the level of suicide risk for an individual over a period of time using Support Vector Machines (SVM) and the features of Term Frequency-Inverse Document Frequency (TF-IDF), word count, unique word count, average word count per tweet, and average character count per tweet. De Choudhury et al. [17] identified linguistic, lexical, and network features that describe a patient suffering from a mental health condition for predicting suicidal ideation. Analysis of content that contains self-reporting posts on Reddit can provide insights on mental health conditions of users. Utilizing propensity score matching, [17] measured the likelihood of a user sharing thoughts on suicide in the future. Another study from Sueki [58] investigated the linguistic variations among different authors on social media, and observed correlation between suicidal behavior and suicide-related tweets. Furthermore, Cavazos-Rehg et al. [8] performed a qualitative analysis of user's content on Tumblr to better understand discourse of self-harm, suicide, and depression. The study highlights Tumblr as a platform for development of suicide prevention efforts through early intervention. Further, people on social media with mental health conditions, often look for similar people [48].

2.2 Analysis of Suicidal Risk Severity

So far, prior research studied the identification of signals for predicting the suicide risk, mental health conditions leading to suicide [15, 16, 47], psychological state and well-being [50, 51]. Nock et al. [40] reported that ∼9% of people have thoughts of suicide, ∼3% map out their suicidal plans, ∼3% make a suicide attempt and ≤1% people constitute what are known as "suicidal completers". Much information extracted from the content of an individual provide explicit, implicit or ambivalent clues for suicide. These clues can help an MHP assess suicide severity, and better structure the treatment process [11].

Shing et al.[54] used 1.5M posts from 11K users on SuicideWatch subreddit. In the study, experts and crowd-source workers annotated the posts from 245 users using labels defined in [12]. The study evaluates the annotation quality of experts and non-experts and performs risk and suicide screening experiments using linguistic and psycho-linguistic features based on machine/deep learning classifiers. The study fails to bring together different mental health conditions that lead to suicide. Inclusion of supportive users on social media, who are not suicidal, as these constitute the negative samples. Further, the rubric for annotating the dataset was not authoritative, whereas, we utilize C-SSRS endorsed by NIH and SAMHSA.

2.3 Models for Suicide Prediction

In a recent study on predicting suicide attempt in adolescents, Bhat et al. employed deep neural networks for predicting the presence of suicide attempts using >500K anonymized Electronic Health Records (EHR) obtained from California Office of Statewide Health Planning and Development (OSHPD). Through a series of experiments, researchers achieved a true positive rate of 70% and a true negative rate of 98.2% [4]. Another study by Walsh et al. [61] on predicting suicidal attempts using temporal analysis, employs Random Forest (RF) over a cohort of 5167 patients. The study segregates the cohort of patients into 3250 cases and 1917 controls. They achieved an F1-score of 86% with a recall of 95% [61]. The study used binary classification scheme for Electronic Health Records (EHR) dataset, which is not suitable for identifying supportive and indicator users. A transfer learning from social media to EHR can improve its effectiveness [62]. Amini et al. utilized SVM, and decision trees besides RF and Neural Networks (NN), for assessing the risk of suicide in a dataset of individuals from Iran [2]. A recent study by Du et al. [19] used deep learning methods to detect psychiatric stressors leading to suicide. They built binary classifier for identifying suicidal tweets from non-suicidal tweets using Convolutional Neural Networks (CNN). Once suicidal tweets are detected,they performed Named Entity Recognition (NER) using Recurrent Neural Networks (RNN) for tagging psychiatric stressors in a tweet classified as suicidal.

3 BACKGROUND STUDY

We detail the medical knowledge bases underlying the suicide risk severity lexicon used in a baseline (see Section 5.2).

3.1 Domain-specific Knowledge Sources

Medical knowledge bases are resources manually curated by domain experts providing concepts and their relationships for processing the content. As our study aims to assess the severity of at-risk suicidal users, the domain knowledge that corresponds to different levels of suicidality of a patient is crucial. In this work, we employ ICD-10, SNOMED-CT, Suicide Ontology, and Drug Abuse Ontology (DAO) [7] for creating a suicide lexicon to be used in one of our baselines.

Concepts in **SNOMED-CT** are categorized into procedure, observable entity, situation, event, assessment scale, therapy, disorder, and finding and can be extracted using "parents", "children", and "sibling" relationships. For example, Suicide by Hanging [SNOMED ID: 287190007] is a child concept of Suicide [SNOMED ID: 44301001] and sibling concept of Assisted Suicide [SNOMED ID: 51709005], Drug Overdose - Suicide [SNOMED ID: 274228002], Suicide while incarcerated [SNOMED ID: 23546003], and Suicide by self-administered Drug [SNOMED ID: 891003]. **ICD-10** is a medical standard that provides information on patient's health state such as severity, complexity, comorbidities, and complications. Concepts in ICD-10 are categorized into signs, symptoms, abnormal findings, and diagnosis. For example, "suicide attempt" is categorized under a Personal history of self-harm [ICD-10 ID: Z91.5]. It is also categorized under Borderline Personality Disorder [ICD-10 ID: F60.3], Intentional self-harm [ICD-10 ID: X60-X84], and Severe depressive episode with psychotic symptoms [ICD-10 ID: F32.3]. "suicidal ideation" is categorized under Post-traumatic stress disorder[ICD-10 ID: F43.1]. **Suicide Ontology** is an ontology, called "suicideonto"[7] built through text mining and manual curation by domain experts. The ontology contains 290 concepts defining the context of suicide. **Drug Abuse Ontology (DAO)** is a domain-specific hierarchical framework developed by Cameron et al. [7] containing 315 entities (814

[7] https://bioportal.bioontology.org/ontologies/suicideo

instances) and 31 relations defining drug-abuse and mental-health concepts. The ontology has been utilized in analyzing web-forum content related to buprenorphine, cannabis, a synthetic cannabinoid, and opioid-related data [13, 14, 34]. In [21] it was expanded using DSM-5 categories covering mental health and applied for improving mental health classification on Reddit.

3.2 Existing Domain Specific Lexicons

Prior research [6, 38] highlighted the disparity between the informal language used by social media users and the concepts defined by domain experts in medical knowledge bases. Medical entity normalization fills such a gap by identifying phrases (n-grams, or topics) within the content and mapping them to concepts in medical knowledge bases [36]. We use (i) two lexicons, namely, TwADR-L and AskaPatient (see Table 1) to map the social media content to medical concepts [36], and (ii) anonymized and annotated suicide notes made available through Informatics for Integrating Biology and the Bedside (i2b2) challenge to identify content with negative emotions (see Table 2).

Table 1: Existing domain specific lexicons used in this study

Lexicon	#SNOMED Concepts	Max. #phrases per Concept	Sample SNOMED to informal terms mapping
TwADR-L	2172	36	**SNOMED Concept:** Acute depression **Phrases:** 'acute depression', 'just want to finally be happy', 'hated my life', 'depression'
AskaPatient	3051	56	**SNOMED Concept:** Anxiety **Phrases:** 'anxious', 'anxiety issues', 'anxiety', 'anzity'

TwADR-L [36] maps medical concepts in SIDER[8] to their corresponding informal terms used in Twitter. The lexicon has 2172 medical concepts, each of which has up to 36 informal Twitter terms. Each informal term is assigned a single medical concept. AskaPatient[9] [36] maps informal terms from AskaPatient web forum to medical concepts in SNOMED-CT and Australian Medical Terminology [35]. Since this lexicon was created from a web forum, it is more informative compared to TwADR-L. i2b2 Suicide Notes is a dataset generated as a part of the emotion recognition task in 2011 [63]. We have ~2K suicide notes annotated for different emotions, and of them with negative emotions were removed, resulting in 817 suicide notes (see Table 2 for examples).

Table 2: Suicide notes aggregated by emotion labels defined in i2b2.

