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 suicidal ideation and behaviors, including suicide attempts. A report by 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 U.S. 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, and using language modeling, medical entity recognition, and normalization, negation detection to interpret posts 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 concerning ideation, behavior, and attempt. Our 5-label classification scheme outperforms the state-of-the-art schemes by 4.2% and 12.5% in graded recall and perceived risk measure, respectively. Convolutional neural network (CNN) provided the best performance in our scheme due to the discriminative features and use of domain-specific knowledge resources, as opposed to SVM-linear that is used in the state-of-the-art.

1 INTRODUCTION

According to recent data from the U.S. Centers for Disease Control and Prevention (CDC), suicide is the second leading cause of death for people aged between 10-34 [40] and fourth leading cause for people aged 35-64, escalating the suicide rate in the US by 30% since 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, and suicidal patients may be hesitant or too embarrassed to bring up the topic of suicide. Clinicians often have no prior warning that the patient in front of them is currently suicidal or will be developing significant suicidality. Hence, novel strategies are necessary to proactively detect, assess, and enable timely intervention to prevent suicide[3].

Mental health conditions have been closely linked to suicide [16]. Depression, bipolar and other mood disorders are known to

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

[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

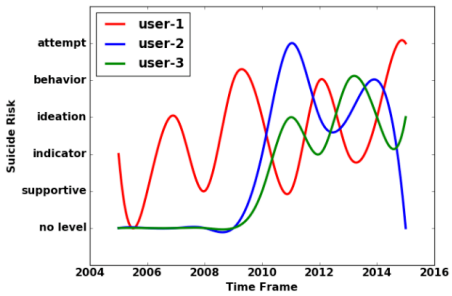

Figure 1: Changing Suicide Risk Severity of 3 Redditors over a period of 11 years

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. Werther effect[34] refers to the influence that a popular figure's suicide can have on an individual, encouraging them to commit suicide, revealing that suicide attempt can be contagious. There are several resources for patients to seek help such as CrisisTextLine, teen line, 7cups.com, imalive.org, and The Trevor Project for LGBTQ. Additional measures are necessary to improve timely intervention[4] further. 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. 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[20].

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 [6, 7, 51]. 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 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.[49] 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. 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 decide on the most appropriate care plan[60]. 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 the quality of actionable insights to enable monitoring progression and alerting MHP as necessary.

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. In this research, we provide a C-SSRS-based five-label (supportive, indicator, ideation, behavior, and attempt) multi-class classification scheme guided by clinical psychiatrists, which allows the MHP to decide on the most appropriate care plan and actionable measure of an individual's suicidality. We generated two variants of 5-label viz, 4-label (indicator, ideation, behavior, and attempt) and (3+1)-label (supportive + indicator, ideation, behavior, and attempt).

Apart from identifying the risk factors of suicide, we can develop approaches to generate answers from the content to the questions 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.

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[55]) and ICD-10, we developed suicide risk severity lexicon, which is 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 [47], MHPs will have the ability to be better informed about the patient's mental health conditions including their suicidality and enable timely intervention and response.

In section 2, we review existing research related to our study. We explain: in section 3, the resources that we use in our approach, in

[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

section 4, the critical components of the approach that we developed. Then, we give details on the experimental design in section 5, and we discuss our results in section 6.

2 RELATED WORK

Prior works relevant to the key areas related to our study.

2.1 Suicide and Social Media

Jashinsky et al. [26], and Christensen et al. [8] predicted the level of suicide risk for an individual over a period of time using SVM and the features of TF-IDF, word count, unique word count average word count per tweet and average character count per tweet. De Choudhary et al. [16] 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, [16] measured the likelihood of a user sharing thoughts on suicide in the future. Another study [53] investigated the linguistic variations among different authors on social media, and observed correlation between suicidal behavior and suicide-related tweets. Furthermore, [7] conducted a qualitative analysis of the Tumblr content shared to gain better understanding of self-harm, suicidal content, and depression, and concluded with a need for suicide prevention. Robinson et al. [43] reports in their review of literature, that people with mental health conditions on social media look for other people with similar problems.

