APPROVAL SHEET

Title of Thesis: CLASSIFICATION OF SOCIAL MEDIA DATA FOR SUICIDAL IDEATION

Name of Candidate: Purva Anand Kulkarni
M.S. Computer Science,
2017

Thesis and Abstract Approved: __________________

Dr. Yelena Yesha
Professor
Department of Computer Science and Electrical Engineering

Date Approved: __________________

ABSTRACT

Title of Thesis: CLASSIFICATION OF SOCIAL MEDIA DATA FOR SUICIDAL IDEATION

Purva Anand Kulkarni, M.S. Computer Science, August 2017

Thesis directed by: Dr. Yelena Yesha, Professor

Department of Computer Science and Electrical Engineering

Social media use continues to grow worldwide and an ever-growing number of people are using various social media platforms to update their social circles on their mental health challenges and suicidal ideations in real time. Number of psychological studies show that expressing suicidal thoughts and attempting suicide happens within a matter of hours and therefore automatic detection and analysis of social media posts by vulnerable users, serves as a critical, real-time window into their health and safety.

In this thesis, we are interested in classifying data from Twitter and Reddit as ‘Suicidal Risky Expression’ and ‘Non-Risky Expression’. We propose a method which includes automatic collection of tweets (Twitter data) and posts (Reddit data) based on suicidal vocabulary, parsing and tokenizing collected textual data, passing this data through a trained neural network and segregating data into two classes ‘Suicidal data’ and ‘Non-suicidal data’. Since this process runs in real time, the classified data can be used to support at-risk users either by reporting suicidal content to behavioral crisis response teams or by connecting people to mental-health support resources in real-time window.

Keywords: Social Media, Suicide Ideation, Suicide Prevention, Twitter, Reddit, Data Classification, Neural Network, Natural Language tokenization

CLASSIFICATION OF SOCIAL MEDIA DATA FOR SUICIDAL IDEATION

By
Purva Anand Kulkarni

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, Baltimore County, in partial fulfillment
of the requirements for the degree of
M.S. Computer Science
2017

Dedicated to my family, friends, and mentors

Acknowledgements

I would like to thank my advisor, Dr. Yelena Yesha for her invaluable guidance in my research. I am greatly indebted to her continuous support, patience, and motivation. I would like to express my sincere thanks to Dr. Tim Oates for all his help and guidance for this thesis. His keen insights and observations were very crucial to the progress of our research. I would also like to thank Dr. Milton Halem and Dr. Tim Oates for agreeing to be on my committee.

I am grateful to Dr. Bruce Herman and Dr. Jason Schiffman for their valuable guidance in analyzing and classifying the data for our experiments. Their expertise in psychology and suicide prevention helped me correctly understand the need and relevance of these experiments.

I would like to express my deepest gratitude towards my parents, Anand Kulkarni and Anagha Kulkarni and my sister Pooja Kulkarni and brother-in-law Mandar Joshi for being my strength and always being there to guide me on the right path, through the good times and the bad! I would like to extend special thanks to my best friend Yashodhan Joshi for his constant support and motivation. A big thanks to all my friends who made this master's journey memorable!

Table of Contents

List of Tables

List of Figures

Chapter 1

INTRODUCTION

Attempting suicide is not just a public health issue, it also has strong socio-economic consequences. According to World Health Organization (World Health Organization, 2017), close to 800,000 people take their own life and there are many more people who attempt suicide every year. Every suicide is a tragedy that affects families, communities, and entire countries. Suicide occurs throughout the lifespan and was the second leading cause of death among 15–29-year-olds globally in 2015. As per American Foundation for Suicide Prevention (American Foundation for Suicide Prevention, 2017), suicide is the 10th leading cause of death in the US. Each year 44,193 Americans die by suicide. Suicide costs the $51 Billion annually. Approximately 25 suicide attempts are made for every reported death by suicide. Many suicide attempts, however, go unreported or untreated. Surveys suggest that at least one million people in the U.S. each year engage in intentionally inflicted self-harm. Suicide is a serious public health problem, however, suicides are preventable with timely, evidence-based and often low-cost interventions. Understanding the ways in which people express their suicidal ideation is the key to prevent such suicidal attempts.