Emotion Label	#Suicide Notes	Example
abuse	9	My son got married to a horrible woman who does not care curses swears and pushes me around.
anger	69	I have no idea why I could let one person hurt me I loved you for so long but I think I hate you now.
fear	25	In this case you would finally meet defeat so crushing will drain strip you off your courage and hope
guilt	208	God is just and it is true that I am a no good but God will see all that I had to pass through
hopelessness	455	Dear Jane Dont think to badly of me for taking this way out but I am frustrated by taking so much pain
sorrow	51	My heart has been hurt hard and grieving.

[8]http://sideeffects.embl.de
[9]https://www.askapatient.com/

3.3 Suicide Risk Severity Lexicon

Besides the existing lexicons (see Section 3.2), we have built a comprehensive lexicon containing terms related to each level of suicide risk severity (see Table 3). The lexicon was created using Suicideonto[10], DSM-5 [21], and concepts in i2b2 suicide notes. Besides these four severity levels, we consider a separate class of "supportive" users who are not suicidal, but use a similar language. The lexicon was created using the aforementioned medical knowledge bases and slang terms from DAO. The lexicon was validated by the domain experts, and used for annotation and for our baseline (see Section 5.2).

Table 3: Suicide Risk Severity lexicon

Suicide Class	# Terms in a class	Examples
Indicator	1535	Pessimistic character, Suicide of relative, Family history of suicide
Ideation	472	Suicidal thoughts, Feeling suicidal, Potential suicide care
Behavior	146	Planning on cutting nerve, Threatening suicide, Loaded Gun, Drug-abuse
Attempt	124	Previous known suicide attempt, Suicidal deliberate poisoning, Goodbye Attempted suicide by self-administered drug, Suicide while incarcerated.

3.4 Columbia Suicide Severity Rating Scale

Each C-SSRS severity class (ideation, behavior, or attempt) is composed of a set of questions that characterize the respective category. Responses to the questions across the C-SSRS classes eventually determine the risk of suicidality of an individual [44]. One of the challenges researchers face when it comes to dealing with social media content is the disparity in the level of emotions expressed. Since the C-SSRS was originally designed for use in clinical settings, adapting the same metric to a social media platform would require changes to address the varying nature of emotions expressed. For instance, while in a clinical setting, it is typically suicidal candidates that see a clinician; on social media, non-suicidal users may participate to offer support to others deemed suicidal. To address these factors, we have defined two additional classes to the existing C-SSRS scale with three classes. We have provided the description of the five classes in Section 4.4.1.

3.5 Suicide Seed Terms

Not all users in subreddit SuicideWatch (SW) are suicidal. We identify suicidal candidates in subreddit SW by looking into the nature of words used in users' posts.

We analyzed the content of SW subreddit against Zipf-Mandelbrot law to precisely identify terms that are 'prominent' in the online discussion of suicidal thoughts balancing frequency and relevance. In Figure 2, the cyan line follows Zipf-distribution while the green line follows the

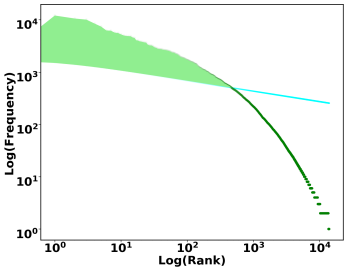

Figure 2: Zipf-Mandelbrot law over SW content for identifying prominent suicide seed terms. Highlighted is the selected region.

[10]https://bioportal.bioontology.org/ontologies/suicideo

Mandelbrot distribution. We are particularly interested in the region of the graph shaded in the top left corner off the cut-off mark between the two lines (light green).

Table 4: Suicide seed terms selected through Figure 2. W: Word, Ph:Phrase, SW: SuicideWatch.

W/Ph.	Freq. W/Ph in SW	W/Ph.	Freq. W/Ph in SW
Commit Suicide	539	Death	215
Hopelessness	130	Gun	577
Sadness	453	Isolation	104
Therapist	194	Trans[11]	96
Kill	577	Sleep	193

This region represents terms in the document that are frequently used by users while also having higher ranks (numerically small values). This effectively eliminates terms that are simply frequently used in the document, but have low ranks. Identified terms were validated by clinical psychiatrists and a curated list of 339 words with a cut-off frequency of 725. A sample list of 10 words is shown in Table 4.

Having identified the suicidally prominent terms, and in conjunction with negation detection technique, we filtered noisy users (users who don't 'positively' use one or more of these terms in their posts) and identified prominently suicidal users.

3.6 Embedding Models

Word embeddings are a set of techniques used to transform a word into a real-valued vector. This allows words with similar meanings to have similar representations and be clustered together in the vector space. Normally, we either generate domain-specific word embeddings local to our problem or employ general purpose word embeddings [32]. We utilize embeddings from ConceptNet[12] (vocabulary= 417193, dimension= 300), a multi-lingual knowledge graph created from expert sources, crowd-sourcing, DBpedia, vocabulary derived from Word2Vec[13] [49], and Glove[14] [43] [57].

4 DATASET CREATION AND ANALYSIS

In this section, we analyze the data, its features and our procedure to identify a small cohort of Redditors that resemble potential candidates for suicidal users (see Figure 3). Our dataset comprises 270,000 users with 8 Million posts from 15 mental health related subreddits; r/StopSelfHarm (SSH), r/selfharm (SLF), bipolar (r/bipolar (BPL), r/BipolarReddit (BPR), r/BipolarSOs, r/opiates (OPT), r/Anxiety (ANX), r/addiction (ADD), r/BPD, r/SuicideWatch (SW), r/schizophrenia (SCZ), r/autism (AUT), r/depression (DPR), r/cripplingalcoholism (CRP), and r/aspergers (ASP) [21]. We used 93K users who actively participated in the SuicideWatch subreddit providing 587466 posts. To further enrich our dataset, we gathered the posts of these users in the remaining 14 subreddits. The timeframe of our dataset is between 2005 and 2016.

4.1 Potential Suicidal Redditors

Subreddit "SuicideWatch" (SW) had nearly 93K redditors as of 2016. To create a representative sample dataset containing users at five-levels of suicide risk, we used seed terms generated using

[11]https://bit.ly/2NEK9bc
[12]http://conceptnet.io
[13]https://code.google.com/archive/p/word2vec/
[14]https://nlp.stanford.edu/projects/glove/

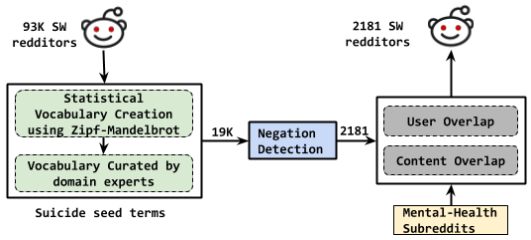

Figure 3: Procedure for generating the annotated dataset. Only 9 mental health subreddits were considered because of significant content overlap (see Figure 4)

Zipf-Mandelbrot (see Section 3.5). We obtained a working set of 19K redditors using such terms. Next, we employed negation detection procedure (see Section 4.2) to eliminate non-suicidal users. Finally, we obtained 2181 users who are potentially suicidal and had participated in other mental health subreddits. For referencing, we denote these users and their content in SW as U^{SW}.

4.2 Negation Detection

Negation detection is a crucial part as the presence of negated sentences can confound a classifier [23]. For example, *I am not going to end my life because I failed a stupid test* is not suicidal, whereas *My daily struggles with depression have driven me to alcohol* reflects user's mental health. The former sentence can give false positive, if we just extract 'going to end my life' as a precursor to a suicide attempt. We employ a negation detection tool and probabilistic context-free grammar that supports negation extraction and negation resolution to improve classifier performance [23].