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 [14, 15, 42], psychological state and well-being [45, 46]. However, monitoring severity of suicide risk is essential for timely intervention. Nock et al. [35] 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 can be extracted from the content of an individual as they provide explicit, implicit or ambivalent clues for suicide. These clues can help a MHP assess suicide severity, and better structure the treatment process [10]. Corbitt-Hall et al. [11] defined a rubric based on a 4-way suicide categorization. We use a rubric (or questionnaire) endorsed by NIH and SAMHSA.

Shing et al.[49] 1.5M posts from 11K users on SuicideWatch, out of which 245 users were annotated by experts and non-experts. The study evaluates the annotation quality of experts and non-experts and performs risk and suicide screening experiments using domain lexicon, linguistic and psycholinguistic features involving machine learning and deep learning classifiers. The study fails to bring together different mental health conditions that lead to suicide indication, suicide ideation and attempt. It is critical to highlight supportive users on social media, who are not suicidal, as these constitute negative samples.

2.3 Models for Suicide Prediction

In a recent study on predicting suicide attempt in adolescents, Bhat et al. employed neural networks (NN) with multi-layer depth 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 tuning the depth of the network, researchers achieved a true positive rate of 70% and a true negative rate of 98.2% [3]. Another study by Walsh et al. [56] on predicting suicidal attempts through temporal analysis, employs Random Forest (RF) over a cohort of 5,167 patients. The study segregates the cohort into cases and controls, in which the former constitutes a set of 3,250 and the latter 1,917 patients. They achieved an F-score of 86% with a recall of 95% [56]. As they used binary classification scheme EHR for dataset, , are their model was oblivious to identification of supportive or indicative users. However, a transfer learning from social media to EHR is worth investigating [57]. Amini et al. utilized SVM, and decision trees besides RF and NN, for assessing the risk of suicide in a dataset of individuals from Iran [1]. Du et al. identified psychiatric stressors for suicide using CNN dichotomously conceptualizing the problem, a precision of 78%[18]. On another study, individuals were classified as at-risk of suicide using CNN while psychiatric stressors were identified through RNN [18].

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 knowledge (e.g., mental health disorder, symptoms, side-effects, drugs) 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) 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 exploited 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[6] containing 315 entities (814 instances) and 31 relations defining drug-abuse and mental-health concepts. The ontology has been utilized in analyzing web-forum content

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

related to buprenorphine, cannabis, a synthetic cannabinoid, and opioid-related data [12, 13, 29]. In [19] it was expanded using DSM-5 categories covering mental health and applied to depression. This mental health coverage includes suicide relevant to our study.

3.2 Existing Domain Specific Lexicons

Prior research [5, 33] highlighted the disparity between the language (e.g., informal terms) 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 social media content and mapping them to concepts in medical knowledge bases [31]. We use (i) the two lexicons, namely, TwADR-L and AskAPatient (see Table 1) to map the social media content to medical concepts [31], and (ii) the anonymized and annotated suicide notes made available through i2b2 challenge to identify social media content with negative emotions (see Table 2).

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'

Table 1: Existing domain specific lexicons used in this study

TwADR-L [31] 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 [31] maps informal terms from AskAPatient web forum to medical concepts in SNOMED-CT and Australian Medical Terminology[30]. Since this lexicon was created from a web forum, it is more informative compared to TwADR-L. Informatics for Integrating Biology and the Bedside (i2b2) Suicide Notes is a dataset generated as a part of the emotion recognition task in 2011 [58]. In this i2b2 dataset, we have 2K suicide notes annotated for different emotions, and we separated the notes with negative emotions resulting in 817 suicide notes (see Table 2).

Emotion Label	#Suicide Notes	Example
abuse	9	Now my son got married to a horrible woman who does not care for me , curses , and 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 that it 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, Don't think to badly of me for taking this way out but I am frustrated by taking so much pain that I can't go on like this
sorrow	51	My heart has been hurt hard and everyday I am grieving.

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

[8]http://sideeffects.embl.de

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). 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).

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.