Suicidal ideation is defined as thoughts about killing oneself, while suicidal behavior is harming oneself with the intention of causing death. Though not all individuals experiencing suicidal ideation will make a suicidal attempt, but such thoughts place these individuals at increased risk of taking their own lives (O'Dea et al., 2015). Some individuals discuss about their suicidal thoughts with their friends or family members, but it is found that many do not disclose their suicidal thoughts or intent. Recently, such vulnerable individuals are found broadcasting their suicidal feelings on social media such as Twitter, Reddit, Facebook, Tumblr, et al., connoting that these social media sites can be used as suicide prevention tools.

Though psychologist and psychiatrists are concerned about challenges in identifying genuine mental health crises from false alarms online, they seem optimistic about using social media data for analyzing various ways of expressing suicidal ideation and for suicide prevention whenever possible. According to Nadine J. Kaslow (PhD, ABPP, the American Psychological Association's 2014 president), it is important for therapists to encourage patients who might be interested to use social media sites or mobile apps as they see fit in a way that integrates them with the treatment (Speaking of Psychology: Preventing Suicide, Episode 14). Dr. Robert E. Accordino mentioned that social media platforms could also provide resources to help individuals understand when they should intervene and how to intervene with a friend online. This new culture of sharing and connectivity provides new opportunities to apply preventive measures and practice timely interventions. He also added that social media can be used to improve mental health and reduce incidents of suicide (Dr. Robert E. Accordino, the White House Fellow to the Secretary of Defense, 2017). As per Glen Coppersmith, a data scientist and psychologist, gleaning suicide related information from social media, as well as other data generated by mobile devices, could be vital to psychologist looking for subtle hints about a patient's risk for suicide (Why scientists think your social media posts can help prevent suicide, 2016).

For this thesis, we are collecting data from Twitter and Reddit. Twitter is a free broadcasting social media site that allows registered users to interact with others in real-time via a web interface or mobile app. Message broadcasted on Twitter is called as a 'tweet' which can be of 140 characters or less. Roughly one-quarter of online adults (24%) use Twitter (Greenwood, Perrin, Duggan, 2016) and around 200 billion tweets are sent per year (Twitter Inc, 2017). Reddit is a free social news aggregation and discussion website. Reddit's registered users can submit text posts or direct links to be read and voted by other registered users via a web interface. Content entries on Reddit are organized by areas of interest called 'subreddits'. As of 2017, Reddit had 234 million unique users and it is ranked as 9th most visited web-site in the world and 4th most visited website in the US (Wikipedia – Reddit).

Both Twitter and Reddit have set practices to deal with self-harm and suicide, but that relies solely on the discretion of networked users, who may have difficulty in understanding genuine risk. Twitter asks its users to alert the team devoted to handling threats of self-harm or suicide if any user encounters such threats on Twitter (Twitter Inc.). Reddit has subreddit called 'Suicide Watch', where Reddit users can post their intent or thoughts about self-harm or suicide and other users and moderators can respond to them as per their understanding of the concerned post, how Reddit defines it – non-judgmental peer support ONLY (Reddit Inc). Though there is a lot of research going on in the field of identification of suicidal risky expressions on social media sites; valid, reliable, and acceptable methods of online detection have not yet been fully established, allowing new researchers to come up with innovative methods and practices.

Our study aimed to automatically detect and classify data (tweets and posts) from Twitter and Reddit in real time, which makes direct or indirect textual reference to suicidal thoughts. For our experiments, we extracted tweets and posts based on suicide vocabulary (Suicide Vocabulary Word List). Under the guidance of subject matter experts (psychologist and psychiatrist), human coders classified these tweets and posts as suicidal-risky content and non-risky content. After parsing and tokenizing this data using Natural Language Toolkit, we passed it through a trained Neural Network to get predicted class values. The feasibility of this automatic prediction was examined using recall and precision metrics.

Automatic detection of suicidal intentions will help in increasing the possibility of arranging timely, evidence-based intervention for vulnerable users. Also, detailed analysis of identified tweets and posts can help psychiatrists, psychologists and mental health workers gain insights in emerging causes and methods for suicide attempts. We have not kept our scope limited only to tweets and posts by individuals, but we are also extracting retweets (responses to the tweets) and comments (for posts), which depicts how individuals and their social network interact about suicide related contents. This may help defining new policies and methods for mental health awareness and suicide prevention on community level.

The following chapters of this thesis are organized as - Chapter 2 discusses background and related work. Chapter 3 describes the datasets we used for our experiments and for training the Neural Network. In Chapter 4, we talk about

Architecture and implementation details of our system. Chapter 5 describes the experimental setup and analyses the results of our experiments and we conclude and discuss future work in Chapter 6.