4.3 User and Content Overlap

As individuals form communities based on shared topics of interest related to mental health conditions [59, 64] in different subreddits, we performed user and content overlap analysis between SW and other mental health subreddits to enrich the contents of users. This analysis provides deeper insight into how potentially suicidal users communicate on problems including causes, symptoms, and treatment solutions. Through user overlap we infer the population level similarity between a mental health subreddit and SW, whereas using content we quantify overlap in context for each user. We calculated the user overlap through the intersection of the users in U^{SW} and i^{th} mental health subreddit (U^{MH_i}). Content overlap was calculated using a cosine similarity measure through domain-specific lexicon, LDA2Vec [37] and ConceptNet.

We leverage the quantified similarity of suicide-related topics between content of the users in U^{SW} and other subreddits (U^{MH_i}), to append the content of users U^{SW}. This procedure will contribute to the holistic nature of the content and enable more discriminative features in the classifier. For example, a post in SW: *I dont think Ive thought about it every day of my entire life. I have for a good portion of it, however, my boyfriend may be able to determine whether I'm worth his time* seems to imply that the user is *non-suicidal*. However, after appending following post taken from "depression" subreddit: *Having a plan for my own suicide has been a long time relief for me as well. I more often than not wish I were dead*, we notice that the user has *suicidal ideations*. As the content in Reddit posts contain slang terms for medical entity, we employed a normalization procedure using standardized lexicons to provide a cleaner interpretation of a

patient's condition, meaningful to a mental health professional or clinician. To perform medical entity normalization, we utilize three lexicons (see Section 3.2), namely, i2b2, TwADR, and AskaPatient, which were created from Medical Records, Twitter, and Web Forum respectively. The normalization used string match.

Content overlap using TwADR-L and AskaPatient: We trained an LDA model with topic coherence over the normalized content to find coherent topics for SW subreddit. Subsequently, using the trained LDA model of SW content, we generated two sets of Topics at user level for U^{SW}, and U^{MH_i}. The topical similarity (TS) was calculated between topics of U^{SW}, and those U^{MH_i}. For the calculation of TS, the user should be present in U^{SW} and U^{MH_i} and should have an average similarity greater than 0.6 (defined empirically). We formalized TS as;

If $u \in U^{SW} \cap U^{MH_i}$, then

$$\left(\frac{1}{|U^{SW}|-1}\sum_{v \in U^{SW} \setminus \{u\}} \cos(\vec{u}^{MH_i}, \vec{v}^{SW})\right) > 0.6 \quad (1)$$

In the above equation, topic vector of users in U^{MH_i} is denoted as $\vec{u}^{MH_i}$ and that of U^{SW} as $\vec{v}^{SW}$.

The resultant column vector contains the similarity between MH_i and SW and has a dimension of **14x1**. Equation 1 used with for two lexicons: TwADR-L and AskaPatient for abstracting the concepts within the reddit posts. To create each column vector, we trained two topic models because TwADR-L lexicon has been created using Twitter and AskaPatient Lexicon using Forum content.

Content Overlap using i2b2: Table 2 shows 6 emotion labels in i2b2 suicide notes dataset. For quantifying the user's content with appropriate emotion label (Table 2), we generated embeddings of content in SW and other MH subreddits for each user using ConceptNet embedding model. We also generated the representations of the emotion labels of the suicide notes through concatenation and dimensionality reduction of the embedding vectors of their corresponding suicide notes [20]. Then, we performed the cosine similarity measure over: (i) embeddings of content from mental health subreddits for each user and the emotion labels, and (ii) embeddings of content from the SW subreddit for each user and the emotion labels. We formalize similarity between i2b2 label and user content embedding as follows:

$$UL(SW, L) = cos(\vec{u}, \vec{l}), u \in U^{SW}, l \in L \quad (2)$$

where UL(SW,L) stores the similarity values between the users in U^{SW} and the emotions labels in i2b2 (L), forming a matrix of dimension 2181 x 6. It is calculated using cosine similarity between the vector of a user ($u \in U^{SW}$) and an emotion label ($l \in L$). Each row of the matrix represents the similarity value for a user embedding generated from all their posts against embedding of each label in i2b2 generated from suicide notes. A similar matrix (using Equation 2) is created for users in other mental health subreddits ($u \in U^{SW} \cap U^{MH_i}$) and emotion labels L. We denote such a matrix as $UL(MH_i, L)$ of dimensions 2181 x 6. $UL(SW, L)$ and $UL(MH_i, L)$ are interpreted as matrices showing to what degree users' contents are close to six emotions. Thereafter, we generate a similarity score ($SS(MH_i, SW)$) as a product of $UL(SW, L)$ and transpose of $UL(MH_i, L)$. Formally we define it as:

$$\text{SS}(MH_i, SW) = \frac{\sum_{\vec{u} \in UL(SW,L)} \sum_{\vec{v} \in UL(MH_i,L)^T} \vec{u}.\vec{v}}{(|U^{SW}|-1)^2} \quad (3)$$

If the users are in U^{SW} and U^{MH_i}, their content will be appended to SW from MH_i only if the content overlap is greater than 0.6 in Equations 1 and 3. The procedure repeated over all MH subreddits and we obtain results shown in Figure 4.

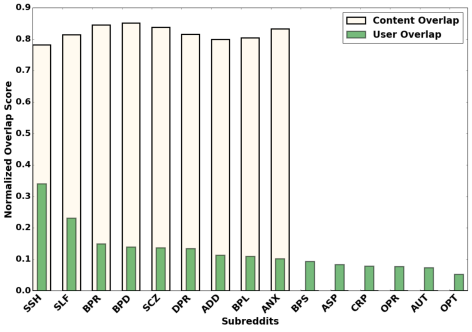

Figure 4: User Overlap and Content Overlap based quantification of influence of other mental health related subreddit to SW. Subreddits *SSH* and *SLF* have the highest content overlap with SW followed by *BPR* and *BPD*.

4.4 Gold Standard Dataset Creation

We describe different classes of suicidality, characterizing users who suffer from mental health conditions or involve themselves in a supportive role on social media. Further, we describe annotated dataset with examples and annotation evaluation using Krippendorff.

4.4.1 5-labels of Suicide Risk Severity: C-SSRS begins with ***Suicidal Ideation*** (*ID*), which is defined as thoughts of suicide including preoccupations with risk factors such as loss of job, loss of a strong relationship, chronic disease, mental illness, or substance abuse. This category can be seen to escalate to ***Suicidal Behavior*** (*BR*), operationalized as actions with higher risk. A user with suicidal behavior confesses active or historical self-harm, or active planning to commit suicide, or a history of being institutionalized for mental health. Actions include cutting or using blunt force violence (self-punching and head strikes), heavy substance abuse, planning for suicide attempt, or actions involving a means of death (holding guns or knives, standing on ledges, musing over pills or poison, or driving recklessly). The last category, an ***Actual Attempt*** (*AT*), is defined as any deliberate action that may result in intentional death, be it a completed attempt or not, including but not limited to attempts where a user called for help, changed their mind or wrote a public "good bye" note. When reviewing users' risk levels for social media adaptation, two additional categories were added to define user behaviors less severe than the above categories.

The first addition was a ***Suicide Indicator*** (*IN*) category which separated those using at-risk language from those actively experiencing general or acute symptoms. Oftentimes, users would engage in conversation in a supportive manner and share personal history while using at-risk words from the clinical lexicon. These users might express a history of divorce, chronic illness, death in the family, or suicide of a loved one, which are risk indicators on the C-SSRS, but would do so relating in empathy to users who expressed ideation or behavior, rather than expressing a personal desire for self-harm. In this case, it was deemed appropriate to flag such users as *IN* because while they expressed known risk factors that could be monitored they would also count as false positives if they were accepted as individuals experiencing active ideation or behavior.

The second additional category was named as ***Supportive*** (*SU*) and is defined as individuals engaging in discussion but with no language that expressed any history of being at-risk in the past

or the present. Some identified themselves as having background in mental health care, while others did not define their motive for interacting at all (as opposed to a family history). Since posting on Reddit is not itself a risk factor, so we give these users a category with even lower risk than those expressing support with a history of risk factors. Any use of language such as a history of depression, or "I've been there" would re-categorize a user as exhibiting suicidal indicator, ideation, or being at greater risk, depending on the language used. These new categories for an adapted C-SSRS should help account for those who communicate in suicide-related forums but were at a low or undefined risk.