Table 3: Suicide Risk Severity lexicon

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 [39]. 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 SW are suicidal. We identify suicidal candidates in subreddit SW by looking into the nature of words used in user's posts. To identify "suicidal" words, we utilized a technique to discover the trade-off between high-frequency (less important) against low-frequency (more important) words. We analyzed the content of SW subreddit against Zipf-Mandelbrot

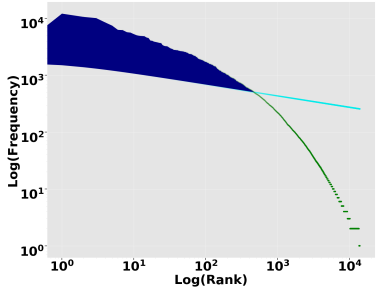

Figure 2: Zipf-Mandelbrot law over SuicideWatch content for identifying prominent suicide seed terms. Highlighted is the selected region.

law to precisely identify terms that are 'prominent' in the online discussion of suicidal thoughts. In figure 2, the red line follows Zipf-distribution while the blue line follows the 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. This

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[9]	96
Kill	577	Sleep	193

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

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 psychiatrist and a curated list of 339 words with a cut-off frequency of 725. A sample list of 10 words are 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. In this study, we utilize embeddings from ConceptNet[10] (vocabulary: 417193, dimension:300), which is a multi-lingual knowledge graph (created using words and phrases which are not necessarily concepts (or entities). The knowledge graph has been created using information from expert sources, crowd-sourcing, DBpedia, vocabulary derived from Word2Vec[11] [44], and Glove[12] [38]. We utilized ConceptNet embeddings as they have proved to improve natural language applications [52].

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 2.5 mental health related subreddits. These mental health subreddits are: 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). 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.

[10]http://conceptnet.io
[11]https://code.google.com/archive/p/word2vec/
[12]https://nlp.stanford.edu/projects/glove/

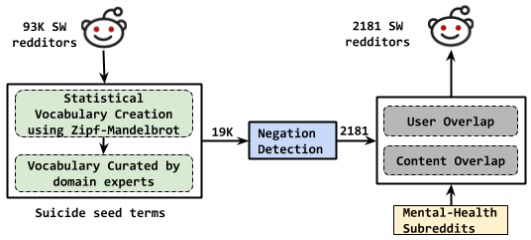

Figure 3: Procedure for generating the annotated dataset. Only 9 mental health subreddits were considered because of significant content overlap (see Figure 4)

4.1 2181 Potential Suicidal Redditors

Subreddit "SuicideWatch" (SW) has nearly 93K redditors as of 2016, and not all these users have suicidal tendencies. Therefore, we need to create a representative sample dataset that will contain users potentially at five-levels of suicide risk. We used a list of seed terms generated using Zipf-Mandelbrot (see section 3.5), to identify users who mentioned these terms in their content. We obtained a working set of 19K redditors. Next, we employed negation detection procedure (see Section 4.2) to further eliminate non-suicidal users. Finally, we obtained 2181 users who are potentially suicidal and have participated in other mental health subreddits. So, we take their cross-posts in other mental health subreddits, to account for the history of mental disorder.

4.2 Negation Detection

Negation detection is a crucial part of the information extraction process as the presence of negated sentences can confound a classifier [21]. For example, *I am not going to end my life because I failed a stupid test* is a negated sentence in suicidal context, 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, while the latter sentence can be analyzed using simple match. The presence of negated sentences can influence false alarm rate of a classifier and deserves proper handling. We employ a negation detection tool developed using a list of negation cues, and probabilistic context-free grammar that supports negation extraction and negation resolution[21]. Such a tool correctly interprets social media content. We eliminate non-suicidal Redditors posting in SW to reduce the false-alarm rate of the classifier.