Chapter 2

RELATED WORK

2.1 Sentiment Analysis for Social Media Data

Sentiment Analysis on social media is nothing but analyzing social media posts and discussions to figure out whether it has positive connotation or negative one. Analyzing sentiment helps us understand what the person posting a message is feeling. Knowing the emotion behind the online message can provide important context for understanding and responding to that post.

Till recently, sole reason behind sentiment analysis on social media was marketing. Sentiment analysis tools such as Hootsuite Insights, Brandwatch (Hootsuite Inc), Twitter Advanced Search (Twitter Inc), Semantria (Lexalytics Inc), Rapidminer (Rapidminer Inc) are developed with focus on understanding customer emotions and reactions for brands, products, and services.

(Godbole, Srinivasaiah, Skiena 2007) presented a system that assigns scores indicating positive or negative opinion to each distinct entity in news and blogs. They achieved it by sentiment lexicon generation. They evaluated this sentiment lexicon generation by a test called un-test. The prefixes un- and im- generally negate the sentiment of a term. Thus, they made the terms of for X and unX to appear on different end of sentiment spectrum. Their plans were to analyze the degree to which their sentiment indices predict future changes in popularity and market behavior.

(Kouloumpis, Wilson, Moore 2011) investigated utility of linguistic features for detecting the sentiment of Twitter messages. They used Twitter hashtags to identify positive, negative, and neutral tweets for training three-way sentiment classifier. Along with hashtag dataset they used emoticon dataset, since emoticons are getting used for expressing various emotions effectively and using less number of characters. Their results show that microblogging features such as positive/negative/neutral emoticons, presence of intensifiers along with part-of-speech are helpful in sentiment analysis.

2.2 Social Media dealing with Suicidal Posts

Individuals at-risk of committing suicide, do want to alert those around them that something is seriously wrong. It is found that many of these warning signs of suicide are visible on social media. Figures 1.1, 1.2, 1.3 and 1.4 show some of the conversation (related to suicide) on various social media sites. Several of the most popular social media sites have ways and policies to respond to users expressing suicidal ideation.

Fig 2.1 Suicide note on Tumblr

Fig 2.2 Discussing suicidal intention on Facebook

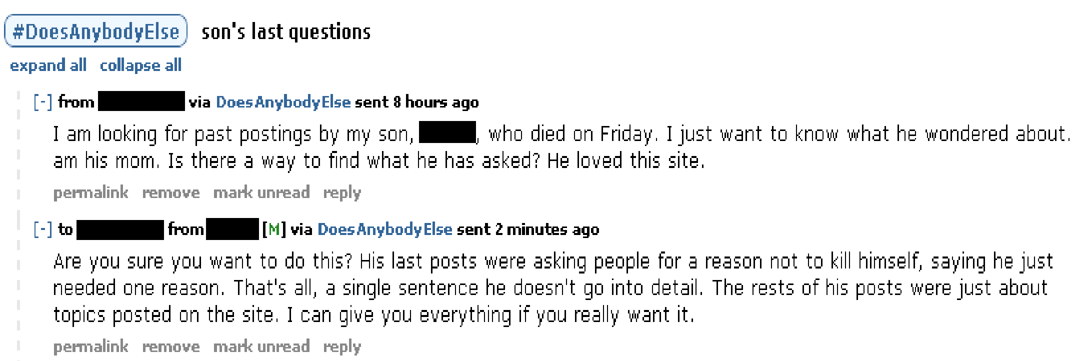


Fig 2.3 Post suicidal discussion on Reddit

Fig 2.4 Suicidal Tweets

Facebook works with National Suicide Prevention Lifeline to help vulnerable Facebook users. If anyone reports a post with suicidal content on Facebook, that post is reviewed by a team at Facebook and a notification is sent to at-risk user that a friend is concerned about their safety and given suggestions to seek help (Facebook Inc). Figure 1.5 shows Facebook response message to at-risk user.

