Prediction and Analysis of Degree of Suicidal Ideation in Online Content

by

Noah C. Jones

B.Sc., Morehouse College (2017)

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning
in partial fulfillment of the requirements for the degree of

Master of Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2020

Author .....
Program in Media Arts and Sciences, School of Architecture and Planning
May 7, 2020

Certified by .....
Rosalind Picard
Professor of Media Arts and Sciences
Thesis Supervisor

Accepted by .....
Tod Machover
Academic Head of Media Arts and Sciences

Prediction and Analysis of Degree of Suicidal Ideation in Online Content

by

Noah C. Jones

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning
on May 7, 2020, in partial fulfillment of the
requirements for the degree of
Master of Science

Abstract

Machine learning (ML) has increasingly been used to address the growing burden of mental illness and lack of access to quality mental health care. Recently such models have been applied to online data, such as social media postings to augment mental health screening. Despite the potential of these methods, online ML classifiers still perform poorly in multi-class settings. In this thesis, we propose the usage of novel document embeddings and mental health based user embeddings for triaged suicide risk screening. Machine learning to infer suicide risk and urgency is applied to a dataset of Reddit users in which the risk and urgency labels were derived from crowdsource consensus. We show that the document embedding approach outperforms count-based baselines and a method based on word importance, where important words were identified by domain experts. We examine interpretable features and methods that help to discern and explain risk labels. Finally, we find, using a Latent Dirichlet Allocation (LDA) topic model, that users labeled at-risk for suicide post about different topics to the rest of Reddit than non-suicidal users.

Thesis Supervisor: Rosalind Picard
Title: Professor of Media Arts and Sciences

Prediction and Analysis of Degree of Suicidal Ideation in Online Content

by

Noah C. Jones

This thesis has been reviewed and approved by the following two readers:

……

Cynthia Breazeal
Associate Professor of Media Arts and Sciences
MIT Media Lab

……

Stephen Schueller
Assistant Professor of Psychological Science
University of California, Irvine

Acknowledgments

To the supportive friends, family, committee members and collaborators who made this possible I love and thank you dearly!

I would like to give a special thanks to Rosalind Picard, Stephen Schueller, Natasha Jaques, Pat Pataranutaporn, and Ian Magnusson who shaped my thinking and engaged me to bring the thesis to another level.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Suicide Prevalence

In the United states, it is estimated that every 13 minutes at least one person will die by suicide [1]. Approximately 43,000 people die each year and it is the 2nd leading cause of death in those between the ages of 10 and 34 [2]. Such events not only affect the individual, but also those closest to them and the wider society. For every suicide, it is estimated that 135 relatives or friends have a significantly increased risk of depression and suicide [3]. In addition, attempted and completed suicides account for $93.5 billion in lost wages and medical expenses each year in the U.S alone [4]. While there have been several large-scale prevention efforts [5], suicide rates have increased 35% from 1999 to 2018 [6]. Thus, it is critical to identify and improve upon the most promising strategies to prevent suicide.

1.2 Risk and Protective Factors

In order to do so, it is important to understand possible causes of suicide. Such causes are often multi-faceted and interdependent. Over the last two decades, researchers [7, 8] and the World Health Organization [9] have enumerated common psychological risk and protective factors. They can largely be classified in terms of individual risk factors, social risk factors and situational factors. Individual factors, include:

history of mental illness, abuse or trauma, chronic illness and pain, and imbalance in neurobiology. The mental health disorders prevalent in most suicidal victims are depression, post-traumatic disorder, drug abuse/dependence and conduct disorder. The most common conditions among those who have made an attempt are depression, hopelessness, and impulsivity [10].

Socio-cultural factors that increase risk include: idealization of suicide in the media, religious beliefs that glorify suicide, social contagion of suicide, and barriers to access adequate mental health care such as, stigma or high costs. Situational factors include financial difficulties, career setbacks, death of a loved one, social isolation or easy access to a suicide means.

Conversely, common protective factors for reducing the likelihood of suicide include: a strong social and moral support system, problem-solving and conflict resolution skills, restricted access to lethal objects, sufficient access to healthcare resources, cultural or religious beliefs that do not stigmatize help seeking behaviors or help sustain and idealize suicidal behaviors [9].

Despite progress in elucidating possible risk and protective factors, it is still difficult to understand which are most important on an individual basis and to determine factors that cause suicidal ideators to make an attempt. This is underscored in a study by Ahmedani et al. where they note that 25% of suicidal patients met with a health professional one week prior to their attempt [11]. Also, Franklin et al. [12] describe in a review of 300 studies that ability to forecast suicide has not significantly changed over the past five decades.

1.3 Challenges of Suicide Prevention

We think two important reasons for the difficulty of maximizing the impact of preventive efforts is under-reporting of suicidal intent, and poor access to expert-level risk assessment and treatment. McHugh et al. [13] reviewed 70 studies, and state that suicidality cannot be predicted effectively using the standard practice of clinicians asking in person about suicidal thoughts: 80% of patients who were not already re-

ceiving psychiatric treatment and who died of suicide denied having suicidal thoughts when asked by a general practitioner. They conclude that and with other recent meta-analyses, "highlight a high degree of uncertainty about the statistical strength of commonly used approaches to suicide risk assessment."

Access to high quality care is another important problem. 124 million Americans live in federally designated mental health care shortage regions [14] and even for those that are able to receive care, many psychotherapists lack the specialized clinical training needed to adequately support these populations [15]. Fewer than 20% of individuals who complete suicide have seen a mental health provider in the few months prior to their deaths [16]. Yet, when suicide risk is identified, and successfully treated it is highly likely to have positive outcomes. In a review of suicide prevention strategies from 15 countries, researchers found that providing professional education in suicide risk evaluation and treatment in the primary care level was one of the most effective methods of reducing completed suicide [17].

Thus implementing ways to improve identification of persons and allocation of treatment for suicide may help to lower suicidal attempt levels in the current decade.

1.4 Risk detection from Health Records

Common risk assessment models use medical records and assessments such as the Patient Health Questionnaire-9 (PHQ-9) from large scale electronic health record databases to identify early warning signs for risk and guide care.

In particular, The National Institutes of Health funded *The Mental Health Research Network*; a network of researchers and 13 health systems serving 13 million patients to advance suicide prevention efforts [18]. Findings from this effort demonstrate that scores of 2 or 3 on item 9 of the PHQ-9 denoting moderate and severe suicidal ideation significantly increase risk of suicide [19]. Item 9 was also found to predict suicidal ideation and suicidal attempts up to 90 days after the first report in ethnic minority groups [20]. In addition, Simon et al. note that cumulative risk of suicide increases from 0.03% among those reporting no suicidal ideation on to .3%

for those reporting some suicidal ideation and is an enduring vulnerability factor [21]. Despite these findings, precision and recall for these predictions are notably low. If this factor is used erroneously, it could have the adverse effect of stigmatizing an individual who doesn't intend to commit suicide or mistake someone who under reports suicidal thoughts.

Moreover, others such as Nock et al. [22] argue that one must examine collective risk factors to significantly enhance the assessment of risk with machine learning. Notable contributions , which seek to do this include prediction of suicide from high-risk hospitalization of army veterans [23], prediction of suicidal ideation from multiple somatic symptoms [24] and prediction of suicidal attempt, death and treatments from electronic health record data [25].

Multi-faceted risk modeling is a promising approach for uncovering profiles which distinguish patients who have high likelihood of committing suicide from those that do not. Yet once, identified there may be only a short window of time or consistent effort may be needed to track individuals at need.

1.5 Risk Detection from Online Content

We propose that automatic risk detection using online social media such as Reddit, may be an important tool to address these challenges. Behavioral research indicates people may be more open online than with a clinician. Fein et al. [26] identified that when adolescents were screened online instead of in-person, it doubled the likelihood of identifying adolescents with psychiatric problems. In a survey of adolescents with self-reported mental health problems, a majority (75%) expressed preference in sharing mental health problems online instead of face-to-face [27]. In addition, many people are spending an increasing amount of time on the internet, and in virtual discussion forums such as Reddit and ReachOut which provide opportunities for people to deal with mental health issues, gain support, and find connections. Because these communities have a much higher proportion of users under the age of 26 [28], analyzing their posts and assisting them may be a promising way to reach the younger

demographic at great risk of suicide [29].

Chapter 2

Related Work

2.1 Suicidal Ideation Detection from Text

Over the past few years, Machine learning (ML) and Natural Language Processing (NLP) have emerged as tools to estimate mental health [30] from passive data. Research has identified linguistic markers for suicide from textual information such as blogs [31], poems [32], clinical notes [33], and suicide notes [34].

Online social media data, in particular, has been found to contain predictive information for a range of mental health conditions including depression and suicide [35, 36].

One of the first investigations to analyze suicidal ideation on social media was done by Masuda et al. [37]. They used a logistic regression model and forum behavior patterns on a Japanese online social network and found that features such as number of communities that a user belongs to and intransivity, were most important to distinguish users participating in suicide-related forums from controls [37].

De Choudhury et al. (2016) [38] focused on modeling temporality of suicidal risk in online forums. The authors used a logistic regression model with N-gram and Linguistic inquiry and Word Count (LIWC) features to analyze factors that influence Reddit users posting in depression communities to shift to suicide-support communities. Linguistic coherence and coordination with the community reduced social engagement and manifestation of hopelessness, impulsiveness, anxiety and loneliness

are some of the factors that characterized these shifts.

