Detecting Suicide Risk on Reddit

Alton Lu altonl/lualton@cmu.edu
Heinz College
Carnegie Mellon University
Pittsburgh, PA, United States

1. Abstract

This paper discusses the implementation of a machine learning task on The University of Maryland Reddit Suicidality Dataset (Shing 2018). The dataset encompasses user-level posts on the social media site, Reddit.com and limited meta-data about the posts, such as time and subreddit (discussion board topic). This paper demonstrates an attempt to predict varying levels of suicidal risk among users of Reddit.com based on posting habits and other text engineered features.

This research project involves classifying users into separate suicide risk categories. The results show that binary classification of users into [no risk, some risk] classes is successful with access to limited meta-data. However, classifying users into specific categories of risk [low risk, moderate risk, severe risk] is much more challenging. This project will discuss the difficulties of this multiclass classification and the importance of meta-data.

2. Introduction

On November 15th, 2018, the Centers for Disease Control and Prevention released a report stating that suicide rates have reached a 50-year peak in the United States. The director of the CDC National Center for Injury Prevention and Control stated that 'Increasing suicide rates in the U.S. are a concerning trend that represent a tragedy for families and communities and impact the American workforce. Knowing who is at greater risk for suicide can help save lives through focused prevention efforts' (Hellman).

Suicide risk is a challenge to classify. Suicide is often a by-product of other mental illness, such as depression and anxiety. In order for patients to be evaluated, they must often seek expensive medical care.

There has been increasing amounts of research around utilizing social media to detect these types of mental illness and suicidal risk factor. The goal is to either speed up potential diagnosis or provide users with information about resources they can utilize.

Social media posts and the associated meta-data can contain useful signals about a users mental health. This paper attempts to leverage data from the social media site, Reddit.com, in order to predict suicide risk of current users.

3. Background and Related Work

This section will discuss the literature around applying computational methods to mental health issues on social media. Namely, previous research around detecting depression on twitter, datasets for study, and many of the common issues that arise.

Coppersmith et al., (2014) continues to influence most research in this area. Coppersmith provided a method to find pseudo-labels for users on Twitter. They utilize user self-diagnosis and announcement as a means of finding ground truth. A user would state 'Ive been clinically diagnosed with depression' as a means of determining depression. On the other hand, a user who stated 'Ive never been so depressed. The Steelers can't beat the Patriots,' as a disingenuous signal.

Balani et al., (2015), and Guntuku et al., (2017) experiment with various ways of extracting useful features from textual and meta-data present within social media data to reasonable success. Balani utilized a self-disclosure keyword mechanism, such as the user stating whether they wanted to die or not, as a key feature. Balani also utilized key metadata such as upvotes and downvotes of posts, response time, and follower details. Guntuku reviewed three separate studies that used screening surveys, public sharing on Twitter, and membership in online forums to predict depression.

Mowery et al., (2017) created a Twitter corpus and tested the predictive power of depression-related keywords. Their findings show that keywords are quite limited in their ability to predict depression due to context-related issues. A common challenge centered on differing context for certain words. A user might discuss the next economic depression, which is is an example of an an inaccurate signal. What appeared to be most predictive was phrases that described depression-like symptoms (loss of energy, cant concentrate, etc).

Shu et al., (2017) performed a study to detect fake news on social media. Shu dissected approaches to this detection of fake that mirror similar approaches in detecting mental illness. Namely, Shu described two overarching methods in social media detection. The first is *content-based* methods, which focus on style, linguistic features, or knowledge-based computational methods of the text. These include the classical natural language processing, such as extracting features from text (n-grams, sentiment, parts-of-speech tagging). The second is the *social-based* method, which describes and quantifies the data around the content. In the fake news context, this may be how quickly a news article is shared or the number of likes by certain political leaning individuals. This can be extended to mental illness context, considering how often a user posts, when they post, what are the social-media reactions (likes, shares, responses) to derive other features that describe a users activity.

Perez et al., (2018) discuss the importance of meta-data in identifying users and how meta-data should be considered as sensitive as the content of a message. Perez demonstrates that the application of a simple supervised learning algorithm with meta-data enables identification of any user in a group of 10,000 with approximately 97 percent accuracy. Even with adding noise to 60 percent of the training data, their algorithm still performed at 95 percent accuracy.