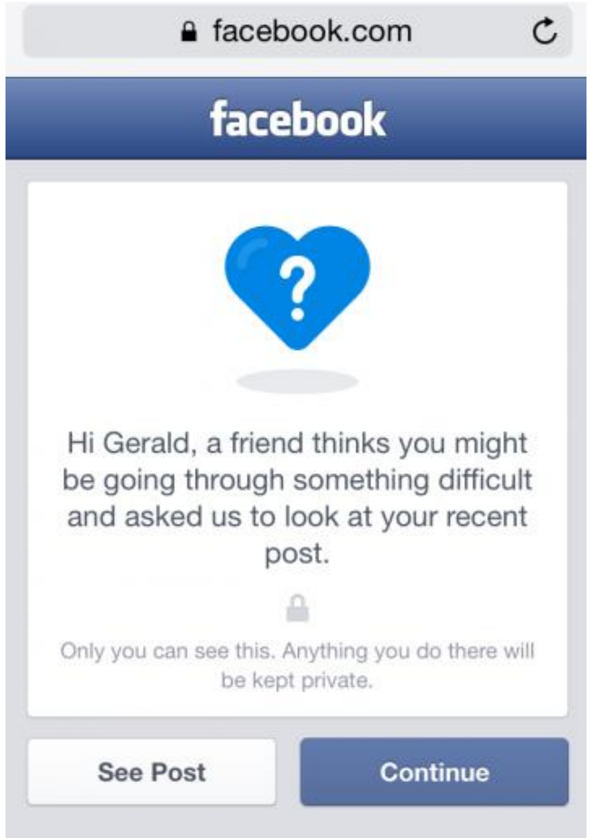


Fig 2.5 Facebook responds to at-risk user

If a tweet is flagged for suicide risk, Twitter's team for handling threats of self-harm and suicide will reach out to let at-risk user to know that someone is concerned about their wellbeing, to share the resources and to encourage them to get help. Twitter also encourages users who are concerned for other vulnerable users to reach out them and help them find support (Twitter Inc).

When a user on Tumblr comes across a suicidal user, s/he can write an email to Tumblr about suicidal user with as much information as possible including the URL of Tumblr blog. A member of Tumblr's Safety Team will send the suicidal user an email with the Lifeline number and possibly a link to chat with a Lifeline counselor (Oath Inc).

Reddit has a subreddit called Suicide Watch. It is not a suicide prevention hotline. It only offers nonjudgmental peer support out of genuine concern. Also, there are other subreddits like 'suicideprevention', where other Reddit users share information about suicide prevention hotline, help centers and counselling centers with at-risk users (Reddit Inc).

All the above-mentioned policies are at users' discretion and chances of suicidal content go unreported are high due to lack of understanding, challenges involved in identifying genuine risk, hesitation in reporting these cases, time elapsed between submission and reading the post, etc.

2.3 Automatic Detection of Social Media Data for Suicidal Ideation

Automatically (with minimal or no involvement of human users) detecting suicidal content and at-risk users on social media is the real problem. A lot of research is going on in this area to come up with valid, reliable, and real-time detection and response methods for the safety and wellbeing of vulnerable users.

(Jashinsky, Burton, Hanson, West, Giraud-Carrier, Barnes, Argyle 2013, United States) worked on identifying suicide-related risk factors through Twitter conversations by matching on geographical suicide rates from vital statistics data. They observed a strong correlation between Twitter data (segregated by states in the US) and actual state age-adjusted suicide data. They concluded that at-risk individuals can be detected through social media and Twitter may be the viable tool for real-time monitoring of suicide risk factors on large scale. (Abboute, Boudjeriou, Entringer, Aze, Bringay, Poncelet 2014, France) worked on mining Twitter data for suicide prevention. They implemented automatic language processing and learning methods to identify Twitter messages indicating suicidal risky content. (Colombo, Burnapa, Hodoroga, Scourfield 2015, United Kingdom) worked on understanding the connectivity and communication of Twitter users expressing suicidal thoughts. They analyzed characteristics of social networks of suicidal users by building social network graphs. Analysis of these graphs showed high degree of reciprocal connectivity between the authors of suicidal content as compared to other Twitter users, indicating a tightly coupled virtual community. Also, analysis of retweet graphs showing connectivity between users posting suicidal content with users who are not suicidal, suggest a potential for information exchange and possibility of contagion effect. (O'dea, Wan, Batterham, Calear, Paris, Christensen 2015, Australia) examined whether the concern for a suicide-related tweet (Twitter post) could be determined solely on the content of the tweet, as judged by human coders, and then replicated by machine learning. They found that individuals use Twitter to express suicidal intent and such Twitter data needs to be investigated thoroughly. However, they further conclude that the predictive power for actual suicidal behavior is not yet known and their findings do not directly identify targets for intervention. (Birjali, Beni-Hssan, Erritali 2016, Morocco) proposed a system which will automatically analyze sentiments in tweets to detect suicidal content using various machine learning algorithms and measured effectiveness of performance in terms of recall, precision and accuracy on sentiment analysis. (Liu, Cheng, Homan, Silenzio 2017, Hong Kong, United States) studied approaches for annotation and label aggregation to train a classifier to detect suicide-related tweets. They collected tweets using keyword and location based queries. They got data annotated by crowdsourced workers and experts. They experimented with different ways for aggregating multiple annotations and comparing their effects. Their results show that, though domain knowledge and experience are necessary in labeling the suicide-related data, multiple crowdsourcing workers (with high inter-annotator agreement), can provide reliable quality of annotations.