In addition, Vioules et al. [39] analyzed domain-derived distress levels of 500 posts on twitter. Each post was given a distinct label where 0 represents, "text discussing everyday concerns" and 3 represents, "text including mention of self-harm, suicidal thoughts, ... not being good enough, etc.". Their best performing model was a random forest with n-grams and lexicon words that included information about symptoms, pronouns, and swear words that achieved moderately high F1 scores on a multi-class task of distinguishing the distress levels of the posts.

2.1.1 Contextualized Models for Suicidal Ideation Detection

In the online mental health estimation literature, and from prior work, it is common to use linguistic features from psychological literature such as LIWC, emotion features, mental disease lexicon, or depression based lexical categories [40]. Many researchers are beginning to explore more complex models from the deep-learning literature, such as Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), and Long Short-Term Memory Neural Network (LSTM) often resulting in significant performance gains. [35].

Methods such as LSTM with attention and Transformers, allow for attention-based explanations in order to learn data-derived explanations of the given risk label [41]. These can be used to find interesting previously unrevealed features that are related to risk. In the Crisis Text Line Challenge, 54 million messages were analyzed and "ibuprofen" and "bridge" appeared as words most indicative of risk [42].

The utility of word representations for downstream text classification and other natural language processing tasks is gauged by their ability to encode syntactical and word sense information. Researchers have explored the usage of deep-learning based features for encoding words which has lead to the development of *word embeddings*. Deep networks that predict word context have enabled improvements in semantic word representations or embeddings. To create word embeddings words are projected to a high dimensional vector space of randomly initialized model. A deep learning based model modifies these weights on a downstream task, such as predicting words

in a context window. Unlike other methods for encoding words such as Bag-of-words, the resulting representation is sparse and has the property of being similar in vector space to semantically similar words. Static embeddings such as Word2Vec and GloVe encode a vector for each word and consider all sentences where a word is present to obtain the global representations. Newer classes of models propose contextualized, contextual, dynamic, or document embeddings [1] that diverge from this concept by aggregating representations from context words before encoding. Word senses for instance of "bank" in the sentences “The river overflowed the bank” and “The bank will not accept cash on Sunday” correspond to different embedding vectors.

Trends in using features from deep neural network models are exemplified in systems for The Computational Linguistics for Psychology (CLPsych) Shared Tasks. The winning submissions of recent years, have used static sentence [43] and word embeddings [44] in their final models and some authors have argued for the importance of contextual factors [44], [45].

Recently, there has been an increase in the usage of contextualized embeddings that encode aspects of the sentence from which the words or sequence of characters (tokens) come from. When facing the more challenging task of predicting degree of suicide risk as opposed to binary risk, these features may play an important role. Pertaining to suicide risk, the most directly relevant work is Matero et al. 2019 [46]. The authors assess the risk degree of Reddit posts utilizing contextualized embeddings from a large scale pre-trained language model known as Bidirectional Encoder Representations from Transformers (BERT) [47]. The embeddings combined with a logistic regression model, achieve state-of-the-art-performance in the competition with access to all posts from a user. Another model by Mohammadi et al. 2019 [48] also made use of a type of contextualized embedding known as Embeddings from Language Models (ELMo) [49] with static embeddings and obtained state-of-the-art performance with access to only suicide-related posts or non suicide-related posts in the same suicide risk classification; however, the large scale model ensembling may

[1] The names for the contextualized embeddings are used interchangeably. Hereafter, this chapter, we will describe them with the term *document embeddings*

explain the performance increase in the latter case.

There have been improvements in contextualized language models over the years [50] and this has led to improved performance on downstream tasks, such as sentiment analysis. We seek to explore how these recent classes of embeddings perform across a variety of models in relation to both static embeddings and traditional domain-knowledge based approaches. In addition to this, we analyze features of the dataset and the models which allow us to explore the benefits and limitations, inform future models and highlight patterns relevant to suicidal risk.

2.2 Contributions

This thesis improves upon existing computational social science approaches for mental healthcare by providing

• Models that leverage recently proposed dynamic document embeddings to obtain a better understanding of language context, and that are able to accurately distinguish which users posting in an online forum are most at risk of committing suicide.

• Analysis of the advantages and disadvantages between leveraging transfer learning from powerful deep learning models trained on large datasets, and using more interpretable features incorporating domain-specific expertise.

• Analysis of text-based signs of suicide risk, including differences in the types of posts made by the general Reddit population versus by users identified as at risk of suicide.

Chapter 3

Detection of Suicide Risk from Online Text

In this chapter, we explore the use of supervised multi-class classification to detect the risk level of users who posted on r/Suicidewatch using an interpretable set of features leveraging domain knowledge and document embedding deep learning based approaches both motivated by existing literature.

3.1 University of Maryland Suicide Risk Dataset

The labeled dataset used to construct user-level feature vectors was developed by Shing et al. [41]. It consists of an anonymized set of every available Reddit posting from 2005-2015, and an extracted set of labels for users who posted on r/SuicideWatch. Reddit data is public and users are anonymized; however, Shing et al. took an extra level of precaution by replacing Reddit ID's with numeric identifiers using Named Entity Recognition.

The authors defined four categories to consider in assessing suicide risk level (T=Thoughts of suicide, L=logistics/access, C=context, and F=feelings) based on Corbitt-Hall et al.'s [51] definitions of risk categories. 865 users were labeled on CrowdFlower by "high performance annotators (as determined by the CrowdFlower platform)" that agreed with " annotations on seven clear test examples." The authors

also created a validation set of labels to compare against.

> "In order to facilitate crowdsourced as well as expert annotation, we divided sequences of more than five SuicideWatch posts for a single user into multiple annotation units containing up to five posts each, yielding a total of 982 annotation units. (For example, a user with 12 posts would yield three annotation units of their first 5 posts, next 5 posts, final 2 posts.)"

The expert annotators consist of a suicide prevention coordinator for the Veteran's Administration, the Co-chair of the National Suicide Prevention Lifelines Standards, a doctoral student with expert training in suicide assessment and treatment, and a clinician in the Department of Emergency Psychiatry at Boston Children's Hospital. Experts either received information about the aforementioned risk factors along with the risk categories (long instructions), or brief descriptions of the risk categories (short instructions). Both were told to choose to label the post that strongly supports their judgment and that choices should never be reduced to a lower label. For the 245 users, experts had an inter-rater reliability in their agreement (Krippendorf's $\alpha = .812$). Crowdsource label consensus was reported to be Krippendorf's $\alpha = 0.5047$ for the 245 users and Krippendorf's $\alpha = 0.554$ for all 865 users. Crowdsource ratings were also notably biased toward labels indicating more severe risk. The labelers misclassify more than half of low-risk users as high risk.

The dataset used in this work was derived from Shing et al. dataset and provided by the Shared Task Organizers from CLPsych. According to the Task Organizers:

According to the authors,

> "it contains two disjoint sets users annotated on four-point scale: crowd: This annotation is done by crowdsourcers using CrowdFlower, and contains annotations for 621 users who posted on SuicideWatch subreddit and 621 control users.
>
> expert: This annotation is done by experts for 245 users who posted on SuicideWatch and 245 control users"

Post Label	Description	Train	Test
a	No Risk: I don't see evidence that this person is at risk for suicide	127	32
b	Low Risk: There may be some factors here that could suggest risk , but I don't really think this person is at much of a risk of suicide	50	13
c	Moderate Risk: I see indications that there could be a genuine risk of this person making a suicide attempt	113	28
d	Severe Risk: I believe this person is at high risk of attempting suicide in the near future	206	52

We use the 919 posts from 496 users and the following class breakdown is as follows *crowd* [41].

3.2 Feature Engineering

3.2.1 Count-based features

Bag-of-words

Bag-of-words (BOW) is a model for representing text documents. Documents d and terms t are represented in a d x t matrix where the rows contain the frequency counts of t in d.

$$BOW_{d,t} = count(d, t) \tag{3.1}$$

Term frequency-inverse document frequency

In the prior approach, words that are especially common, such as *the*, *it*, or *they* have high frequencies in many documents. This reduces the encoded similarity of semantically similar words. Term frequency inverse document frequency (tf-idf) featurization seeks to resolve this by penalizing counts of words that occur very frequently across documents.

To compute the term frequency, we take the matrix transpose of the document-term matrix from the prior BOW model.

$$tf_{t,d} = BOW_{d,t}^{T} \tag{3.2}$$

We also compute the inverse document frequency for each term (idf_t) , where N is

the total number of documents in the collection, and df_t is the number of documents in which term t occurs.

$$idf_t = N/df_t \tag{3.3}$$

The tf-idf matrix[1] is then computed as:

$$w_{t,d} = tf_{t,d} idf_t \tag{3.4}$$

3.2.2 Domain Knowledge Features

For our first approach, we capitalized on expert knowledge to create a set of features based on semantic similarity to known terms related to suicide risk. Jashinky and colleagues [52] developed a dictionary of common Twitter search terms based on known suicide risk factors such as family violence and prior suicide attempts. Experts then filtered this list by determining whether these terms were linked to posts related to genuine suicide risk.

Expert knowledge was also used in creating the CLPsych Reddit dataset used in this paper [41]. Both experts and crowd-sourced workers on the platform CrowdFlower were instructed to assess suicide risk based on four families of risk factors: thoughts of suicide, logistics (methods/access), context, and feelings. We combined these two sources of knowledge by filtering the terms proposed by Jashinsky and colleagues (2016) [52] and retaining only those terms that pertained to the four categories of risk assessed in the CLPsych dataset. The final list of terms and their associated categories that we derived is shown in the Table 3.1.