4. The Dataset

The data was provided by Philip Resnik at the University of Maryland and was developed in Shing et al., (2018). The data comes from user posts on Reddit.com.

4.1 Reddit.com

Reddit is a social media website that centers around text discussions. Users can post a link to a URL or a personal note that other users on the website can discuss. This is a very different format than other social media sites like Twitter and Facebook. Twitter caps posts at 280 characters and the focus of content is more on short thoughts or updates. This is reflected in Twitters continuing evolution into a news based social media site. Facebook has always been about connection to friends, family, and pages you find interesting. The focus is on what those individual parts are doing.

Reddit differs in a few key ways. The first is anonymity. Unlike Twitter and Facebook where most users have a profile with their real name and pictures, Reddit does not. It takes a couple seconds to create a new Reddit account and all posts are anonymous unless the users personally give identifying information. The second is that Reddit is focused around discussions. The cap for posts is 40,000 characters. The third is that the structure of Reddit is around topics . Reddit hosts subreddits which are specific discussion sub-forums centered around a topic and any user can create a new one. For example, the NFL subreddit focuses on discussions around the National Football League (NFL). Posts that arent relevant to discussions around the NFL are not allowed in this forum. There is a coffee subreddit, where users discuss everything around coffee. As of 2018, there are an estimated 1.2 million different subreddits.

Researchers at the University of Maryland developed this dataset as a means of providing more possibilities to research mental health on social media. The data was created in the following process:

1. They took a corpus of every publicly available Reddit post from January 1, 2008 to August 31, 2015.

2. They marked a positive signal for each user that posted in the SuicideWatch subreddit.

3. Eliminated users with less than 10 posts.

4. Aggregated a control group of users that did not post in any mental health subreddit.

4.2 Annotating the Data

There were approximately 934 users that were signaled using the Coppersmith 2014 methodology as being positive for suicide risk. The data was then annotated by a

panel of experts and a crowdsourced group of users from CrowdFlower.com according to suicide risk categorization in Corbitt-Hall et al. (2016). The panel and the Crowd were asked to annotate subsets of the 934 users for the categories: **(a) No Risk**: no evidence for suicide risk; **(b) Low Risk**: There are factors that suggest risk, but probably not; **(c) Moderate Risk**: There are signs that there is a genuine risk of this person making a suicide attempt; **(d) Severe Risk**: This person is at high risk of attempting suicide in the near future.

4.3 Structure of the Data

The dataset includes user Id, post title, post body, timestamp, subreddit. There are also separate risk labels for experts and crowds. An example of user data is below:

Table A: Simplified User Data.

User ID	Subreddit	Post Title	Post Body
100	Awww	Look at how cute...	
100	Movies	Star Wars is not...	Episode 8 ruined Star...
200	News	Apple announces three new...	
200	Politics	Trump announces Supreme...	What do people...
203	Coffee	Cold Brew?	Cold Brew is the superior...

Table 1: User Data

Table B: Suicide Postings by Users.

User ID	Subreddit	Post Title	Post Body
809	SuicideWatch	Giving up and why we...	Does anyone else just...
9282	SuicideWatch	Whats the point to any...	Every day i wake up...
2132	SuicideWatch	I almost did it	I'm a coward for not...
203	SuicideWatch	Thank you to SuicideW...	I looked down a shotgun...

Table 2: User Data

4.4 Preprocessing Decisions

There are several decisions made around preprocessing that would impact overall results. For the purposes of simplification, I assigned users to the Expert risk label. The Crowd label was only used in cases where there was no Expert label available. In cases where the Crowd and Expert had both annotated a user, I used the Expert label.

URLs and emojis were simplified to the token url or the emojis to corresponding text.

Post Titles and Post Body were kept separate but went through normal tokenization and stemming processes.

Where post title or post body was null, I filled in with an empty string character.

5. Two Tasks

5.1 Binary Classification

The first is a binary classification of users as a suicide risk or not suicide risk. This classification tries to determine which users have at least a low level of suicide risk or absolutely no suicide risk at all. The context for Reddit.com would be to try and understand which users are more likely to be a suicide risk. Often, we want to be able to simply classify users based on whether they have any level of risk against those who have none.

5.2 Multiclass Classification

The second is a multiclass classification task. Users have been previously defined by the experts and crowdsourced assessment as one of [No risk, low risk, moderate risk, severe risk]. This multiclass classification task can be constructed in two separate ways: **a)** classify the entire sample into either one of the four categories, **b)** classify the sample that has been previously classified as a potential suicide risk.