4.3 User and Content Overlap

As individuals form communities based on shared topics of interest related to mental health conditions [54, 59] through different subreddits, we performed user and content overlap analysis between SW and other mental health subreddits to enrich the content of users. This analysis provides deeper insight into how potentially suicidal users communicate on problems including causes, symptoms, and treatment solutions. We extracted content of the 2181 redditors of SW in other subreddits, and identified their suicide-related posts employing a similarity assessment technique through domain-specific lexicons, LDA2Vec[32] and ConceptNet. Then, we appended the injected the content from other subreddits into the content of each of the 2181 redditor in SW. Formally, we define the

User overlap between SW and other subreddits as:

$$\text{User-Overlap}(SW, MH_i) = \frac{U^{SW} \setminus U^{MH_i}}{|U^{MH_i}|} \quad i = 1 : |MH| \quad (1)$$

where U^{SW} represent users in 2181 redditors set, U^{MH_i} represent users in i^{th} mental health subreddit and 2181 redditors set. We computed scores to quantify the overlap in users between SW and MH_i (excluding SW) (see Figure 4). We leverage the quantified similarity of suicide-related topics between content of the 2181 users in SW and other subreddits, to enrich the content through appending the content in other subreddits into the SW content. 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 patients 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 was performed through string match.

Content overlap using domain-specific lexicons. We trained an LDA model with topic coherence over the normalized content to find coherent topics for SW subreddit. Similarly, we processed posts in other mental health related subreddits. Subsequently, we generated two sets of Topics: SW subreddit, and related MH subreddits. The topical similarity was calculated between topics of SW and MH. We formalized the topic similarity(TS) between content of 2181 users in MH_i and SW as

$$TS(Lex) = \underset{users, i=1}{\overset{(|MH_i|, 1)}{}} \frac{cos(\vec{U}^{SW}, \vec{U}^{MH_i})}{|U^{MH_i} \setminus U^{SW}|} \quad users := 1 : |U^{SW}| \quad (2)$$

$TS(Lex)$ is calculated as the ratio of averaged cosine similarity between topic vectors users in U^{MH} ($\vec{U}^{MH_i}$) and topic vectors of a users in U^{SW} ($\vec{U}^{SW}$) over number of user in the intersection of U^{MH_i} and U^{SW}. The resultant matrix TS(Lex) has a dimension of 2181 x 1, representing users and their similarity scores. The parameter Lex takes as input: TwADR-L (TS(TwADR-L)) and AskaPatient(TS(AskaPatient)). For such task, we trained two topic models because TwADR lexicon is created using twitter content and AskaPatient Lexicon is created using Forum content.

For quantifying the user's content with appropriate emotion label (Table 3), we generated embeddings of content in SW and other MH subreddits for each user using ConceptNet word embedding model, we also generated the representations of the emotion labels used in the suicide notes, by combining the embedding vectors of their corresponding notes 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:

$$UL^{SW} = \{cos(\vec{u}, \vec{l}) \mid u \in U^{SW}, l \in L\} \quad (3)$$

where UL^{SW} represents the set of users in SW subreddit (U) and their cosine similarity values with labels in L forming a matrix of dimension 2181 x 6. 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 is created for 2181 users using their content in other MH subreddits. We denote such a matrix as UL^{MH} of dimensions 2181 x 6. UL^{SW} and UL^{MH} is interpreted as a matrix showing to what degree users' content are close to abuse, anger, fear, guilt, hopelessness, and sorrow. Thereafter, we generated a user similarity matrix ($Usim$) as a normalized product of UL^{SW} and UL^{MH}. $Usim$ has a dimension of 2181 X 1. Formally we define it as:

$$Usim = \frac{cos(UL_u^{MH}, UL_v^{SW})}{|U^{SW}|} \quad u, v \in U^{SW} \quad (4)$$

We define an empirical threshold of 0.6 over similarity scores from from matrices: $Usim$, $TS(TwADR\ L)$, and $TS(AskaPatient)$. Content of 2181 users in MH subreddits having similarity greater than or equal to 0.6 across all the three matrices are appended to their respective

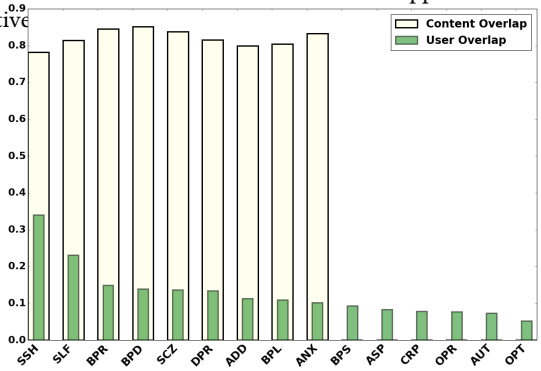

Figure 4: User Overlap and Content Overlap based quantification of influence of other mental health related subreddit to SW. Subreddits *SSH* and *SLF* have the highest content overlap with SW followed by *BPR* and *BPD*.