To employ these terms to create a suicide risk classifier, we compute the Word2Vec [53] embedding of each term in the list, and store it in a feature dictionary. For each post, we create 71 features based on the cosine similarity between the Word2Vec embedding of the words in the post, and the Word2Vec embeddings of the terms and topics stored in the feature dictionary.

To create the domain-knowledge features, for each topic or category of suicidal

[1] In practice, both the tf and idf terms can be normalized by a log_{10} function

Table 3.1: Categories of suicidal linguistic terms proposed by Jashinsky et al. and used in the domain-specific semantic similarity classifier.

Topic	**Term**
Thought	thoughts, used, multiple, past, suicide, thought, killing,
Methods/Access	shooting, prozac, gun, suicide, went, zoloft, alcohol, pills, range, sertraline
Context	attempted, fight, sister, parents, abused, friend, brother, tried, suicide, dad, pain
Feelings	hopeless, depressed, alone, anxious, abused, empty, impulsive, worthless, sad, feel, hurt, helpless

term (*thought, methods/access, context, feelings*), we compute aggregate statistics about the similarity of words in the post to words in the topic, including the average, median, skew, mode and max. We also compute the cosine similarity between the embeddings of each word in the post and embeddings of each word in the feature dictionary, and use the median of these values to create a similarity feature for each dictionary word.

3.2.3 Document Embedding Features

We also explored the use of new forms of deep neural network word embedding models which are able to obtain a more reliable representation of longer pieces of text, and can thus create more reliable document embeddings.

Embeddings from Language Models (ELMo)

The first large scale method to develop contextualized embeddings is known as ELMo [49]. ELMo is a bidirectional LSTM that takes input representations x_k^{LM} for token k and applies L layers of forward and backward language models (LM), which encode left and right contexts of the kth tokens to get representations $h_{k,j}^{LM}$ for each token for each layer.

$$h_{k,j}^{LM} = \left[\overrightarrow{\mathbf{h}}_{k,j}^{LM}, \overleftarrow{\mathbf{h}}_{k,j}^{LM}\right] \forall j = 1, \ldots, L \tag{3.5}$$

The initial embedding as well as the hidden representations for each layer are concatenated to create a final representation.

$$\begin{aligned} R_k &= \left\{\mathbf{x}_k^{LM}, \overrightarrow{\mathbf{h}}_{k,j}^{LM}, \overleftarrow{\mathbf{h}}_{k,j}^{LM} | j = 1, \ldots, L\right\} \\ &= \left\{\mathbf{h}_{k,j}^{LM} | j = 0, \ldots, L\right\} \end{aligned} \tag{3.6}$$

Finally for a supervised problem, ELMo flattens all layers in R_k in a single vector where the parameters γ^{task} and s_j^{task} are learned for the downstream task

$$\mathbf{ELMo}_k^{task} = E\left(R_k; \Theta^{task}\right) = \gamma^{task} \sum_{j=0}^{L} s_j^{task} \mathbf{h}_{k,j}^{LM} \tag{3.7}$$

In practice, it is recommended to utilize the contextualized embedding with the input representation of a static embedding such as GloVe x_k^{LM} [49]. For each token, it extracts the intermediate representation.

Contextual String Embeddings (FLAIR)

The recently proposed FLAIR embeddings [54] are similar to ELMo yet instead of a bidirectional word-level LSTM, the authors use a bidirectional lstm and a conditional random field (CRF). The embeddings are generated by passing sentences as sequences of characters into a *character-level* language model to obtain word-level embeddings. The word embeddings are further trained on a sequence labeling task with a CRF. [54]. Because the model operates directly on characters and does not need to limit its vocabulary by stemming words, it is able to represent the context of the sentence (for example, the tense of the words in the sentence). FLAIR embeddings have provided state of the art performance on sequence labeling tasks. When combined with hyperparameter tuning, FLAIR embeddings have shown superior performance on tasks even above fastText embeddings and Google AutoML [55].

Bidirectional Encoder Representations from Transformers (BERT)

Bidrectional Encoder Representations from Transformers (BERT) is another recent contextualized language model [47]. Unlike ELMo and FLAIR, BERT improves upon the approaches by using a bidirectional transformer language model to simultaneously attend to two contexts. It is trained using two different pre-training tasks. In the masked language modeling task, a percentage of tokens in a sentence are randomly masked, and BERT predicts the masked tokens. In the next-sentence-prediction task, BERT takes a sentence as input and predicts the next sentence. There have been numerous works on improving BERT. BERT obtained new state-of-the-art results on eleven natural language processing tasks, e.g. improving the GLUE (Wang et al.,

2018) score to 80.5%. We use the version of BERT known as Robustly Optimized BERT Pretraining Approach (RoBERTa) [56]. It was able to make a few changes to the BERT model to achieve substantial improvements. It is trained using much more training data, longer sequences and a next sentence prediction objective.

Generalized Autogregressive Pretraining for Language Understanding (XLNet)

XLNet [50] is another contextualized language model that improved upon the performance of BERT by overcoming limitations of the masked pre-training task, which assumes conditional independence in tokens, is unable to handle dependencies when predicting consecutively masked tokens and is unable to fine-tune masked tokens on the training set. To train a model that incorporates bidirectional context without the mask token and parallel independent predictions, XLNet utilizes permutation language modeling. The order of the prediction is not necessarily left to right, but sampled randomly instead. The Maximum Likelihood Estimation objective is calculated as:

$$\max_{\theta} E_{\mathbf{z}\in Z_N}\left[\sum_{j=1}^{N}\log p_\theta\left(t_{z_j}|t_{z_1},t_{z_2},\ldots,t_{z_{j-1}}\right)\right] \tag{3.8}$$

Here, XLNet samples a permutation with the order z = [z1, z2, ..., zN] from the set of all permutations Z_N. The probability of a sequence is factorized according to z, such that the z_jth token is conditioned on the previous tokens according to permutation order. The cardinality of Z_n is factorial, so XLNet conditions on part of the input:

$$\max_{\theta} E_{\mathbf{z}\in Z_N}\left[\sum_{j=c+1}^{N}\log p_\theta\left(t_{z_j}|t_{z_1},t_{z_2},\ldots,t_{z_{j-1}}\right)\right] \tag{3.9}$$

where c is the cutting point of the sequence.

3.3 Preprocessing

We removed de-identification tokens. For the count-based features, we stemmed the words to reduce the dimensionality of the features given the large vocabulary size. For the domain knowledge and document embedding features, we added spaces between

the words and before or after punctuation to improve tokenization.

We concatenated all posts belonging to a specific user and added a space between posts for the domain knowledge and count-based features. The concatenated documents ranged from 70-520 tokens. For the embedding features, we obtained token embeddings for a single post and averaged them to get a post embedding. To get a user embedding, we averaged all of the user's post embeddings.

3.4 Post processing

Given the embedding sizes ranged from, 3000 to 4096 for embedding features and our vocabulary size was 7510, we applied a principal component analysis (PCA) to reduce the dimensionality of the count-based and embedding features, while still explaining 95% of the variance in the data. For the embeddings, the average dimension size was 120.

3.5 Models

We experimented with classifiers including Random Forests, Logistic regression, SVM and neural network models. We performed hyper-parameter tuning using the validation set.

3.5.1 Classical Machine Learning Models

Support vector machine

A support vector machine is a classification algorithm that attempts to find a hyperplane with a margin that maximally separates the classes

$$\begin{aligned} &\min_{w,b,\gamma} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \gamma_i \\ &\text{such that } y_i \left(w^T \phi\left(x_i\right) + b \right) \geq 1 - \gamma_i \\ &\gamma_i \geq 0, i = 1, \cdots, n \end{aligned} \tag{3.10}$$

Where x_i, for $i = 1$ through n are training data points, C is a regularization constant, y_i for $i = 1 \ldots, n$ are the data labels, , $\phi(.)$ is the feature map of the kernel, and w, b, are the parameters we seek to learn. The decision function is

$$\hat{y}(x) = \operatorname{sgn}\left(w^{*T}x + b^{*T}\right) \tag{3.11}$$

We experimented with two kernels: a linear kernel defined as $K_1(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T\mathbf{x}' + c$ and Gaussian kernel defined as $K_2(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{2\sigma^2}\right)$

Logistic regression

Logistic regression is a generalized linear model that tries to model the posterior probabilities of K classes. The model is as follows:

$$p(y = c|x) = \frac{e^{w_c \cdot x + b_c}}{\sum_{j=1}^{k} e^{w_j \cdot x + b_j}} \tag{3.12}$$

where w is a weight vector, x is an input vector and b is the bias.

Random forest

Random forests are groups of trees. The trees are developed through bootstrap sampling of the training data, and grown by selecting a random subset of features, picking the best and splitting them. Given the ensemble of trees, a prediction is made based on voting or averaging. Depending on the complexity, random forests can be difficult to interpret, but this model class can approximate more complex functions than a decision tree.

3.5.2 Deep Learning Models

Multi Layer Perceptron (MLP)

The multi layer perceptron consists of a series of densely connected layers

$$a(x) = f(Wx + b)$$

where a is the output (also known as the activation), $f(\cdot)$ is an activation function,

W is the weight matrix, and b is a bias vector. Without applying f, this is equivalent to a linear regression model, so we use a nonlinear activation function in order to incorporate nonlinearity in the network. We experimented with the rectified linear unit (ReLU), defined as

$$f(x) = \max(0, x)$$

in the intermediate layers, and the softmax function, defined as

$$f(\mathrm{x}) = \frac{e^{x_j}}{\sum_{k=1}^{n} e^{x_k}}$$

in the last layer to obtain an output between 0 and 1.