4.4.2 Description of the Annotated Dataset: For the purpose of annotation, we randomly picked 500 users from a set of 2181 potential suicidal users. In the annotated data, each user on an average has 31.5 posts within the time frame of 2005 to 2016.

The annotated data comprises of 22% supportive users, 20% users with some suicidal indication but cannot be classified as suicidal, 34% users with suicidal ideation, 15% users with suicidal behaviors, and 9% users have made an attempt (success or fail) to commit suicide. Supportive users constitutes 1/5th of the total data size and prior studies have ignored them. Table 5 shows posts from redditors and their associated suicide risk severity level. To identify which mental health subreddits (except SW) contributed most to suicidality, we mapped potential suicidal Redditors to their subreddits (see Figure 5).

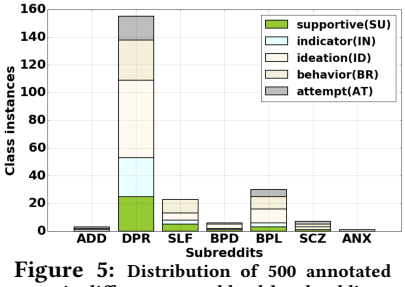

Figure 5: Distribution of 500 annotated users in different mental health subreddits

Table 5: Paraphrased posts from candidate suicidal redditors and associated suicide risk severity level

Always time for you to write your happy ending .... doesnt need to be spelled out with alcohol and Xanax.... keep an open mind	SU
Ive never really had a regular sleep schedule....no energy to hold a conversation....no focus on study....barely eat and sleep....fluffy puppy dog face	IN
Sometimes I literally cant bear to move....my depression....since I was 14....suffering rest of my life....only Death is reserved for me.	ID
Driving a sharp thing over my nerve. Extreme depression and loneliness.... worthless excuse for a life....used everything from wiring to knife blades	BR
I am going to off myself today...loaded gun to my head..determined....huge disappointment....screwed family life....breaks my heart everyday.	AT

4.4.3 Evaluation of Annotation: Four practicing clinical psychiatrists were involved in the annotation process. Each expert received 500 users dataset comprising of 15755 posts. We perform two annotation analysis defined for ordinal labels: **(1)** A pair-wise annotator agreement using Krippendorff metric (α) to identify the annotator with highest agreement with others, **(2)** An incremental group wise annotator agreement to find the robustness of the earlier annotator [55]. For group wise agreement, we denote a set of annotators as G with cardinality ($|G|$) range from 2 to 4. α is calculated as $1 - (\frac{D_o(A_j,S)}{D_e})$, where $D_o(A_j, S)$ is observed disagreement and D_e is expected disagreement. The pairwise annotator agreement is a subset of groupwise and we formally define it as:

Table 6: (left). Pairwise annotator agreement, (right). Group wise annotator agreement. A,B,C,and D are annotators.

	B	C	D
A	0.79	0.73	0.68
B	-	0.68	0.61
C	-	-	0.65

	B	B&C	B&C&D
A	0.79	0.70	0.69

$$D_o(A_j, S) = \frac{1}{N.|S|} \sum_{i=1}^{N} \sum_{m \in S} |A_j^i - S_m^i|^2, S \subset G \setminus \{A_j\} \quad (4)$$

$$D_e = \frac{2}{N \cdot |G|(|G|-1)} \sum_{i=1}^{N} \sum_{m,q \in G, m \neq q} |G_m^i - G_q^i|^2 \quad (5)$$

where A_j is the annotator having highest agreement in pairwise α. S is the subset of a group of annotators G that excludes A_j. G_m^i and G_q^i represents the two annotators m and q within the group G^i. i is the index over all the users in the dataset. Results of pairwise and group wise annotators agreement is in Table 6. We observe a substantial agreement between the annotators[15].

5 EXPERIMENTAL DESIGN

5.1 Characteristic Features

Prior research has shown the importance of psycholinguistics, lexical, syntactic, and emotion features in enhancing the efficacy of the classifier [24, 46]. We further improve our feature set with information provided by Reddit. In training our models we used *AFINN*[16], which is a list of words scored for sentiment, emotions, mood, feeling, or attitude. Posts on Reddit may have nearly equal number of upvotes and downvotes making them controversial. We computed *controversiality score (CScore)* as the ratio of the maximum value of the difference, between *upvotes* and *downvotes*, and 1, over *totalvotes*.

$$\text{CSscore} = \max\left(1, \frac{\#upvotes - \#downvotes}{\#totalvotes}\right)$$

We factored in *Intra-Subreddit Similarity* with and without nouns and pronouns as a measure of content similarity of posts between a user and others in a subreddit. To determine the level of personal experience in the social media text, we utilize *First Person Pronouns Ratio* that measures the extent to which a Redditor talks about his/her own experience compared to other Redditors' experience [10]. We used *Language Assessment by Mechanical Turk (LabMT)*, a list of 10,222 words with happiness, rank, internet usage scores, employing strict match and soft match with Reddit posts [18]. On social media, readability is an important factor. We use *height of the dependency parse tree* to measure readability, with parse tree height being proportional to readability [25]. We employ *maximum length of verb phrase* [26] to capture suicidality of individuals. Similarly, *number of pronouns* was used to determine whether they are sharing a direct experience or second hand experience [42]. The value of this feature was high for users classified as supportive or indicator, as these users usually help others. Moreover, *number of sentences* and *number of definite articles* are also discriminative [27].

[15] http://homepages.inf.ed.ac.uk/jeanc/maptask-coding-html/node23.html
[16] http://neuro.imm.dtu.dk/wiki/AFINN

5.2 Baselines

In this study, we use two baselines; (1) 4-class scheme for predicting the suicide risk [54] (2) an empirical baseline based on the suicide risk severity lexicon. We provide details of our lexicon-based empirical baseline. Suicide lexicon developed as a part of the study for initial filtering of users and annotation process is a suitable resource for a baseline. This baseline is a rule-based model for classifying a user based on a strict and soft match criteria according to presence of a concept in the user's content and the suicide risk severity lexicon. For a competitive baseline, we compared this baseline with word-embedding and TF-IDF based approaches for suicide classification [29]. As we also experimented with word-embedding models trained over suicide and non-suicide related content, using compositions of word vectors [3, 32, 41], the baseline based on suicide risk severity lexicon outperformed these competitive approaches.

5.3 Convolutional Neural Network

We have implemented a convolutional neural network (CNN) as proposed in [30] for our contextual classification task [53].

The model takes embeddings of user posts as input and classifies into one of the suicide risk severity levels. We combine embeddings of posts for each user through concatenation, and pass into the model.

$$posts_u = post_{u,1} \bigoplus post_{u,2} \bigoplus ..post_{u,p}.. \bigoplus post_{u,P} \quad (6)$$

$$post_{u,p} = \vec{v}_{u,p,1} \bigoplus \vec{v}_{u,p,2} \bigoplus ..\vec{v}_{u,p,w}.. \bigoplus \vec{v}_{u,p,W} \quad (7)$$

Here $\bigoplus$ represents the concatenation operation of P posts of user u, where each post p of user u ($post_{u,p}$) is the concatenation of vectors of each word w ($\vec{v}_{u,p,w}$) where W is the total number of words in a post. Embeddings of the posts for each user ($posts_u$) have variable length. Hence, we use minimum length padding to make the dimensions of the representations uniform. The model has a convolution layer with filter window {3, 4, 5} and 100 filters for each. After getting the convoluted features, we apply max-pooling and concatenate the representative pooled features. We pass the pooled features through a dropout layer with dropout probability of 0.3, followed by an output softmax layer. The learning rate was set to 0.001 with adam optimizer [31]. While training the model, we have used mini batch of size 4 and trained for 50 epochs. CNN's performance is compared and evaluated in Section 6.