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, disease, or substance abuse. A user with suicidal ideation expresses current or a history of thoughts of suicide, the desire to die, and/or suffering from high-risk mental illness. This category can be seen to escalate to *suicidal behavior*(*BR*), which can be operationalized as actions which elevate ideation into a higher risk category. 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. This includes but is not limited to self-harm such as cutting or 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 a empt*(*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 this social media adaptation, two additional categories were added to define user behavior that was 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 a 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 *suicide indicator* because while they expressed known risk factors that could be monitored they would also count as a false positives (by classifier) 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 didn't define their motive for interacting at all (as opposed to a family history). Since posting on Reddit is not itself a risk factor, it was deemed appropriate to 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 social media users 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.

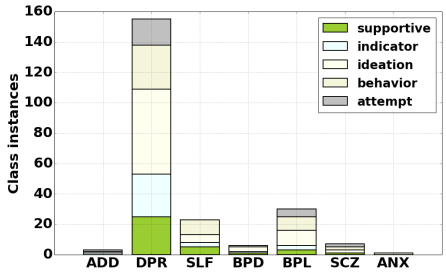

Figure 5: Distribution of 500 annotated users across different mental health subreddits

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 subreddit (except SW) contributed most to suicidality, we mapped potential suicidal Redditors to their subreddits (see Figure 5).

Always time for you to write your happy ending .... doesnt need to be spelled out with alcohol and Xanax. Youre more than worth our time. You still need to know what to do with your current relationship. All Im saying is, once you make a decision: 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....Jumping spiders....fluffy puppy dog face	IN
I sometimes have days when I literally cant bear to move....thats me every day....my depression....since I was 14....likely suffer the rest of my life....Death is the only thing 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 and carry with me most of the time	BR
I am going to off myself today and had a loaded gun to my head, feeling determined....felt myself as a huge disappointment....good for nothing....I screwed family life or how it could be better....breaks my heart everyday.	AT

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

	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

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

4.4.3 Evaluation of Annotation. Four mental health experts were involved in the annotation process. Each expert received 500 users dataset comprising of 15755 posts. The annotator agreement was performed using Krippendorff metric as it is appropriate for ordinal labels. α is calculated as $1 \quad D_o \bullet De$, where D_o is observed disagreement and D_e is expected disagreement. Formally, we define D_o and D_e as:

$$D_o = \frac{1}{N \quad \mathrm{j}G\mathrm{j}^2} \tilde{\mathrm{O}} \quad \mathrm{j}\tilde{\mathrm{O}}\mathrm{j} \quad \mathrm{j}A_j \quad G_k\mathrm{j}^2;\ D_e = \frac{1}{N \quad \mathrm{j}G\mathrm{j}^1 N \quad \mathrm{j}G\mathrm{j} \quad 1^{\circ}} \tilde{\mathrm{O}} \quad \mathrm{j}\mathrm{G}^j \quad \mathrm{G}^k\mathrm{j}^2 \tag{5}$$

where N is the number of users in the dataset, A_j represents a label from a j^{th} annotator, and G_k represents a label from a k^{th} annotator in a group G. $\mathrm{j}\mathrm{G}^j \quad \mathrm{G}^k\mathrm{j}^2$ represents pointwise difference across annotations by group of annotators ($\mathrm{G}^j; \mathrm{G}^k$).

We perform two analysis: A pair-wise annotator agreement to identify the annotator with highest agreement with others, setting $\mathrm{j}G\mathrm{j}$ to 2 in 5, and an incremental group wise annotator agreement to find the robustness of the earlier annotator[50]. For group wise annotator agreement, we set $\mathrm{j}G\mathrm{j}$ = {2,3,4}. Results of pairwise and group wise annotators agreement is in Table 6.