Long Short Term Memory (LSTM)

Sequence or recurrent neural network models (RNN) are commonly used in NLP because they permit remembering values at previous time iterations. In a recurrent neural network, each element of an input embedding x_t is processed sequentially.

$$h_t = f\left(Ux_t + Wh_{t-1}\right) \tag{3.13}$$

U and W represent the weight matrices between an input and hidden states (h_t) of the recurrent connection at timestep t and the function f is a non-linear transformation such as $tanh$, $ReLU$. RNN allows for variable length processing while maintaining the sequence order. However, it is limited when it comes to long sentences due to the exponentially growing or decaying gradients. Long short term memory (LSTM) is a common way to handle such a limitation using gating mechanisms. It has additional "forget" gates over the simple RNN which enables the network to encode longer term dependencies without the vanishing gradient problem. A cell consists of three gates: input, forget and output gates. Let x_t be the input vector to the cell. The three gates at time step t are represented by i_t for input gate vector, f_t for the forget gate vector, and o_t for the output gate vector. Let C_t be the cell state and h_t be the "hidden state" vector output from the cell. Letting W and U be weight matrices and b a bias

vector, with $\odot$ representing the Hadamard (entry- wise) product, we compute the hidden state as follows:

$$\begin{aligned} f_t &= \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right) \\ i_t &= \sigma\left(W_i x_t + U_i h_{t-1} + b_i\right) \\ C_t &= f_t \odot C_{t-1} + i_t \odot \tanh\left(W_c x_t + U_c h_{t-1} + b_c\right) \\ o_t &= \sigma\left(W_o x_t + U_o h_{t-1} + b_o\right) \\ h_t &= o_t \odot \tanh\left(C_t\right) \end{aligned} \tag{3.14}$$

Gated Recurrent Units

A gated recurrent unit (GRU) is another gated RNN variant [57].

It is less complex than the LSTM, but has similar performance in most tasks. Unlike the LSTM, it only has two gates: reset gate and update gate, which handle information flow similar to an LSTM without a memory unit. This exposes the hidden content without any control. Being less complex, the GRU can be a more efficient RNN than the LSTM. Let x_t is the input vector. The gates at time step t are represented by z_t for the update gate, r_t for the reset gate, h_t for the output gate. Let W, U and b represent the weight matrices and bias vectors. The hidden state of a GRU is computed as follows:

$$\begin{aligned} z_t &= \sigma\left(W_z x_t + U_z h_{t-1} + b_z\right) \\ r_t &= \sigma\left(W_r x_t + U_r h_{t-1} + b_r\right) \\ h_t &= z_t \circ h_{t-1} + (1 - z_t) \circ \tanh\left(W_h x_t + U_h\left(r_t \circ h_{t-1}\right) + b_h\right) \end{aligned} \tag{3.15}$$

3.5.3 Hyperparameter tuning

For all model and feature combinations, we performed an extensive hyperparameter search over the parameters in Table 3.2 using a Random search over a maximum of 20 parameter combinations and sampling without replacement.

Table 3.2: Bounds for hyperparameters for benchmarking models

Model	**Hyperparameter Bounds**
Support Vector Machine	C = (1,10), Kernel = (Linear, Gaussian)
Logistic Regression	C = (.01, .09, 1, 10, 25)
Random Forest	Max depth = (10, 60, 100), Min. samples per leaf = (3,4,5), Max features = (all features, square-root of features), Min. samples per split (2,4), number of estimators = (100, 200)
Multi-Layer Perceptron	Hidden layer = ([200], [128], [256], [128, 128], [256, 256], [256, 128], [128, 256]), Activation = (Sigmoid, ReLU, Tanh)
LSTM, GRU, BiGRU, BiLSTM	Number of epochs = (2,4,6,8,10), Hidden Size = (64,128,256), Learning rate = (1E-3, 5E-4, 1E-4),

3.6 Model Evaluation

We compare classifiers based on count-based features, document embeddings derived from deep neural networks, and hand-engineered features based on domain knowledge. Table 3.3 presents the results.

To compare all of the feature sets and models, we first compute a random baseline that predicts a label from the prior distribution of classes. This classifier achieves a performance of F1 = .271, Precision = 0.267 and Recall = 0.2811.

The baseline count based features improve on this performance and have low to moderate macro average F1 scores across the models. This may be because the feature vectors do not account for word order and do not include sufficient context. Interestingly, the tfidf outperforms the BOW and represents a strong baseline. The domain-knowledge model which uses a dictionary of words to attempt to build on expert knowledge and transfer it to the task, does not demonstrate good performance for language in context as well. As an example, consider the following paraphrased post from the data set which was labeled as 'a - No risk': *''I don't really want to die, I just want the pain to stop"*. This post contains words like *pain* and *die*, which are highly similar (or the same) as words like *pain* and *killing* in the domain-knowledge feature dictionary. However, the user is explaining that they do not actually wish to commit suicide, although they are in pain. The domain-knowledge classifier is not able to understand the grammatical and semantic meaning of the words in context, and so is out-performed by the deep learning features, which are able to encode this information.

The document embeddings models achieve good performance on this task. This can be explained by the fact that the document embeddings are trained on much larger text datasets and so are able to build a more robust general understanding of language, which can be effectively transferred to the current task where labeled data are limited.

With the large embedding sizes, we applied a PCA that explains 95% of the variance to both count-based and embedding features. We find that ELMo embeddings reduced by 95% or 100 contribute the largest gains in model performance and are robust across models.

Table 3.3: F1, Precision, Recall across feature types with standard errors from 10 random initializations

Feature	SVM			Logistic Regression			Random Forest			MLP		
	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall
Domain (median)	0.327 (0.00)	0.327 (0.00)	0.314 (0.0)	0.396 (0.00)	0.278 (0.0)	0.377 (0.0)	0.321 (.015)	0.382 (.016)	0.294 (.032)	0.26 (.040)	0.252 (.014)	0.320 (.041)
Domain (mean)	0.298 (0.00)	0.363 (0.00)	0.307 (0.00)	0.257 (0.00)	0.371 (0.00)	0.265 (0.00)	0.356 (.017)	0.411 (.012)	0.337 (.037)	0.256 (.058)	0.324 (.047)	0.234 (.034)
Domain (mode)	.232 (0.00)	0.293 (0.00)	0.259 (0.00)	0.235 (0.00)	0.303 (0.00)	0.486 (0.00)	0.232 (0.00)	0.293 (0.00)	0.231 (.006)	0.209 (.020)	0.276 (0.015)	0.205 (.016)
FLAIR	0.361 (0.00)	0.389 (0.00)	0.363 (0.00)	.401 (0.00)	0.415 (0.00)	0.406 (0.00)	0.356 (.007)	0.415 (.008)	0.361 (.051)	0.289 (.025)	0.346 (.028)	0.293 (.025)
ELMo	0.377 (0.00)	0.386 (0.00)	0.372 (0.00)	0.42 (0.00)	0.43 (6.2E-17)	0.462 (0.00)	0.4 (0.00)	0.50 (0.01)	0.54 (0.03)	0.33 (.018)	0.386 (.023)	0.340 (.077)
XLNet	0.373 (0.00)	0.376 (0.00)	0.375 (0.00)	0.407 (6.21E-17)	0.439 (0.00)	0.414 (6.21E-17)	0.366 (.103)	0.415 (.009)	0.382 (.051)	0.339 (.043)	0.407 (.036)	0.300 (.046)
RoBERTa	.0.361 (0.00)	0.388 (0.00)	0.364 (0.00)	0.401 (0.00)	0.416 (0.00)	0.406 (0.00)	0.368 (.009)	0.421 (.005)	0.423 (.076)	0.326 (.018)	0.375 (.013)	0.318 (.057)
BOW	0.403 (0.00)	0.400 (6.2E-17)	0.408 (0.00)	0.371 (0.00)	0.393 (0.00)	0.354 (0.00)	0.31 (0.011)	0.377 (.013)	0.271 (.009)	0.371 (.012)	0.422 (.022)	0.378 (.042)
TFIDF	0.409 (0.00)	0.417 (0.00)	.433 (6.2E-17)	0.406 (0.00)	0.418 (0.00)	0.457 (0.00)	0.329 (.012)	0.395 (.013)	0.293 (.012)	0.33 (0.043)	0.377 (.034)	0.384 (.065)
Flair (PCA)	0.4 (0.00)	0.415 (0.00)	0.393 (0.00)	0.419 (0.00)	0.434 (0.00)	0.43 (0.00)	0.313 (.020)	0.36 (.017)	0.330 (.053)	0.302 (.057)	0.361 (.048)	0.305 (.043)
ELMo (PCA)	**0.42 (6.2E-17)**	**0.42 (0.00)**	**0.423 (0.00)**	**0.457 (0.00)**	**0.457 (0.00)**	**0.51 (0.00)**	**0.33 (.017)**	**0.383 (.015)**	**0.370 (.122)**	**0.377 (.04)}**	**0.405 (.025)**	**.401 (.060)**
XLNet (PCA)	0.352 (0.00)	0.361 (0.00)	0.35 (0.00)9	.418 (6.2E-17)	0.420 (6.2E-17)	0.452 (0.00)	0.31 (0.00)4	0.359 (.018)	.466 (.110)	0.391 (.04)	.416 (.036)	.409 (.075)
RoBERTa (PCA)	.400 (0.00)	.415 (0.00)	.393 (0.00)	.419 (0.00)	.434 (0.00)	.430 (0.00)	.323 (0.019)	.377 (0.014)	.363 (0.110)	.301 (.016)	.366(.020)	.262 (.011)

Table 3.4: F1, Precision, Recall for ELMo embeddings with standard errors from 10 random initializations

	Precision	Recall	F1
LSTM	0.46 (0.048)	0.549 (0.128)	0.444 (0.051)
Bi-LSTM	0.436 (0.044)	0.52 (0.092)	0.407 (0.05)
GRU	**0.475 (0.046)**	**0.62 (0.106)**	**0.457 (0.048)**
Bi-GRU	0.438 (0.059)	0.508 (0.131)	0.387 (0.064)

We also experiment with sequence models that fully incorporate positional information utilizing the best performing embedding features. We implemented a few RNN models such as LSTM and a unidirectional and bidirectional GRU as input shown in table 3.4. We see that among these models, the GRU with ELMo embeddings performs best with an equivalent F1 score, and better precision and recall than the logistic regression. We also note that the bidirectional models perform worse than the unidirectional models in terms of macro average F1 scores. This may be because the additional parameters are causing the models to overfit .