5.4 Evaluation Metrics

We alter the formulation of False Positive (FP) and False Negative (FN) to better evaluate the model performance. FP is defined as the ratio of the number of times the predicted suicide risk severity level (r') is greater than actual level (r^o) over the size of test data (N_T). FN is defined as the ratio of the number of times r' is less than r^o over N_T. Since the numerators of FP and FN involves comparison between r' and r^o suicide risk severity levels, we termed the metrics as graded precision, and recall as graded recall. Ordinal Error (OE) is defined as the ratio of the number of samples where difference between r^o and r' is greater than 1. In our study it represents the model's tendency to label a person as having no-severity or low degree of severity, when he/she is actually at risk.

We formally define FP, FN, and OE as:

$$FP = \frac{\sum_{i=1}^{N_T} I(r'_i > r^o_i)}{N_T},\ FN = \frac{\sum_{i=1}^{N_T} I(r^o_i > r'_i)}{N_T},\ OE = \frac{\sum_{i=1}^{N_T} I(\Delta(r^o_i, r'_i) > 1)}{N_T}$$

where $\Delta(r^o_i, r'_i)$ is the difference between r^o_i and r'_i. r'_i and r^o_i are the predicted and actual response for i^{th} test sample.

5.4.1 Perceived Risk Measure (PRM): It is defined to better characterize the difficulty in classifying a data item while developing a robust classifier in the face of *difficult to unambiguously* annotate datasets. It captures the intuition that if a data item is difficult for human annotators to classify unambiguously, it is unreasonable to expect a machine algorithm to do it well, or in other words, misclassifications will receive reduced penalty. On the other hand, if the human annotators are in strong agreement about a classification of a data item, then we would increase the penalty for any misclassification. This measure captures the biases in the data using disagreement among annotators. Based on this intuition, we define PRM as the ratio of disagreement between the predicted and actual outcomes summed over disagreements between the annotators multiplied by a reduction factor that reduces the penalty if the prediction matches any other annotator. We formally define it as;

$$PRM = \frac{1}{N_T} \sum_{i=1}^{N_T} \left(\frac{1 + \Delta(r'_i, r^o_i)}{1 + \sum_{m,q \in G^i, m \neq q} \Delta(G^i_m, G^i_q)} \cdot \frac{\sum_{m \in G^i} I(r'_i = G^i_m)}{|\mathbf{G^i}|} \right) \quad (8)$$

Where the denominator is the disagreement between G^i_m and G^i_q annotators summed over all annotators in a group $\mathbf{G^i}$ (notations are same as in equation 5). $\frac{\sum_{m \in G^i} I(r'_i = G^i_m)}{|\mathbf{G^i}|}$ is the risk reducing factor calculated as the ratio of agreement of prediction with any of the annotators over the total number of annotators. In cases where r' disagrees with all the annotators in $\mathbf{G}$, the risk reducing factor is set to 1.

6 RESULTS AND ANALYSIS

We evaluate the model performance over different levels of suicide severity. We categorize our experiments into three schemes: **Experiment 1** evaluates the performance of the models over 5 labels (supportive, indicator, ideation, behavior, and attempt); **Experiment 2** evaluates models' performances over 4 labels in which supportive (or negative) samples are removed, and **Experiment 3** comprises labels defined according to 4-label categorization (where supportive and indicator classes are merged into one class : no-risk). Further, for each experiment, the input data is of two forms: **(I1)** Only textual features (TF) represented as vectors of 300 dimensions generated using ConceptNet embeddings , **(I2)** having Characteristics features (CF) (see Section 5.1) and textual features (CF+TF). All experiments were performed with 5 fold hold-out cross-validation. It was defined empirically, observing results at various folds. We show that the proposed 5-label classification scheme has better recall, and the perceived risk measure of the 5-label classification scheme is low compared to other reduced classification schemes. All the experiments have been performed with 5-fold cross validation and results are reported on hold-out test set.

6.1 Experiment 1: 5-Label Classification

For evaluation, we consider five learning models (SVM with Radial Basis Function (SVM-RBF), SVM with Linear Kernel (SVM-L), Random Forest (RF), Feed-Forward Neural Network (FFNN), CNN) that have been used in similar studies (see Section 2.3) over two types of inputs: I1 and I2. For input I1, the baseline is a suicide-lexicon based

classifier which is content-based, and for input I2, the baseline is SVM-L which is the best performing model in Shing et al. [54].

Table 7: Experiment with 5-label Classification

Approach	Input	With Supportive Class			
		Graded Precision	Graded Recall	F-Score	OE
Baseline	text	0.56	0.36	0.44	0.38
SVM-RBF	I1	0.53	0.51	0.52	0.12
	I2	0.57	0.62	0.61	0.12
SVM-L	I1	0.60	0.45	0.52	0.12
	I2	0.77	0.40	0.53	0.09
RF	I1	0.68	0.49	0.57	0.19
	I2	0.62	0.45	0.52	0.11
FFNN	I1	0.45	0.59	0.51	0.15
	I2	0.52	0.63	0.57	0.12
CNN	I1	0.71	**0.60**	0.65	0.10
	I2	0.70	**0.59**	0.64	0.09

Input type I1: Table 7 reports that CNN outperforms the baseline with an improvement of 40% in precision, 5% in recall, and 25% in F-score. Based on small improvement in recall, it is inferred that CNN has a tendency to predict a low risk level (e.g; Supportive) for a user who has an observed high risk (e.g; Behavior). SVM-RBF and SVM-L show an improvement in precision compared to baseline; however, there is 12% and 27% reduction in recall respectively. Further, RF showed a 40% increase in precision at a cost of 16% reduction in recall. On the contrary, FFNN performed relatively well in comparison to baseline concerning recall. Hence, at a fine-grained level of comparison, CNN outperforms the baseline with a considerable improvement in precision and recall. To better characterize the comparison between the models, we analyze them using OE. Such a measure is coarse-grained and focuses more on FN as opposed to acceptable FP. Based on Table 7, we observed that CNN showed the least error based on OE calculation, reporting that 1% of the people have been predicted with a severity level of difference 2 or more compared to observed. Such a measure of evaluation is important because it ignores the biases in the gold standard data. As a result, CNN correctly predicted the severity of 90% of users.

Input Type I2: In comparison to the second baseline, CNN outperforms SVM-L with an improvement of 32% in recall with reduction of 10% in precision. We infer from Table 7 that SVM penalized false positives more than false negatives because of its linearity and i.i.d (independent, identically distributed)[17] assumptions. Whereas, CNN's convoluted representation ignores i.i.d assumptions, the non-linearity induced by ReLU tries to balance FP and FN. It can be seen from recall of SVM-RBF for I2 which is higher than SVM-L. However, SVM-RBF fails to balance FP and FN because of i.i.d considerations. Further, from column OE in Table 7, we infer that CNN predicted a suicidality level >1 compared to observed, for 9% of the users, whereas SVM-L did for 10% of the users.

6.2 Experiment 2: 4-label Classification

To evaluate the models over 4-label classification scheme, we use the same approach as applied in Experiment 1 for the purpose of consistency. In addition, in this experiment, the baseline model created over suicide lexicon disregards supportive labels.

[17] https://bit.ly/2Rw9i5Z

Table 8: Experiment with 4-label Classification

Approach	Input	Without Supportive Class			
		Graded Precision	Graded Recall	F-Score	OE
Baseline	text	0.43	0.57	0.49	0.20
SVM-RBF	I1	0.63	0.47	0.54	0.12
	I2	0.66	0.59	0.62	0.12
SVM-L	I1	0.62	0.53	0.57	0.12
	I2	0.68	0.57	0.61	0.09
RF	I1	0.67	0.41	0.51	0.22
	I2	0.64	0.47	0.54	0.18
FFNN	I1	0.63	0.58	0.60	0.15
	I2	0.67	0.62	0.64	0.12
CNN	I1	0.72	**0.59**	0.65	0.11
	I2	0.70	**0.57**	0.62	0.1

Observing Tables 7 and 8, there is a noticeable improvement in the precision of the models due to reduction in the degree of freedom of the outcome variable (removal of supportive class). Moreover, Tables 7 and 8 show the reduction in recall and an increase in OE. Hence, 5-label scheme supports lower OE for best performing model than does 4-label scheme.