5 EXPERIMENTAL DESIGN

5.1 Characteristic Features

Prior research has shown the importance of psycholinguistics, lexical, syntactic, and emotion features in the enhancing the efficacy of the classifier [22, 41]. We further improve our feature set with information provided by Reddit. We used following features in training our models: *AFINN*[13] 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

[13] http://neuro.imm.dtu.dk/wiki/AFINN

controversial. We computed *controversiality score* as the ratio of the maximum value of the difference, between upvotes and downvotes, and 1, over total votes. 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 [9]. 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 [17]. On social media, readability is an important factor for understandability of the text. We use *height of the dependency parse tree* to measure readability, with parse tree height being proportional to readability [23]. We employ *maximum length of verb phrase* [24] to capture less suicidal individuals. Verb phrase length is negatively correlated with the length of the subject in the content; therefore, length of the verb phrase is greater when the subject is a pronoun, rather a noun. Similarly, *number of pronouns* was used to determine whether they are sharing a direct experience or second hand experience [37]. The value of this feature was high for users classified as supportive or indicative, as these users usually help others. Moreover, *number of Sentences* and *Number of definite articles* are also discriminative [25].

5.2 Baseline

Suicide lexicon developed as a part of the study for initial filtering of users and annotation process is a suitable resource for the baseline. Our empirical baseline model is a rule-based model classifying the user based on a strict and soft match criteria according to the presence of a concept in the user's content and the suicide risk severity lexicon. To have a competitive baseline, we compared our rule-based baseline with word-embedding and TF-IDF based approaches to suicide classification [27]. Considering the experiments using word-embedding model trained over suicide and non-suicide related content, compositionality of word vectors was put to test following the works of [2, 36]. However, the suicide risk severity lexicon based baseline outperformed these competitive experiments. Apart from lexicon-based baseline, we consider the second baseline based on Shing et al.' s work [49] employing machine learning models for predicting the suicide risk.

5.3 Convolutional Neural Network

We have implemented a convolutional neural network (CNN) as proposed in [28] for our contextual classification task [48].

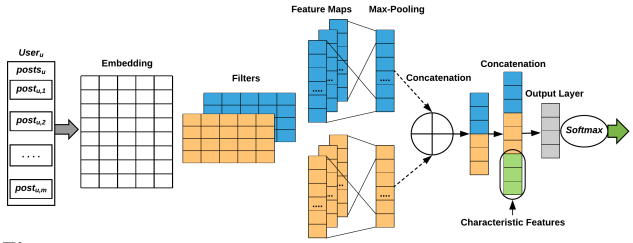

Figure 6: CNN architecture created over aggregated posts from a user.

The model in Figure 6 takes the embedding of user posts as input and classifies it into one of the suicide risk severity level. We merge the user posts and pass the concatenated embeddings of the words in the merged content as the input to the model.

$$posts_u = post_{u;1} \oplus post_{u;2} \oplus :: post_{u;p} :: \oplus post_{u;P} \quad (6)$$

Here $\oplus$ represents the concatenation operation of P posts of user u, where each post p of user u ($post_{u;p}$) is the concatenation of the word vectors of each word w ($\vec{w}_{u;p;w}$) in W (total number of words in a post).

$$post_{u;p} = \vec{w}_{u;p;1} \oplus \vec{w}_{u;p;2} \oplus :: \vec{w}_{u;p;w} :: \oplus \vec{w}_{u;p;W} \quad (7)$$

Model has a convolution layer with filter window {3, 4, 5} and 100 filters for each. After getting the convoluted features, we apply maxpooling and concatenate the representative pooled features. We pass the pooled features through a dropout layer with dropout probability 0.3, followed by a output softmax layer. Model's learning rate was set to 0.001 with adam optimizer. While training the model we have used mini batch of size 4 and trained for 50 epoch. CNN's performance is compared and evaluated in Section 6.