Interestingly, these models can also be used to make a risk level prediction after every post a user posts to make an analysis of how the user's thought changes through time.

Here's an example of how the model is able to predict the risk level after every post:

user 18233 at post 1e6gv8 is labeled *b*

user 18233 at post 1jto27 is labeled *b*

user 18233 at post 1p0bt5 is labeled *c*

user 18233 at post 244hbb is labeled *c*

The actual label for the user was "c" so the model is able to understand that after the third post the user posted in the SuicideWatch subreddit. Looking at the posts, there was some discussion of suicidal-ideation from the first few posts as well and the model is doing a good estimation of the risk after each post. This feature of sequential models can be explored further to perform a better analysis on the risk per timestep per user.

Given the model architecture, it is also possible to fine tune the weights on a

downstream task and train with smaller embeddings; however, we did not do that here because of limited data and computational resources. After doing this, we may notice an increase in accuracy.

Chapter 4

Interpretable Text-Based Risk Factors

To discern factors relevant to risk degree, we use this section to interpret patterns from the posts and classifiers.

4.1 Post Behavior

We compute the top unigrams and bigrams for each class in the training set excluding stop words and personal pronouns in Table 4.1. From the unigrams, we can see that the more frequently appearing words relate to the expression of desires and emotions, such as "help" and "want". Words are commonly used across categories; however, there is progressively more suicidal language for instance, "suicidal thoughts", "want die," and "want end" in the more severe risk categories.

We also compute the top most collocated N-grams using point-wise mutual information (PMI) in Table 4.1. This metric ranks word sets, such that the items in the set which more commonly appear together are ranked highest. This method can showcase meaningful terms or phrases in the risk categories that provide insights into language patterns. Here we apply a filter that selects collocations that appear at least three times in order to reduce the appearance of very uncommon phrases in the corpus. In the lower-risk categories, we see phrases about actions taken by other people, for instance "killed himself", "to kill herself" and "she constantly says". In the higher-risk categories, there is noticeably more collocations related to specific men-

Table 4.1: Most frequent N-grams (N=1,2)

Category	Most frequent N-grams (Ordered by frequency)
No risk unigrams	help, life, know, like, friend, get, feel, people, want, really
Low risk unigrams	know, like, life, feel, get, want, would, people, really, one
Moderate risk unigrams	like, want, know, feel, life, would, get, really, time, one
Severe risk unigrams	want, like, know, life, feel, get, time, even, one, would
No risk bigrams	(feel, like), (need, help), (get, better), (please, help), (every, day), (best, friend), (tl, dr), (felt, like), (r, suicidewatch), (someone, talk)
Low risk bigrams	(feel, like), (year, old), (need, help), (dont, know), (suicidal, thoughts), (get, better), (commit, suicide), (disgust, disgust), (friends, family), (really, bad)
Moderate risk bigrams	(feel, like), (high, school), (want, die), (want, live), (suicidal, thoughts), (need, help), (really, want), (get, better), (know, anymore), (know, want)
Severe risk bigrams	(feel, like), (want, die), (every, day), (get, better), (long, time), (really, want), (need, help), (want, end), (feels, like), (even, though)

Table 4.2: Most collocated N-grams (N=2,3)

Category	Most collocated N-grams (Ordered by PMI)
No risk bigrams	(butterfly, project), (e, mail), (tl, dr), (lt, 3), (mental, health), (multiple, times), (r, suicidewatch), (killed, himself), (9, months), (constantly, says)
Low risk bigrams	(mixed, race), (video, games), (older, brother), (kill, herself), (thursday, night), (commit, suicide), (high, school), (old, male), (suicidal, thoughts), (step, dad)
Moderate risk bigrams	(rock, bottom), (panic, attacks), (social, skills), (socially, awkward), (video, games), (moving, forward), (mental, health), (rage, against), (rage, rage), (middle, class)
Severe risk bigrams	(makeup, runnin), (san, diego), (carbon, monoxide), (downward, spiral), (minimum, wage), (gas, station), (golden, gate), (brown, belt), (tl, dr), (anti, depressants)
No risk trigrams	(entirely, your, own), (thinking, about, killing), (would, be, appreciated), (not, entirely, your), (she, constantly, says), (no, matter, how), (about, an, hour), (spend, time, with), (if, anyone, needs), (thinking, about, suicide)
Low risk trigrams	(league, of, legends), (year, old, girl), (attention, in, class), (in, high, school), (another, thing, is), (we, weren, t), (my, step, dad), (to, kill, herself), (get, away, from), (t, make, sense)
Moderate risk trigrams	(rage, rage, against), (only, thing, keeping), (go, gentle, into), (being, taken, away), (20, year, old), (rage, against, the), (not, go, gentle), (thing, keeping, me), (over, 2, years), (how, much, longer)
Severe risk trigrams	(golden, gate, bridge), (borderline, personality, disorder), (makeup, runnin, down), (playing, video, games), (falling, awayno, longer), (foster, brother, died), (low, self, esteem), (personality, disorder, anxiety), (everthe, scars, will), (signyour, tears, are)

Table 4.3: Most popular subreddits

Category	Most frequent Subreddit
Control	AskReddit (159), funny (153), pics (127), gaming (98), AdviceAnimals (89), aww (86), WTF (76), videos (75), reddit.com (59), Music (50)
No risk	SuicideWatch (127), AskReddit (79), funny (52), pics (49), aww (38), WTF (37), Music (34), gaming (34), AdviceAnimals (33), trees (29)
Low risk	SuicideWatch (50), AskReddit (28), gaming (15), pics (14), depression (13). aww (12), funny (12), videos (9), AdviceAnimals (9), askscience (9)
Moderate risk	SuicideWatch (113), AskReddit (59), funny (43), depression (31), AdviceAnimals (30), pics (28), aww (25), offmychest (23), WTF (23), explainlikeimfive (20)
Severe risk	SuicideWatch (206), AskReddit (114), depression (80), funny (63), pics (55), aww (47), AdviceAnimals (44), gaming (41), offmychest (41), explainlikeimfive (36)

tal health symptoms, reasons for suicide or suicidal plans such as, "panic attacks", "foster brother died" or "golden gate bridge".

Finally, to understand more about the behavior of the users in each risk category, we computed the top 10 subreddits (forums) for each category ordered by the number of users posting in them as shown in Table 4.1. Users in all categories post forums with entertainment themes *r/funny*, *r/gaming* and *r/aww*. Users in higher-risk categories, post in forums specific to the disclosure of mental health and otherwise personal concerns, such as *r/Depression* and *r/offmychest*.

4.2 Topic Model

To better enhance comprehension about the types of discussions taking place on Reddit and how these relate to suicide risk, we used Latent Dirichlet Allocation (LDA) to create a topic model of the dataset [58]. In this case, LDA learns a generative model of Reddit posts, in which each post can be described by a mixture of latent topics that are discovered by the model. Each topic represents a distribution over possible words. LDA has proven extremely successful for text modeling in part because it assumes that the topic distribution has a sparse Dirichlet prior, meaning that it assumes each document covers only a small number of topics, and each topic is related to a small set of important words.

The dataset included 919 posts of users on r/SuicideWatch. To train an LDA model on Reddit, we combined these posts with 919 other randomly sampled posts from Reddit users that had never posted on r/SuicideWatch. We train the LDA model on the combined dataset, to determine a broad set of topics discussed on Reddit by both suicidal and non-suicidal users. We use the average topic coherence score [59] to determine the best number of topics to describe the data. The model which obtained the best average topic coherence score of -2.207 had 7 topics. Table 4.4 presents those topics, including the words most important to each topic. Salient words that were used to decide how to label each topic are bolded. Reddit is male-dominated (69% of users are male) [60], and the topics reveal a focus on video games and technology.

Table 4.4: Reddit topics found by LDA. Bolded words represent salient terms that were used in determining the topic label.