The context for this task would be to stratify users based on risk level. While low risk of suicide is still unfortunate, the users arent in any immediate danger or harm. However, understanding users at severe risk may give Reddit the opportunity to stage a limited intervention, such as a support message similar to Facebook and Twitter. The importance of moderate risk users are around worsening conditions. These are the users that we would have an opportunity to provide real treatment before conditions reach a point where harm is very likely in the near future.

6. Feature Extraction

I utilize a few natural language processing techniques to extract features from the text. All features were extracted using NLTK (Loper), Spacy, Gensim (Radim), textstat, or scikit-learn (Pedregosa) in Python.

6.1 Sentiment

For each title and post body, I extracted a sentiment score of negativity, neutrality, and positivity.

6.2 Subjectivity

For each title and post body, I calculated a score that depicted where the text was on a subjectivity range. Low numbers indicated a text was extremely objective (meaning

free from emotions or personal feelings) and high numbers indicated a subjective text (one influenced by personal feelings and opinions).

6.3 Readability Ease

This is a method developed in Farr et al., (1951). Reading ease is calculated with the following formula: Reading Ease = 206.835 - (1.015 x Average Sentence Length) (84.6 x Average Syllables per Word).

6.4 Grade-level of Text

This method was developed in Kincaid et al., (1975) to score the difficulty of material. The Grade-level is calculated with the following formula: Grade = (0.39 x Average Sentence Length) + (11.8 x Average Syllables per Word) - 15.59.

6.5 Bag of Words

The post title was represented as a bag of words vectors, composed of unigrams and bigrams.

6.6 Syntactic Features

- Punctuation Ratio: A ratio of punctuation to words.
- Average length: The average length of posts and titles.
- Average sentence length

6.7 Frequency of Posts

How often a user posted on Reddit.com during the specified timescale of our data. This a potentially a meaningless feature, as a user can certainly be commenting on Reddit without making posts. This is a limitation of our data.

6.8 Mental Disease Lexicon

A count of a post title and body that matched a mental disease lexicon as created in Yazdavar et al., (2017).

6.9 Encoding Subreddits

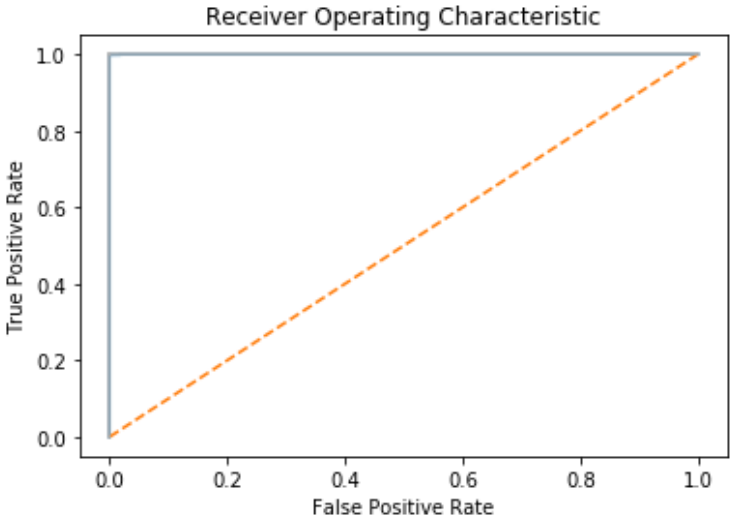
I utilized one-hot-encoded specific mental health subreddits. These included depression, SuicideWatch, anxiety, Schizophrenia.

7. Binary Classification

This section discusses the binary classification task. Due to the nature of the data, this task is extremely easy. Fitting a simple Nave Bayes algorithm can get near 100 percent accuracy in classifying users to either suicide risk or non-risk. Shown below is the confusion matrix and ROC curve. As shown the ROC curve can be considered a perfect classifier, as it completely overlaps with a perfect true positive rate.

Table 1: Binary Classification Confusion Matrix

Predicted	No Risk	Suicide Risk
Actual		
No Risk	5573	21
Suicide Risk	4	5531



This accuracy is due to the presence of the subreddit feature. Whether a user posts in mental health subreddits, such as SuicideWatch and Depression, are highly indictive of at least modest levels of suicide risk. Part of this is due to the nature of the data collection. This dataset has higher samples of those with some mental illness who are also at risk for suicide because part of the selection criteria came from posting on SuicideWatch.