5.4 Evaluation Metrics

We alter the formulation of False Positives, and False Negatives to better evaluate the model performance. False Positives (FP) is defined as the ratio of the number of times the predicted suicide risk severity level (r^0) is greater than actual level(r^o) over the size of test data (N_T). False Negatives (FN) is defined as the ratio of the number of times r^0 is greater than r^o over N_T. Since the numerator of FP and FN is compare suicide risk severity levels (r^0, r^o), we termed precision as graded precision, and recall as graded recall. Ordinal Error (OE): It is defined as the ratio of the number of samples where difference between r^o and r^0 is greater than 1. we formally define FP, FN, and OE as:

$$FP = \frac{\sum_{i=1}^{N_T} \mathbb{1}[r_i^0 > r_i^o]}{N_T};\ FN = \frac{\sum_{i=1}^{N_T} \mathbb{1}[r_i^o > r_i^0]}{N_T};\ OE = \frac{\sum_{i=1}^{N_T} \mathbb{1}[\Delta(r_i^o, r_i^0) > 1]}{N_T} \quad (8)$$

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 over summation of disagreements between the annotators multiplied by a reducing factor that reduces the penalty if the prediction matches any other annotator.

$$PRM = \frac{1}{N_T} \sum_{i=1}^{N_T} \frac{1 + \Delta(r_i^0, r_i^o)}{1 + \sum_{p;q=1}^{|\mathbb{G}^i|} \Delta(A_p, A_q)} \frac{\sum_{p=1}^{|\mathbb{G}^i|} \mathbb{1}[r_i^0 = A_p]}{|\mathbb{G}^i|} \quad (9)$$

Where r_i^0 and r_i^o are the prediction and actual response for i^{th} sample. The denominator is the disagreement between A_p and A_q annotators summed over all annotators in a group $\mathbb{G}$. $\frac{\sum_{p=1}^{|\mathbb{G}^i|} \mathbb{1}[r_i^0 = A_p]}{|\mathbb{G}^i|}$ 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 when the r^0 disagrees with all the annotators in $\mathbb{G}$, 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' performance over 4 labels in which supportive (or negative) samples are removed, and **Experiment 3** comprises labels defined according to 4-level categorization (where supportive and indicator classes are merged into one class : no-risk). Further, for each experiment, the input data is of three forms: (I1) Only textual features (TF) represented as vectors of 300 dimensions generated using ConceptNet embeddings , and (I2) 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 four learning models (SVM, RF, Feed-Forward Neural Network(NN), 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-linear which has been recorded as the best performing model in a study by Shing et al. [49].

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-Linear	I1	0.60	0.45	0.52	0.12
	I2	0.77	0.40	0.53	0.09
Random Forest	I1	0.68	0.49	0.57	0.19
	I2	0.62	0.45	0.52	0.11
Feed-Forward NN	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

Table 7: Experiment with 5-label Classification

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-linear show an improvement in precision compared to baseline; however, there is 12% and 27% reduction in recall respectively. Further, Random forest 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 model, we analyze them from the perspective of OE. Such a measure is coarse-grained, and focuses more on false negatives as opposed to acceptable false positives. 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 of it ignores the biases in the gold standard data. As a result, CNN correctly (or closely) predicted the severity of 90% of users.

Input Type I2: In comparison to the second baseline, CNN outperforms SVM linear 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[14] assumptions. Whereas, CNN's convoluted representation ignores i.i.d assumptions and the non-linearity induced by ReLU tries to balance false positives and false negatives. It can be seen from recall fo SVM-rbf for I2 which is higher than SVM-linear. However, SVM-rbf fails to balance false positives and false negatives because of i.i.d considerations. Further, from OE, we infer that for 9% of the users CNN and SVM-linear predicted a suicidality by a margin of 2 or more from the observed (actual).

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.

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-Linear	I1	0.62	0.53	0.57	0.12
	I2	0.68	0.57	0.61	0.09
Random Forest	I1	0.67	0.41	0.51	0.22
	I2	0.64	0.47	0.54	0.18
Feed-Forward NN	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

Table 8: Experiment with 4-label Classification

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 4-label.