Topic	**Terms ordered by importance**
Suicide help	**feel,** go, like, get, want, know, think, life, time, make, even, one, my, **live, try,** would, people, **feel like, never, fuck,** tell, day, take, **die,** see, say, anymore, much, thing, really, **end, kill,** way, **work, friend,** everything, leave, year, noth, love, **help,** anything, still, **depress, suicide,** better
Social relationships	want, get, know, go, like, **friend,** think, help, thing, really, feel, year, people, time, one, would, life, make, say, talk, **suicide,** even, day, tell, try, my, need, take, start, never, **work, someone,** much, back, live, see, **care,** good, **family,** find, **love,** give, could, come, **school,** something
Tech review	**game, play, use,** get, look, would, **buy, new,** one, like, time, go, be, know, run, what, if, question, this, so, good, find, work, want, people, see, **post,** thank, day, do, **cost, pc,** make, also, **team**, could, allow, guy, around, way, **reddit,**, you, **link**, how, **player**, try, any
Human rights	get, need, time, help, make, **prison**, use, **homeless,** like, one, would, thank, want, look, what, people, new, if, be, good, game, post, take, also, how, love, think, go, this, us, point, say, work, **give, human,** they, lot, nan, to, **state, fund,** many, **join,** see, find, much, still
Video games	**charge, damage,** time, have, **enemy, range,** like, **hit, counter,** use, **deal,** make, would, average, get, people, high, **play, sub,** come, need, send, you, can, this, **infinity, multiply,** wind, home, slow, also, long, **move,** first, non, thing, really, look, something, may, see, any, around, be, interest, help, **tip, fire**
General advice	gt, get, **anyone,** make, sd, **know,** need, could, say, find, use, **help,** if, like, do, work, look, one, you, be, hd, **remember, today,** how, see, give, way, **god,** new, so, build, **pharmacy,** back, think, much, come, try, start, guy, **believe,** go, also, my, month, seem, else, want
Services / sales	**free, via, craigslist, craigslist via, ifttt,** need, help, since, first, one, state, **fb,** look, **book, part,** would, be, great, use, **item, card, pick,** come, gb, **drive, level,** want, **stuff,** team, case, **game, price,** get, please, **include,** if, build, know, side, play, thank experience, **video, cpu,** table, every, **list, year, power**

There are also topics related to selling items via Craigslist, general advice, social relationships, human rights, and suicide. While the topics have some overlap (i.e. the word *game* appears in both the *tech review* and *sales* topics, they appear to be largely distinct.

Figure 4-1 shows which topics are discussed by users at different levels of suicide risk. *Non suicide watch* users have never posted on r/SuicideWatch. We see that these users, which can be conceptualized as the average Reddit user, discuss all topics fairly equally, with a particular focus on social relationships and technology. Potentially suicidal users appear to focus much more strongly on discussing social relationships and seeking help for suicide than any of the other topics. This could suggest that they use Reddit mainly as a way of seeking help, rather than engaging with its other communities. Users who have posted on r/SuicideWatch but were deemed no risk

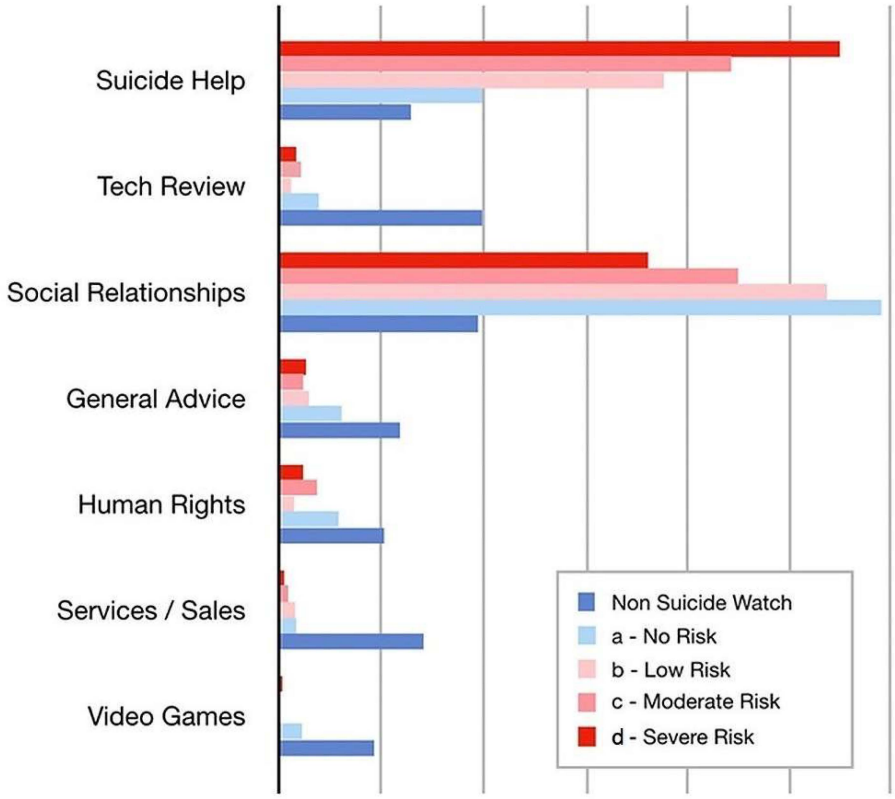


Figure 4-1: Proportion of topics discussed by Reddit users by suicide risk

(category 'a'), appear to be especially focused on social relationships.

As suicide risk increases, the focus on discussing social relationships decreases and the focus on suicide itself increases. This could suggest that problematic relationships may be a potential factor separating the groups by risk level. It also aligns with our earlier findings that words related to relationships e.g, *she constantly says* are particularly important in assessing suicide risk. Previous literature on suicide has also emphasized the importance of relationships [52].

4.3 Interpreting Classifier Outputs

In this section, we explore interpretability methods that explain model behavior (global) and explain particular examples (local) to obtain insights about features.

4.3.1 Global Interpretability

A strength of the domain-knowledge classifier is that it can provide more interpretable results. Specifically, we can assess the importance of each of the words proposed by Jashinsky and colleagues [52] in predicting suicide risk on Reddit. The words are clustered into one of four categories corresponding to the labeling criteria: feeling (F), context (C), methods/access (L) and T (thoughts) so we can see patterns among the clusters as well. While a common approach is to assess feature importance using a metric like information gain, some authors have criticized this measure as biased [61]. Therefore, we adopt the approach of Parr and colleagues (2018) in assessing word importance using *permutation importance* [62]. This is done by training a classifier and assessing how much the prediction error increases when the values of a specific feature are scrambled.

We assess *permutation importance* of the domain-knowledge features, and present the results in Figure 4-2. As stated in Section 3.We find that words about objects and substances that could actually be used to carry out a suicide (*alcohol, zoloft, gun, sertraline, prozac*) are some of the most important features. This mirrors a previous finding that *ibuprofen* and *bridge* were the most important words in determining suicide risk for Crisis Text Line [42].

Interestingly, the two most important words are *worthless* and *parents*. For the

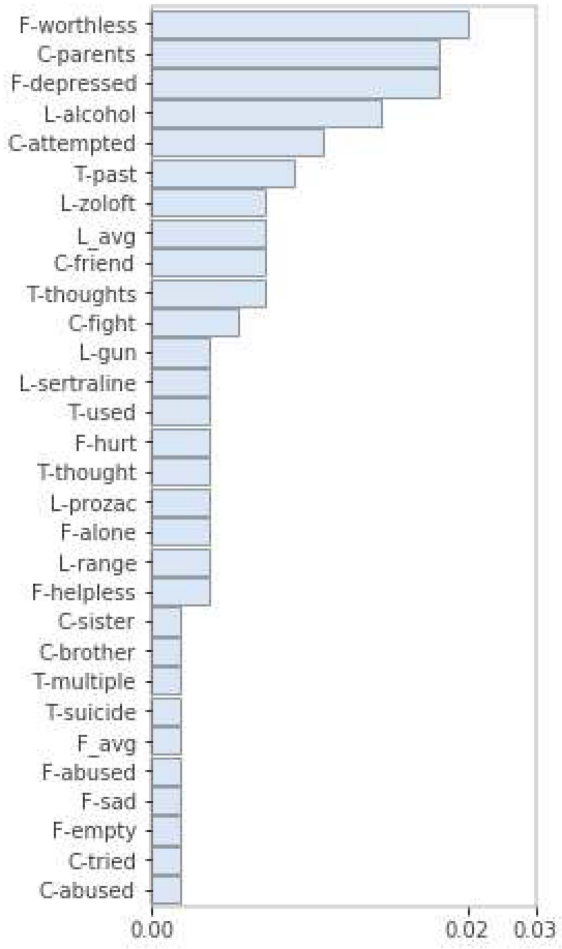


Figure 4-2: Feature importance in the domain-knowledge classifier

former, feelings of worthlessness could be an important factor given work demonstrating that among 20 symptoms for depression, worthlessness was the only symptom associated with a lifetime suicide attempt [63].

The importance of parents to suicide risk is likely to be higher among individuals less than 26 years old, a demographic among whom the risk of suicide has been increasing dramatically [29]. Because 58% of Reddit users are under the age of 29 (compared to only 22% of adults in the U.S. population) [28], the Reddit dataset is likely to be representative of this population. Note that other words related to social relationships (*friend, fight, alone, sister, brother, abused*) also feature prominently among the most important terms to the classifier.