However, the predictive power of this simple feature presents the power that meta-data has in a simple classification algorithm.

8. Multiclass Classification

This section discusses the multiclass classification task and pulls in more features from the feature extraction process. The goal is to stratify users in the classes of no risk, low risk, moderate risk, and severe risk. The following table represents a random forest algorithm with 10-fold cross validation and normalized features. The cross-validation score is 35

Table 2: Multi-class Classification Confusion Matrix

Predicted	No Risk	Low Risk	Moderate Risk	Severe Risk
Actual				
No Risk	41	6	26	20
Low Risk	17	4	16	16
Moderate Risk	31	17	40	39
Severe Risk	38	12	58	52

From this accuracy, we see that there isnt a specific pattern of inaccuracy. Misclassifications occur roughly equally across all classes. However, in particular, Im more concerned with he large number of moderate and severe risk individuals that were predicted as no risk. From our context-setting, these are the individuals that we would want to be able to track with reasonable certainty in order to act. But if our model has such poor sensitivity, there is no value in the predictions.

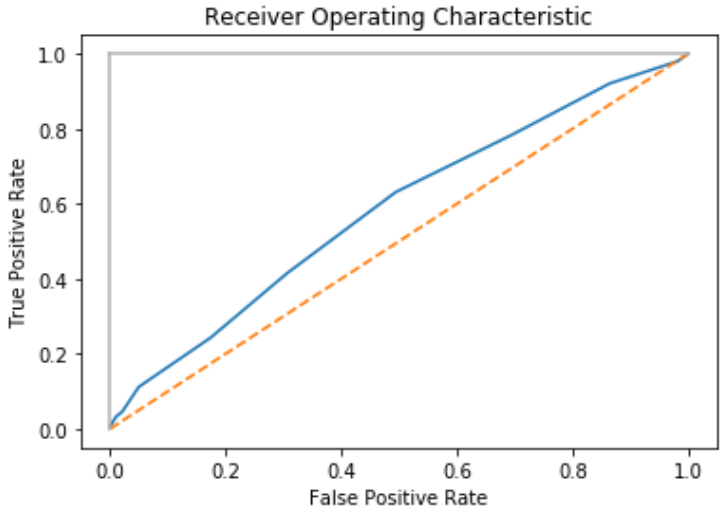
8.1 Combining classes

From this outcome, I considered binarizing the multi-class task. While this paper first discussed the possibility of predicting no risk vs suicide risk, this multiclass binarizing will bin the classes into two separate groups. The first combined class will be [no risk, low risk] into a single minimal group. The second class combines [moderate risk, severe risk] into a single class of high risk.

We run the exact same algorithm with 10-fold cross validation and get the following results.

Table 3: Binary (Multiclass) Confusion Matrix

Predicted / Actual	Moderate or Severe	Low or No Risk
Moderate or Severe	232	49
Low or No Risk	115	37



9. Discussion

This project demonstrates the difficulty in differentiating suicidal risk based only textual features. Because of the limited scope of the data, determining the difference between two individuals relies on the assumption and hope that there is a fundamental difference in how suicidal and non-suicidal users talk online. This doesnt seem to be the case and is demonstrated very simply with exploratory data analysis.

Many of the previous studies in this area, such as work with Twitter and Facebook, make extensive use of meta-data. This is intuitively more useful, as meta-data tends to describe how a user acts online and the previous studies have demonstrated that there is a fundamental difference in how depressed people act compared to non-depressed users.

The Reddit Suicide Dataset is limited because of this lack of meta-data. In the binary classification, we see the value of meta-data with the fact that users post in the common mental health subreddits [SuicideWatch, Depression, etc] as signal enough to classify as risky.

In the multiclassification, we see very poor performance, barely above random guessing. And these inaccuracies do not follow any type of pattern. The model crosses up many terms and doesnt particularly over or under predict a specific class.

10. Conclusions and Future Work

10.1 Improvements in Data

Continued improvement in data. This paper demonstrates the limitation of content-based approaches to detecting mental illness. With a pure content-based approach, the hypothesis is whether users at suicide risk talk differently than those not-at-risk

on social media. And as demonstrated, there are small differences, but the differences disappear when we begin stratifying based on severity of suicide risk.