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 (not much change in OE) of SVM-linear. It can be inferred that SVM-linear is vulnerable to predict 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 level; 4-label scheme does not bring in major change in OE for CNN.

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

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 [11]. For this experiment, we considered two top performing models from previous experiments: SVM-Linear and CNN.

Approach	Input	Collapsed Supp. and Ind. Class			
		Graded Precision	Graded Recall	F-Score	OE
SVM-Linear	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**

Table 9: Experiment with 3+1-label Classification

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 Table 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 Group: Supportive and Indicator User, (2) Treatment Group: Ideation, Behavior, Attempt.

Figure 7: Confusion Matrix of 5-label classification. (left) CNN, and (right) SVM-Linear. Y-Axis: True Level, X-Axis: Predicted Level

From figure 7, out of 36 No-treatment users, CNN correctly classifies 20 users (56%) whereas SVM-Linear correctly classifies 22 users (61%). However, observing a larger 64 Treatment users, CNN correctly classifies 51 users (80%) whereas SVM-Linear correctly classifies 46 users (72%). Hence, CNN provides more suitable class for the user compared to SVM-linear.

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 8, we observe that CNN classifies 59 out of 64 users (92%) annotated under Treatment whereas SVM-Linear classify 53 out of 64 users (83%).

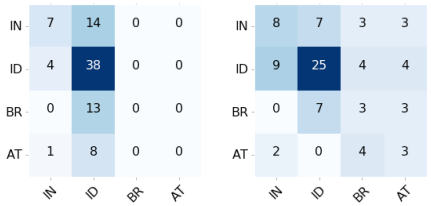

Figure 8: Confusion Matrix of 4-label classification. (left) CNN, and (right) SVM-Linear

6.6 3+1 Label Confusion Matrix Analysis

There are 36 users under No-Treatment (supportive+indicator) group and 64 users under Treatment group. Based on figure 9, we noticed that CNN correctly classifies 26 out 36 (72%) No-Treatment users whereas SVM-Linear scored 16 out of 36 (44%). Further, CNN and SVM-linear 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

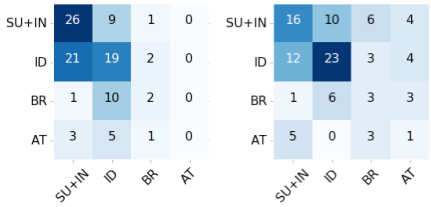

Figure 9: Confusion Matrix of (3+1)-label classification. (left) CNN, and (right) SVM-Linear

for SVM-linear but on comparing 5-label and 3+1 label classification schemes, we observed that collapsing of the supportive and indicator class can lead to increase in the false positive as SVM-linear predicts them as behavior and attempt. There is a reduction in the true positive score for predictive and actual ideation classes, and number of users marked as "attempt" have been classified as "supportive+indicator". As a result, the false negative of the models have increased. Although this analysis proves the efficacy of 5-label classification over 3+1, and CNN being a conservative model, there is a possibility of annotator bias in the data. So below we perform PRM analysis of SVM-Linear and CNN over 2 classification schemes: 5-label and 3+1 label.

6.7 Perceived Risk Measure Analysis

On analyzing models behavior using PRM, Table 10 shows that there is a 12.5% difference between 5-label and 3+1 label classification scheme. 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 SVM-linear has a high risk score compared to CNN in the both the classification scheme. From Table 10, we believe 5-label classification is an adaptable scheme over the other schemes for assessment of suicide.

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

Table 10: PRM based comparison of 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 investigations.

REFERENCES

[1] 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.

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

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

[4] 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.

[5] 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.

[6] 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.

[7] 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).

[8] 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.

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

[10] 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*.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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).

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

[25] 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.

[26] 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).

[27] 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).

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

[29] 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.

[30] 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.

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

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

[33] 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.

[34] 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.

[35] 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.

[36] 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.

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

[38] 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.

[39] 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.

[40] 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).

[41] 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.

[42] 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.

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

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

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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*.

[51] 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.

[52] Robert 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).

[53] 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.

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

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.