Input Type I1 and Input Type I2: For the content-based input, all the models outperform the baseline in terms of precision, however, only CNN model outperforms baseline in terms of recall. Interestingly, there is a decrease in the recall of the models with non-linear kernel from 5-label to 4-label classification scheme; yet, there is a marginal increase in true positives of SVM-L. It can be inferred that SVM-L is vulnerable to predicting some of indicator users as supportive and ideation users as indicator in experiment 1. However, CNN was able to identify supportive users and most of the classification was centered around ideation and indicator levels; 4-label scheme does not bring in major change in OE for CNN.

6.3 Experiment 3: 3+1 Classification

In this classification scheme, we collapsed the supportive and indicator classes into a common class: "control group". It allows us to create the classification structure as defined in [12]. For this experiment, we considered two top performing models from previous experiments: SVM-L and CNN.

Table 9: Experiment with 3+1-label Classification

Approach	Input	Collapsed Supportive and Indicator Class			
		Graded Precision	Graded Recall	F-Score	OE
SVM-L	I1	0.81	0.54	0.65	0.12
	I2	0.74	0.54	0.63	**0.09**
CNN	I1	0.83	**0.57**	0.676	0.07
	I2	0.85	**0.57**	0.68	**0.06**

Input type I1 and I2: Using such a classification scheme (see Table 9), we observe a significant improvement in precision of SVM-linear and CNN in comparison to previous experiments. Apart from the decrease in the degree of freedom of outcome, the model tries to predict the supportive+indicator and ideation classes as opposed to "behavior" and "attempt". Since supportive+indicator and ideation classes are in majority, they boost the precision of the model. However, the model shows a reduction in recall in this scheme compared to 5-label or 4-label classification scheme. Table 10 shows reduction in OE for CNN from 0.1 to 0.07 for I1 and 0.09 to 0.06 for I2 compared to 5-label classification. It is because 3+1 classification

scheme forces the model to compromise with the popular classes and affect the selection of suitable class. Moreover, through our 5-label classification scheme, we achieved an improvement of 4.2% in graded recall over the (3+1) scheme (see Tables 7 and 9).

6.4 5-label Confusion Matrix Analysis

In this evaluation metric we categorize our suicidality labels into two groups; (1) No-Treatment Groups: Supportive and Indicator User, (2) Treatment Groups: Ideation, Behavior, Attempt.

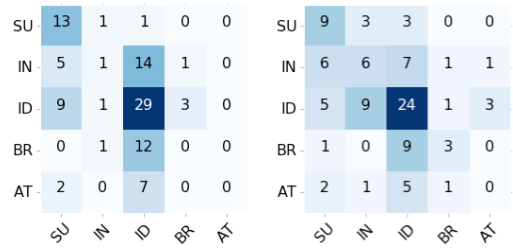

Figure 6: Confusion Matrix of 5-label scheme. (left) CNN, and (right) SVM-L. Y-Axis: True Level, X-Axis: Predicted Level

From Figure 6, out of 36 No-treatment users, CNN correctly classifies 20 users (56%) whereas SVM-L correctly classifies 22 users (61%). However, observing a larger 64 Treatment users, CNN correctly classifies 51 users (80%) whereas SVM-L correctly classifies 46 users (72%). Hence, CNN provides more suitable class for the users compared to SVM-L.

6.5 4-Label Confusion Matrix Analysis

Under the 4-label classification scheme, the No-treatment population involves users annotated as indicator whereas Treatment population contains users annotated as ideation, behavior and attempt. From Figure 7, we observe that CNN correctly classifies 59 out of 64 users (92%) annotated under Treatment whereas SVM-L classifies 53 out of 64 users (83%).

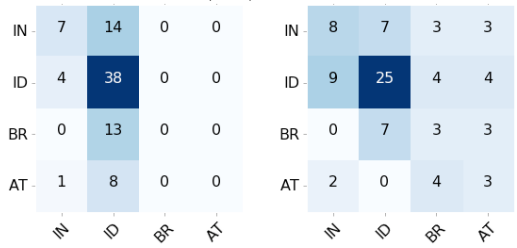

Figure 7: Confusion Matrix of 4-label scheme. (left) CNN, and (right) SVM-L

6.6 3+1 Label Confusion Matrix Analysis

There are 36 users under No-Treatment (supportive and indicator (SU+IN)) group and 64 users under Treatment group. Based on figure 8, we noticed that CNN correctly classifies 26 out of 36 (72%) No-Treatment users whereas SVM-L scored 16 out of 36 (44%). Further, CNN and SVM-L recognized 39 users (61%) and 46 users (72%) in the Treatment group. The decrease in CNN from 80% (5-label) to 61% is attributed to the increase in attempt, behavior, and ideation users classified as No-Treatment.

However, there was no change for SVM-L. But, on comparing 5-label and 3+1 label classification schemes, we observed that collapsing of the supportive and indicator classes can lead to increase in the false positive as SVM-L predicts them as behavior and attempt. There is a reduction in the true positive score for predictive

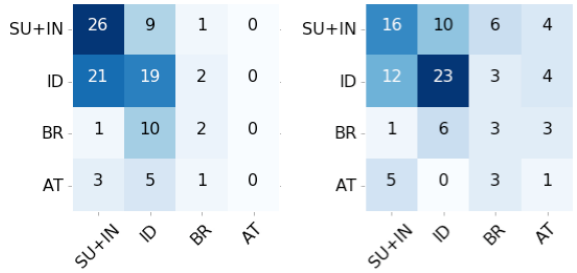

Figure 8: Confusion Matrix of (3+1)-label scheme. (left) CNN, and (right) SVM-L

and actual ideation classes, and users marked as "attempt" have been classified as "supportive and indicator (SU+IN)". As a result, the false negatives of the models have increased. Although this analysis proves the efficacy of 5-label classification over 3+1, CNN being a conservative model, there is a possibility of annotator bias in the data. So below, we perform PRM analysis of SVM-L and CNN over 2 classification schemes: 5-label and 3+1 label.

6.7 Perceived Risk Measure Analysis

Table 10: PRM based comparison of classification schemes

Scheme	Models	PRM
5-Label	CNN	0.14
	SVM-L	0.61
(3+1) Label	CNN	0.16
	SVM-L	0.54

On analyzing models behavior using PRM (Equation 8), Table 10 shows that there is a 12.5% difference between 5-label and 3+1 label classification schemes. Results can be interpreted as: For CNN under 5-label, there is 14% chance that model will provide an outcome that disagrees with every annotator, whereas, for (3+1)-label, it is 16%. Further, we observe that SVM-L has a high risk score compared to CNN in both classification schemes.

7 CONCLUSION

In this study, we presented an approach to predict severity of suicide risk of an individual using Reddit posts, which will allow medical health professionals to make more informed and timely decisions on diagnosis and treatment. A gold standard dataset of 500 suicidal redditors with varying severity of suicidal risk was developed using suicide risk severity lexicon. We then devised a 5-label classification scheme to differentiate non-suicidal users from suicidal ones, as well as suicidal users at different severity levels of suicide risk (e.g., ideation, behavior, attempt). Our 5-label classification scheme outperformed the two baselines. We specifically noted that CNN provided best performance among others including SVM and Random Forest. We make both the gold standard dataset and the suicide risk severity lexicon publicly available to the research community for further suicide-related research.