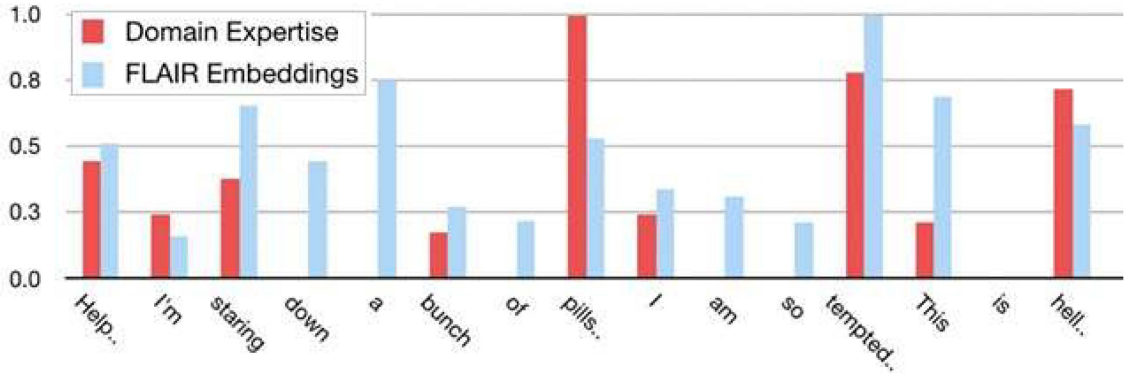
4.3.2 Local Interpretability

A generalization of the permutation importance approach can actually be applied to both black-box and domain-engineered classifiers. We can take a sample post in our dataset, and iteratively remove each word from the post and re-compute the predicted suicide risk. As an estimation of word importance, we assume that words that lead to a higher change in the predicted risk when removed have greater importance to the prediction. We use this approach to assess the importance of words in the context of the larger post with both our black-box embedding classifiers, and domain-engineered classifiers, and compare the results. A similar approach was taken by Felbo et al. [64].

The downside of this approach is it is highly computationally expensive, making it difficult to assess overall word importance across many posts. Further, even if we created such an aggregated statistic, it would only be a simplification of what the FLAIR model is actually using to make a decision. However, this technique is still useful for gaining insight into how the decision functions learned by both classifiers differ.

Using the extension to permutation importance described in Section 4.3.2, we assess the importance of each word to an example post for both the FLAIR-based and domain-engineered classifiers. As is evident in Figure 4-3, the domain-engineered classifier focuses heavily on words that are close in the embedding space to words

Figure 4-3: Word importance comparing domain knowledge and domain agnostic classifiers



in the list in Table 4.4, such as *pills*. In contrast, the predictions of the document-embeddings classifier change to at least some degree no matter which word is replaced, and more importance is placed on words that might help distinguish degree of risk, such as *tempted*. However, there are also artifacts that the model places a lot of importance on, such as a, and *This*. These words are likely common to all posts. The way the embeddings are averaged may have caused the method to place more importance on these words. Models where other averaging methods are used such as an RNN may provide a more principled way to combine the token embeddings. This will be the subject of future work.

Chapter 5

Conclusion

Under-reporting and poor access to healthcare resources limit the impact of effective treatment for suicidal patients. Automatic detection with fine-grained risk degree models may be an important tool to combat these challenges and augment suicide prevention. This work has made several contributions. This work used Reddit posts from the CLPsych competition to classify users of *r/SuicideWatch* into one of four types of suicide risk categories: no risk, low risk, moderate risk, and severe risk.

• Unlike prior work [48], [46] in online suicide risk detection, we compared the embeddings of novel models, such as RoBERTa and XLNet inspired by recent progress in transfer learning. We found that the class of features which perform the best are ELMo embeddings with a reduced number of dimensions. We built a model that takes into account the sequential nature of the text with ELMo embeddings and achieved a macro average F1 score of 0.457 on a held out test set that is a 1.68 fold increase from a random baseline. When the problem was simplified to binary classification, the binary model performance in our classifier achieved a macro average F1 of 0.92. The transfer learning methods presented here could lead to improved models which can effectively triage users that are most at risk of suicide.

• We analyzed post-level phrases and trained an LDA topic model to assess which topics are most frequently discussed by users that have varying degrees of suicide risk. Notably, we showed that there is characteristically less discussion of social relationships in the higher risk categories compared to the lower risk categories. This

mirrors work that shows that negative mood and victimization can lead to more selfish behavior [65]. We also analyzed the domain specific classifier, and we found some evidence that learned helplessness exemplified by connection to word *worthlessness*, discussion of methods for carrying out a suicide, and social relationship discussion may be important for characterizing higher risk groups. Notably, the word *parent* also appeared among features most important for the model. This finding in addition to other patterns from most frequent N-grams, *high school*, *video games*, *older brother* in the groups showcased that there may be more discussion from a younger demographic present on the forum. Because suicide rates are increasing dramatically among those 26 years and younger [29], and because Reddit users predominantly belong to this demographic [28], we believe there is promising potential for classifiers which can automatically determine suicide risk from online posts to provide help for this sub-population.

• Finally, we developed a method that allows us to interpret why the contextualized models may perform better than an ontology based-model and showcased how the model attends to all words in a sequence instead of only words relevant to the risk class.

5.1 Limitations and Future Work

Some modeling limitations of this work include not exploring alternate types of pooling operations for the models and not performing fine-tuning to benefit from the task-specific embeddings. We also did not have the resources to compute smaller embeddings, and resorted to a sub-optimal method of computing smaller embeddings using principal component analysis. For our domain-specific models, we could have used a more expansive ontology to understand terms specific to each risk class. Finally, our interpretability method could have been compared against other common methods, such as saliency gradients [66] and Locally Interpretable Model Agnostic Explanations (LIME) [67] to along with experts to further validate their utility.

Perhaps the most important limitation of this work is the limited ecological validity

of the data due to crowd-sourcing the labels for the four levels of risk, and not relating them to actual suicidal behaviors. When the labeling task was presented to experts, it achieved a high concordance, indicating that the chosen four categories can help validate the *risk*. However, labeling of text, even by an expert, is not as strong of a ground truth as having outside validated data that an individual actually was suicidal. Although, there is much promise in providing online models that can directly respond to potentially suicidal victims, computational social scientists studying suicide risk should strive to work with psychiatrists to obtain better sources of ground truth data, including data from health records such as in Jordan et al. (2018) [24] and Kessler et al. (2019) [25].

In the future, we would like to experiment with other features and architectures along with additional ways of combining expert knowledge with transfer learning from deep neural networks trained on large data sets to improve performance. For instance, one line of work could be to directly utilize the large scale models from which we extracted embeddings and fine tune them on the dataset. A second line of work could investigate the usage of graph neural networks which can incorporate expert knowledge or structure, and benefit from the expressivity of neural networks, as preliminarily demonstrated in [68]. Given the difficulty to distinguish between similar levels of risk, we would also consider training a hierarchical ensemble of classifiers to first distinguish whether a user is at risk of suicide or not, and then determine the level of risk within those users deemed suicidal. All of these classes of models may still benefit over time with external information such as digital sensors, and we likely can achieve higher accuracies for detecting risk levels with the augmented data.

Bibliography

[1] Centers for Disease Control, Prevention, et al. Suicide rising across the us. *Retrieved from*, 2018.

[2] Kenneth D Kochanek, Sherry L Murphy, Jiaquan Xu, and Elizabeth Arias. Deaths: final data for 2017. 2019.

[3] Julie Cerel, Margaret M Brown, Myfanwy Maple, Michael Singleton, Judy van de Venne, Melinda Moore, and Chris Flaherty. How many people are exposed to suicide? not six. *Suicide and Life-Threatening Behavior*, 49(2):529–534, 2019.

[4] Donald S Shepard, Deborah Gurewich, Aung K Lwin, Gerald A Reed Jr, and Morton M Silverman. Suicide and suicidal attempts in the united states: costs and policy implications. *Suicide and Life-Threatening Behavior*, 46(3):352–362, 2016.

[5] World Health Organization. Mental health: Suicide data, 2016.

[6] Holly Hedegaard, Sally C Curtin, and Margaret Warner. Increase in suicide mortality in the united states, 1999–2018. 2020.

[7] Rory C O'Connor and Matthew K Nock. The psychology of suicidal behaviour. *The Lancet Psychiatry*, 1(1):73–85, 2014.

[8] Keith Hawton, Carolina Casañas i Comabella, Camilla Haw, and Kate Saunders. Risk factors for suicide in individuals with depression: a systematic review. *Journal of affective disorders*, 147(1-3):17–28, 2013.

[9] World Health Organization et al. *Preventing suicide: A global imperative*. World Health Organization, 2014.

[10] E David Klonsky, Alexis M May, and Boaz Y Saffer. Suicide, suicide attempts, and suicidal ideation. *Annual review of clinical psychology*, 12:307–330, 2016.

[11] Brian K Ahmedani, Christine Stewart, Gregory E Simon, Frances Lynch, Christine Y Lu, Beth E Waitzfelder, Leif I Solberg, Ashli A Owen-Smith, Arne Beck, Laurel A Copeland, et al. Racial/ethnic differences in healthcare visits made prior to suicide attempt across the united states. *Medical care*, 53(5):430, 2015.

[12] Joseph C Franklin, Jessica D Ribeiro, Kathryn R Fox, Kate H Bentley, Evan M Kleiman, Xieyining Huang, Katherine M Musacchio, Adam C Jaroszewski, Bernard P Chang, and Matthew K Nock. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. *Psychological Bulletin*, 143(2):187, 2017.