Thus, continued improvement in access to richer data is important. The Reddit dataset in this paper is an amazing, but limited resource without the metadata about the posting, location of user, etc. However, as noted in the background section, there are continued privacy concerns that should not be overlooked.

10.2 Reframing as Anomaly Detection

From the simple task of identifying individuals based on no risk or some risk, the use of metadata is shown to be valuable. Thus, there is not much value in computational methods in that area. However, there are specific parts of mental illness that these data can attempt to improve.

We should start thinking about this detection of suicide risk as an anomaly detection problem. Given the type of textual features Ive extracted, we get a linguistic characteristic of each user. Some users are more negative on average, some are more positive. Some write with a lot of clarity and others might need to re-take a 5th English class. However, these are averaged characteristics over time. Thus, a potential way of thinking is to detect when a certain timeframe of text drastically changes negative or increases in mental disease keywords.

The idea is that we wish to intervene or at least support users that may be struggling with suicide ideation. But as previously stated, low risk ideation is more of a manageable symptom. Its at the point when the low risk turns moderate, or the moderate turns severe, that intervention should be had. Research in this area has suggested that many suicides are based on event-related causes, such as bullying, child abuse, or substance abuse. Capturing these changes in behavior is possible through computational methods and is an area to consider researching.

11. Acknowledgements

This paper was written with IRB approval from Carnegie Mellon Universitys Institutional Review Board. The data set was provided by the University of Maryland in collaboration with the American Association of Suicidality.

This research was conducted after reviewing Gaffney and Matias (2018). Gaffney investigated and found large-scale missing data in the Reddit corpus from which this suicide dataset was derived. As I mentioned previously, my approach to missing data was to fill in empty strings. This has an impact on the computational methods and is a weakness of Reddit data.

This research was also conducted after reviewing Benton et al, 2017. Benton discussed ethical research protocols when dealing with health data.

References

[1] Sairam Balani and Munmun De Choundhury. 2015. Detecting and Characterizing Mental Health Related Self-Disclosure in Social Media. Association for Computing Machinerys (ACM) CHI conference 15.

[2] Adrian Benton, Glen Coppersmith, Mark Dredze. 2017. Ethical Research Protocols for Social Media Health Research. Proceedings of the First Workshop on Ethics in Natural Language Processing pages 94-102

[3] Scottye J. Cash and Jeffrey A. Bridge. 2009. Epidemiology of Youth Suicide and Suicidal Behavior. Current Opinion in pediatrics vol. 21 pages 613-619.

[4] Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting Depression and mental illness on social media: an integrative review. Current Opinion in Behavioral Sciences, pages 43-49.

[5] Glen Coppersmith, Mark Dredze, Craig Harman. 2014. Quantifying Mental Health Signals in Twitter. Workshop on Computational Linguistics and Clinical Psychology, pages 51-60.

[6] Devin Gaffney and J. Nathan Matias. 2018. Caveat emptor, computational social science: Large-scale missing data in a widely-published Reddit corpus. PLoS ONE 13(7): e0200162.

[7] Jessie Hellmann. 2018. CDC: Suicide rates increasing among American workers. The Hill.

[8] Matthew Honnibal and Mark Johnson. 2015. An Improved Non-monotonic Transition System for Dependency Parsing. Association for Computational Linguistics. Pages 1373-1378

[9] Edward Loper and Steven Bird. 2002. NLTK: the Natural Language Toolkit. Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistic pages 63-70.

[10] Danielle Mowery, Hilary Smith, Mike Conway. 2017. Understanding Depressive Symptoms and Psychosocial Stressors on Twitter: A Corpus-Based Study. J Med Internet Res.

[11] Pedregosa et al 2011. Scikit-learn: Machine Learning in PythonJMLR 12, pp. 2825-2830.

[12] Beatrice Perez, Mirco Musolesi, Gianluca Stringhini. 2018. You are your Metadata: Identification and Obfuscation of Social Media Users using Metadata Information. Association for the Advancement of Artificial Intelligence.

[13] Rehurek Radim and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks pages 45-50.

[14] Han-Chin Shing, Philip Resnik, et al. 2018. Expert, Crowdsourced, and Machine Assessment of Suicide Risk via Online Postings. Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, pages 2536.

[15] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. ACM SIGKDD Explorations Newsletter: Social and Information Networks.