ACKNOWLEDGEMENT

We acknowledge partial support from the National Science Foundation (NSF) award CNS-1513721: "Context-Aware Harassment Detection on Social Media", National Institutes of Health (NIH) award: MH105384-01A1: "Modeling Social Behavior for Healthcare Utilization in Depression", and National Institute on Drug

Abuse (NIDA) Grant No. 5R01DA039454-02 “Trending: Social media analysis to monitor cannabis and synthetic cannabinoid use”. Dr. Ramakanth Kavuluru was supported by the NIH National Cancer Institute through grant R21CA218231. Any opinions, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF, NIH, or NIDA.

REFERENCES

[1] Amanuel Alambo, Manas Gaur, Usha Lokala, Ugur Kursuncu, Krishnaprasad Thirunarayan, Amelie Gyrard, Randon S Welton, Jyotishman Pathak, and Amit Sheth. 2019. Question Answering for Suicide Risk Assessment using Reddit. *IEEE International Conference on Semantic Computing 2019* (2019).

[2] Payam Amini, Hasan Ahmadinia, Jalal Poorolajal, and Mohammad Moqaddasi Amiri. 2016. Evaluating the high risk groups for suicide: A comparison of logistic regression, support vector machine, decision tree and artificial neural network. *Iranian journal of public health* 45, 9 (2016), 1179.

[3] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2016. A simple but tough-to-beat baseline for sentence embeddings. (2016).

[4] Harish S Bhat and Sidra J Goldman-Mellor. 2017. Predicting Adolescent Suicide Attempts with Neural Networks. *arXiv preprint arXiv:1711.10057* (2017).

[5] J Michael Bostwick, Chaitanya Pabbati, Jennifer R Geske, and Alastair J McKean. 2016. Suicide attempt as a risk factor for completed suicide: even more lethal than we knew. *American journal of psychiatry* 173, 11 (2016), 1094–1100.

[6] Camille Brisset, Yvan Leanza, Ellen Rosenberg, Bilkis Vissandjée, Laurence J Kirmayer, Gina Muckle, Spyridoula Xenocostas, and Hugues Laforce. 2014. Language barriers in mental health care: A survey of primary care practitioners. *Journal of immigrant and minority health* 16, 6 (2014), 1238–1246.

[7] Delroy Cameron, Gary A Smith, Raminta Daniulaityte, Amit P Sheth, Drashti Dave, Lu Chen, Gaurish Anand, Robert Carlson, Kera Z Watkins, and Russel Falck. 2013. PREDOSE: a semantic web platform for drug abuse epidemiology using social media. *Journal of biomedical informatics* 46, 6 (2013), 985–997.

[8] Patricia A Cavazos-Rehg, Melissa J Krauss, Shaina J Sowles, Sarah Connolly, Carlos Rosas, Meghana Bharadwaj, Richard Grucza, and Laura J Bierut. 2016. An analysis of depression, self-harm, and suicidal ideation content on Tumblr. *Crisis* (2016).

[9] Helen Christensen, Philip Batterham, and Bridianne O'Dea. 2014. E-health interventions for suicide prevention. *International journal of environmental research and public health* 11, 8 (2014), 8193–8212.

[10] Kevin Bretonnel Cohen, Dina Demner-Fushman, Sophia Ananiadou, and Jun-ichi Tsujii. 2016. Proceedings of the 15th Workshop on Biomedical Natural Language Processing.

[11] Glen Coppersmith, Ryan Leary, Eric Whyne, and Tony Wood. 2015. Quantifying suicidal ideation via language usage on social media. In *Joint Statistics Meetings Proceedings, Statistical Computing Section, JSM*.

[12] Darcy J Corbitt-Hall, Jami M Gauthier, Margaret T Davis, and Tracy K Witte. 2016. College students' responses to suicidal content on social networking sites: an examination using a simulated facebook newsfeed. *Suicide and Life-Threatening Behavior* 46, 5 (2016), 609–624.

[13] Raminta Daniulaityte, Robert Carlson, Gregory Brigham, Delroy Cameron, and Amit Sheth. 2015. “Sub is a weird drug:” A web-based study of lay attitudes about use of buprenorphine to self-treat opioid withdrawal symptoms. *The American journal on addictions* 24, 5 (2015), 403–409.

[14] Raminta Daniulaityte, Francois R Lamy, G Alan Smith, Ramzi W Nahhas, Robert G Carlson, Krishnaprasad Thirunarayan, Silvia S Martins, Edward W Boyer, and Amit Sheth. 2017. “Retweet to Pass the Blunt”: Analyzing Geographic and Content Features of Cannabis-Related Tweeting Across the United States. *Journal of studies on alcohol and drugs* 78, 6 (2017), 910–915.

[15] Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th Annual ACM Web Science Conference*. ACM, 47–56.

[16] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. *ICWSM* 13 (2013), 1–10.

[17] Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. ACM, 2098–2110.

[18] Peter Sheridan Dodds, Kameron Decker Harris, Isabel M Kloumann, Catherine A Bliss, and Christopher M Danforth. 2011. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PloS one* 6, 12 (2011), e26752.

[19] Jingcheng Du, Yaoyun Zhang, Jianhong Luo, Yuxi Jia, Qiang Wei, Cui Tao, and Hua Xu. 2018. Extracting psychiatric stressors for suicide from social media using deep learning. *BMC medical informatics and decision making* 18, 2 (2018), 43.

[20] Tianfan Fu, Cheng Zhang, and Stephan Mandt. 2018. Continuous Word Embedding Fusion via Spectral Decomposition. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*. 11–20.

[21] Manas Gaur, Ugur Kursuncu, Amanuel Alambo, Amit Sheth, Raminta Daniulaityte, Krishnaprasad Thirunarayan, and Jyotishman Pathak. 2018. Let Me Tell You About Your Mental Health!: Contextualized Classification of Reddit Posts to DSM-5 for Web-based Intervention. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. ACM, 753–762.

[22] George Gkotsis, Anika Oellrich, Sumithra Velupillai, Maria Liakata, Tim JP Hubbard, Richard JB Dobson, and Rina Dutta. 2017. Characterisation of mental health conditions in social media using Informed Deep Learning. *Scientific reports* 7 (2017), 45141.

[23] George Gkotsis, Sumithra Velupillai, Anika Oellrich, Harry Dean, Maria Liakata, and Rina Dutta. 2016. Don't Let Notes Be Misunderstood: A Negation Detection Method for Assessing Risk of Suicide in Mental Health Records. In *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*. 95–105.

[24] David M Howcroft and Vera Demberg. 2017. Psycholinguistic Models of Sentence Processing Improve Sentence Readability Ranking. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, Vol. 1. 958–968.

[25] Yi-Ting Huang, Meng Chang Chen, and Yeali S Sun. 2018. Characterizing the Influence of Features on Reading Difficulty Estimation for Non-native Readers. *arXiv preprint arXiv:1808.09718* (2018).

[26] Nina Hyams and Kenneth Wexler. 1993. On the grammatical basis of null subjects in child language. *Linguistic inquiry* (1993), 421–459.

[27] Zahurul Islam and Alexander Mehler. 2013. Automatic readability classification of crowd-sourced data based on linguistic and information-theoretic features. *Computación y Sistemas* 17, 2 (2013), 113–123.

[28] Jared Jashinsky, Scott H Burton, Carl L Hanson, Josh West, Christophe Giraud-Carrier, Michael D Barnes, and Trenton Argyle. 2014. Tracking suicide risk factors through Twitter in the US. *Crisis* (2014).

[29] Shaoxiong Ji, Celina Ping Yu, Sai-fu Fung, Shirui Pan, and Guodong Long. 2018. Supervised Learning for Suicidal Ideation Detection in Online User Content. *Complexity* 2018 (2018).

[30] Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* (2014).