[13] Catherine M McHugh, Amy Corderoy, Christopher James Ryan, Ian B Hickie, and Matthew Michael Large. Association between suicidal ideation and suicide: meta-analyses of odds ratios, sensitivity, specificity and positive predictive value. *BJPsych open*, 5(2), 2019.

[14]

[15] Rebecca Resnick. Psychological assessment: The 'not good enough' state of the art. In *Veterans Affairs Suicide Prevention Innovations Conference (VASPI)*, 2016.

[16] David Shaffer, Madelyn S Gould, Prudence Fisher, Paul Trautman, Donna Moreau, Marjorie Kleinman, and Michael Flory. Psychiatric diagnosis in child and adolescent suicide. *Archives of general psychiatry*, 53(4):339–348, 1996.

[17] J John Mann, Alan Apter, Jose Bertolote, Annette Beautrais, Dianne Currier, Ann Haas, Ulrich Hegerl, Jouko Lonnqvist, Kevin Malone, Andrej Marusic, et al. Suicide prevention strategies: a systematic review. *Jama*, 294(16):2064–2074, 2005.

[18] Rebecca Rossom, Greg Simon, Arne Beck, Brian Ahmedani, Bradley Steinfeld, Michael Trangle, and Leif Solberg. Facilitating action by learning healthcare systems for suicide prevention. *Psychiatric services (Washington, DC)*, 67(8):830, 2016.

[19] Gregory E Simon, Karen J Coleman, Rebecca C Rossom, Arne Beck, Malia Oliver, Eric Johnson, Ursula Whiteside, Belinda Operskalski, Robert B Penfold, Susan M Shortreed, et al. Risk of suicide attempt and suicide death following completion of the patient health questionnaire depression module in community practice. *The Journal of clinical psychiatry*, 77(2):221, 2016.

[20] Karen J Coleman, Eric Johnson, Brian K Ahmedani, Arne Beck, Rebecca C Rossom, Susan M Shortreed, and Greg E Simon. Predicting suicide attempts for racial and ethnic groups of patients during routine clinical care. *Suicide and Life-Threatening Behavior*, 49(3):724–734, 2019.

[21] Gregory E Simon, Carolyn M Rutter, Do Peterson, Malia Oliver, Ursula Whiteside, Belinda Operskalski, and Evette J Ludman. Does response on the phq-9 depression questionnaire predict subsequent suicide attempt or suicide death? *Psychiatric services*, 64(12):1195–1202, 2013.

[22] Matthew K Nock, Franchesca Ramirez, and Osiris Rankin. Advancing our understanding of the who, when, and why of suicide risk. *JAMA psychiatry*, 76(1):11–12, 2019.

[23] Robert J Ursano, Lisa J Colpe, Steven G Heeringa, Ronald C Kessler, Michael Schoenbaum, Murray B Stein, and Army STARRS Collaborators. The army study to assess risk and resilience in servicemembers (army starrs). *Psychiatry: Interpersonal and Biological Processes*, 77(2):107–119, 2014.

[24] Pascal Jordan, Meike C Shedden-Mora, and Bernd Löwe. Predicting suicidal ideation in primary care: An approach to identify easily assessable key variables. *General hospital psychiatry*, 51:106–111, 2018.

[25] Ronald C Kessler, Samantha L Bernecker, Robert M Bossarte, Alex R Luedtke, John F McCarthy, Matthew K Nock, Wilfred R Pigeon, Maria V Petukhova, Ekaterina Sadikova, Tyler J VanderWeele, et al. The role of big data analytics in predicting suicide. In *Personalized psychiatry*, pages 77–98. Springer, 2019.

[26] Joel A Fein, Megan E Pailler, Frances K Barg, Matthew B Wintersteen, Katie Hayes, Allen Y Tien, and Guy S Diamond. Feasibility and effects of a web-based adolescent psychiatric assessment administered by clinical staff in the pediatric emergency department. *Archives of pediatrics & adolescent medicine*, 164(12):1112–1117, 2010.

[27] Per E Kummervold, Deede Gammon, Svein Bergvik, Jan-Are K Johnsen, Toralf Hasvold, and Jan H Rosenvinge. Social support in a wired world: use of online mental health forums in norway. *Nordic journal of psychiatry*, 56(1):59–65, 2002.

[28] Pew Research Center. Distribution of reddit users in the united states as of february 2016, by age group, 2016.

[29] Jean M Twenge, Thomas E Joiner, Megan L Rogers, and Gabrielle N Martin. Increases in depressive symptoms, suicide-related outcomes, and suicide rates among us adolescents after 2010 and links to increased new media screen time. *Clinical Psychological Science*, 6(1):3–17, 2018.

[30] David C Mohr, Mi Zhang, and Stephen M Schueller. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. *Annual review of clinical psychology*, 13:23–47, 2017.

[31] Lei Zhang, Xiaolei Huang, Tianli Liu, Ang Li, Zhenxiang Chen, and Tingshao Zhu. Using linguistic features to estimate suicide probability of chinese microblog users. In *International Conference on Human Centered Computing*, pages 549–559. Springer, 2014.

[32] Carla Agurto, Pat Pataranutaporn, Elif K Eyigoz, Gustavo Stolovitzky, and Guillermo Cecchi. Predictive linguistic markers of suicidality in poets. In *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, pages 282–285. IEEE, 2018.

[33] Chris Poulin, Brian Shiner, Paul Thompson, Linas Vepstas, Yinong Young-Xu, Benjamin Goertzel, Bradley Watts, Laura Flashman, and Thomas McAllister. Predicting the risk of suicide by analyzing the text of clinical notes. *PloS one*, 9(1):e85733, 2014.

[34] John Pestian, Henry Nasrallah, Pawel Matykiewicz, Aurora Bennett, and Antoon Leenaars. Suicide note classification using natural language processing: A content analysis. *Biomedical informatics insights*, 3:BII–S4706, 2010.

[35] Glen Coppersmith, Ryan Leary, Patrick Crutchley, and Alex Fine. Natural language processing of social media as screening for suicide risk. *Biomedical informatics insights*, 10:1178222618792860, 2018.

[36] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. In *Seventh international AAAI conference on weblogs and social media*, 2013.

[37] Naoki Masuda, Issei Kurahashi, and Hiroko Onari. Suicide ideation of individuals in online social networks. *PloS one*, 8(4), 2013.

[38] Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 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*, pages 2098–2110, 2016.

[39] M Johnson Vioulès, Bilel Moulahi, Jérôme Azé, and Sandra Bringay. Detection of suicide-related posts in twitter data streams. *IBM Journal of Research and Development*, 62(1):7–1, 2018.

[40] Stevie Chancellor and Munmun De Choudhury. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine*, 3(1):1–11, 2020.

[41] Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 25–36, 2018.

[42] S Reardon. Ai algorithms to prevent suicide gain traction. *Nature*, 64, 2017.

[43] Sunghwan Mac Kim, Yufei Wang, Stephen Wan, and Cecile Paris. Data61-csiro systems at the clpsych 2016 shared task. In *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*, pages 128–132, 2016.

[44] Edgar Altszyler, Ariel J Berenstein, David Milne, Rafael A Calvo, and Diego Fernandez Slezak. Using contextual information for automatic triage of posts in a peer-support forum. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 57–68, 2018.

[45] Yufei Wang, Stephen Wan, and Cécile Paris. The role of features and context on suicide ideation detection. In *Proceedings of the Australasian Language Technology Association Workshop 2016*, pages 94–102, 2016.

[46] Matthew Matero, Akash Idnani, Youngseo Son, Salvatore Giorgi, Huy Vu, Mohammad Zamani, Parth Limbachiya, Sharath Chandra Guntuku, and H Andrew Schwartz. Suicide risk assessment with multi-level dual-context language and bert. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 39–44, 2019.

[47] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[48] Elham Mohammadi, Hessam Amini, and Leila Kosseim. Clac at clpsych 2019: Fusion of neural features and predicted class probabilities for suicide risk assessment based on online posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 34–38, 2019.

[49] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

[50] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5754–5764, 2019.

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

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

[53] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[54] Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, 2018.

[55] Tadej Magajna. How to beat google's automl - hyperparameter optimisation with flair.

[56] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[57] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[58] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[59] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408. ACM, 2015.

[60] Pew Research Center. Distribution of reddit users in the united states as of february 2016, by gender, 2016.

[61] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, 8(1):25, 2007.

[62] Terence Parr, Kerem Turgutlu, Christopher Csiszar, and Jeremy Howard. Beware default random forest importances, 2018.

[63] Hong Jin Jeon, Jong-Ik Park, Maurizio Fava, David Mischoulon, Jee Hoon Sohn, Sujeong Seong, Jee Eun Park, Ikki Yoo, and Maeng Je Cho. Feelings of worthlessness, traumatic experience, and their comorbidity in relation to lifetime suicide attempt in community adults with major depressive disorder. *Journal of affective disorders*, 166:206–212, 2014.

[64] Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *arXiv preprint arXiv:1708.00524*, 2017.

[65] Emily M Zitek, Alexander H Jordan, Benoît Monin, and Frederick R Leach. Victim entitlement to behave selfishly. *Journal of personality and social psychology*, 98(2):245, 2010.

[66] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[67] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[68] Manas Gaur, Amanuel Alambo, Joy Prakash Sain, Ugur Kursuncu, Krishnaprasad Thirunarayan, Ramakanth Kavuluru, Amit Sheth, Randy Welton, and Jyotishman Pathak. Knowledge-aware assessment of severity of suicide risk for early intervention. In *The World Wide Web Conference*, pages 514–525, 2019.