[31] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[32] Ugur Kursuncu, Manas Gaur, Usha Lokala, Anurag Illendula, Krishnaprasad Thirunarayan, Raminta Daniulaityte, Amit Sheth, and I Budak Arpinar. 2018. "What's ur type?" Contextualized Classification of User Types in Marijuana-related Communications using Compositional Multiview Embedding. *IEEE/WIC/ACM International Conference on Web Intelligence (WI)* (2018).

[33] Ugur Kursuncu, Manas Gaur, Usha Lokala, Krishnaprasad Thirunarayan, Amit Sheth, and I Budak Arpinar. 2019. Predictive Analysis on Twitter: Techniques and Applications. In *Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining*. Springer, 67–104.

[34] Francois R Lamy, Raminta Daniulaityte, Ramzi W Nahhas, Monica J Barratt, Alan G Smith, Amit Sheth, Silvia S Martins, Edward W Boyer, and Robert G Carlson. 2017. Increases in synthetic cannabinoids-related harms: Results from a longitudinal web-based content analysis. *International Journal of Drug Policy* 44 (2017), 121–129.

[35] Hugo Leroux and Laurent Lefort. 2012. Using CDISC ODM and the RDF Data Cube for the Semantic Enrichment of Longitudinal Clinical Trial Data.. In *SWAT4LS*. Citeseer.

[36] Nut Limsopatham and Nigel Henry Collier. 2016. Normalising medical concepts in social media texts by learning semantic representation. (2016).

[37] Christopher E Moody. 2016. Mixing dirichlet topic models and word embeddings to make lda2vec. *arXiv preprint arXiv:1605.02019* (2016).

[38] Liqiang Nie, Yi-Liang Zhao, Mohammad Akbari, Jialie Shen, and Tat-Seng Chua. 2015. Bridging the vocabulary gap between health seekers and healthcare knowledge. *IEEE Transactions on Knowledge and Data Engineering* 27, 2 (2015), 396–409.

[39] Thomas Niederkrotenthaler, Arno Herberth, and Gernot Sonneck. 2007. The" Werther-effect": legend or reality? *Neuropsychiatrie: Klinik, Diagnostik, Therapie und Rehabilitation: Organ der Gesellschaft Osterreichischer Nervenarzte und Psychiater* 21, 4 (2007), 284–290.

[40] Matthew K Nock, Guilherme Borges, Evelyn J Bromet, Jordi Alonso, Matthias Angermeyer, Annette Beautrais, Ronny Bruffaerts, Wai Tat Chiu, Giovanni De Girolamo, Semyon Gluzman, et al. 2008. Cross-national prevalence and risk factors for suicidal ideation, plans and attempts. *The British Journal of Psychiatry* 192, 2 (2008), 98–105.

[41] Denis Paperno and Marco Baroni. 2016. When the whole is less than the sum of its parts: How composition affects pmi values in distributional semantic vectors. *Computational Linguistics* 42, 2 (2016), 345–350.

[42] James W Pennebaker. 2011. The secret life of pronouns. *New Scientist* 211, 2828 (2011), 42–45.

[43] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on*

empirical methods in natural language processing (EMNLP). 1532–1543.

[44] Kelly Posner, Gregory K Brown, Barbara Stanley, David A Brent, Kseniya V Yershova, Maria A Oquendo, Glenn W Currier, Glenn A Melvin, Laurence Greenhill, Sa Shen, et al. 2011. The Columbia–Suicide Severity Rating Scale: initial validity and internal consistency findings from three multisite studies with adolescents and adults. *American Journal of Psychiatry* 168, 12 (2011), 1266–1277.

[45] Ali Pourmand, Jeffrey Roberson, Amy Caggiula, Natalia Monsalve, Murwarit Rahimi, and Vanessa Torres-Llenza. 2018. Social Media and Suicide: A Review of Technology-Based Epidemiology and Risk Assessment. *Telemedicine and e-Health* (2018).

[46] Hemant Purohit, Andrew Hampton, Valerie L Shalin, Amit P Sheth, John Flach, and Shreyansh Bhatt. 2013. What kind of# conversation is Twitter? Mining# psycholinguistic cues for emergency coordination. *Computers in Human Behavior* 29, 6 (2013), 2438–2447.

[47] Philip Resnik, Anderson Garron, and Rebecca Resnik. 2013. Using topic modeling to improve prediction of neuroticism and depression in college students. In *Proceedings of the 2013 conference on empirical methods in natural language processing*. 1348–1353.

[48] Jo Robinson, Maria Rodrigues, Steve Fisher, and Helen Herrman. 2014. Suicide and social media. *Melbourne, Australia: Young and Well Cooperative Research Centre* (2014).

[49] Xin Rong. 2014. word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738* (2014).

[50] H Andrew Schwartz, Johannes Eichstaedt, Margaret L Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. 2014. Towards assessing changes in degree of depression through facebook. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. 118–125.

[51] H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Richard E Lucas, Megha Agrawal, Gregory J Park, Shrinidhi K Lakshmikanth, Sneha Jha, Martin EP Seligman, et al. 2013. Characterizing Geographic Variation in Well-Being Using Tweets.. In *ICWSM*. 583–591.

[52] Amit Sheth, Utkarshani Jaimini, and Hong Yung Yip. 2018. How Will the Internet of Things Enable Augmented Personalized Health? *IEEE intelligent systems* 33, 1 (2018), 89–97.

[53] Joongbo Shin, Yanghoon Kim, Seunghyun Yoon, and Kyomin Jung. 2018. Contextual-CNN: A Novel Architecture Capturing Unified Meaning for Sentence Classification. In *Big Data and Smart Computing (BigComp), 2018 IEEE International Conference on*. IEEE, 491–494.

[54] Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. 2018. 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*. 25–36.

[55] Guillermo Soberón, Lora Aroyo, Chris Welty, Oana Inel, Hui Lin, and Manfred Overmeen. 2013. Measuring crowd truth: Disagreement metrics combined with worker behavior filters. In *CrowdSem 2013 Workshop*.

[56] Shaina J Sowles, Melissa J Krauss, Lewam Gebremedhn, and Patricia A Cavazos-Rehg. 2017. "I feel like I've hit the bottom and have no idea what to do": supportive social networking on Reddit for individuals with a desire to quit cannabis use. *Substance abuse* 38, 4 (2017), 477–482.

[57] Robyn Speer and Joanna Lowry-Duda. 2017. Conceptnet at semeval-2017 task 2: Extending word embeddings with multilingual relational knowledge. *arXiv preprint arXiv:1704.03560* (2017).

[58] Hajime Sueki. 2015. 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.

[59] C Lee Ventola. 2014. Social media and health care professionals: benefits, risks, and best practices. *Pharmacy and Therapeutics* 39, 7 (2014), 491.

[60] Eduard Vieta and Marc Valentí. 2013. Mixed states in DSM-5: implications for clinical care, education, and research. *Journal of affective disorders* 148, 1 (2013), 28–36.

[61] Colin G Walsh, Jessica D Ribeiro, and Joseph C Franklin. 2017. Predicting risk of suicide attempts over time through machine learning. *Clinical Psychological Science* 5, 3 (2017), 457–469.

[62] Wenbo Wang, Lu Chen, Keke Chen, Krishnaprasad Thirunarayan, and Amit P Sheth. 2017. Adaptive training instance selection for cross-domain emotion identification. In *Proceedings of the International Conference on Web Intelligence*. ACM, 525–532.

[63] Wenbo Wang, Lu Chen, Ming Tan, Shaojun Wang, and Amit P Sheth. 2012. Discovering fine-grained sentiment in suicide notes. *Biomedical informatics insights* 5 (2012), BII–S8963.

[64] Xufei Wang, Lei Tang, Huiji Gao, and Huan Liu. 2010. Discovering overlapping groups in social media. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. IEEE, 569–578.

[65] Andrea N Weber, Maria Michail, Alex Thompson, and Jess G Fiedorowicz. 2017. Psychiatric emergencies: assessing and managing suicidal ideation. *Medical Clinics* 101, 3 (2017), 553–571.