Although, there are available resources and tools for sentiment analysis and opinion mining, and there is a lot of research going on, worldwide, in the field of finding various methods to automatically detect suicide related content online, no system/platform/method can be considered as the best practice yet. This gives an opportunity to new researchers to explore innovative ways and experiment with data at hand to come up with promising results.

In this thesis, we are collecting data from Twitter and Reddit, predict the class label (suicidal risky or non-risky) for given textual message using a trained Neural Network and compare the prediction results for Twitter and Reddit data. We are extracting data from these two social media sites in real-time. We are interested in understanding the differences and similarities in the prediction results for the data extracted from Twitter and Reddit, when the data is passed through similar process. In a way, we are trying to find out a process for predicting suicide related content on social media, which can be applied to the textual data from different sources and gives results with comparable accuracy.

Chapter 3

DATASET

This chapter describes the data collection methods and datasets used in our experiments. To the best of our knowledge, definitive datasets containing tweets and posts (Reddit posts) with suicidal content, where user attempted suicide after posting the content on social media (Twitter and Reddit), is not available. We devised a data collection process, which is described in following sections, to collect tweets and posts in real time. We created suicidal vocabulary to search related content online. Sentiment Polarity Movie review dataset (Pang and Lee, Cornell University) used to train Neural Network, is the only dataset that is publicly available.

3.1 Suicide Vocabulary

Since we are collecting data from Twitter and Reddit in real-time, we are using keyword based queries to collect data containing words getting used in suicidal expressions. We created a list of 170 keywords. We obtained a list of 149 words occurring frequently in suicidal content (Suicide Vocabulary Word List). We also added some more words which were lacking in the list and found in suicidal contents online. Table 3.1 lists all the 170 words that constitute suicidal vocabulary used in our experiments.

Table 3.1 Suicide Terms

act	crying	feelings	life	problem	suspicion	never
addiction	culture	fight against	likelihood	psychiatry	suicide	not
advice	curable	finality	loss	puzzle	suicidal	strong
advise	data	finances	love	questions	ocd	feel
aftermath	death	frenetic	medicine	raw emotion	tears	fate
anger	demons	fret	memorial	recovery	teenagers	unfortunate
assistance	depression	frustration	men	rehabilitation	tendency	no one
attempt	despair	funding	mental	research	terminal	nobody
attention	devastating	genes	mental health	revelation	tragedy	I want to die
behavior	diagnosis	grief	misunderstood	risk	treatable	signing-off forever
bipolar	die	happening	mourn	sad	treatment	cut myself
brain	difficult	hard	mystery	scared	troubled	kill myself
brood	discovery	harm	needs	schizophrenia	thought	better-off without me
causes	discuss	haunting	normality	sense	unresponsive	last day of my life
claim	disease	health	numbers	sensitive	unworthy	final goodbye
commit	disorder	heartbreak	obsessive-compulsive	shock	veterans	I will be gone forever
compassion	dramatic	help	oscillation	socially	vexing	suicde
comprehension	education	hopelessness	overdose	solution	discrimination	suicid
compulsion	emotion	humankind	pain	sorrow	why	sucide
conflicted	end	illness	percentage	stigma	woe	selfharm
confront	enough	impact	personal	stress	women	
consequences	estimate	incoherent	practitioner	struggles	wrists	
control	experience	injury	presage	stunned	wrong	
controversy	failure	inside	prescription	suffering	want	
counselor	fear	legend	prevention	support	do not	

3.2 Twitter Data Collection

We retrieved our Twitter Dataset by using Twython 3.4.0 API for Python. Twython is actively maintained, pure Python wrapper for the Twitter API. It supports both normal and streaming Twitter APIs (Twython 3.4.0 documentation). We extracted tweets by executing keyword (from suicide vocabulary) based search queries. Also, for our experiments, we are extracting tweets in English only. Our process retrieves 100 tweets per keywords at a time. Fig 3.1 shows sample tweet format after execution of a search query using Twython. The retrieved tweet has multiple fields as seen in Fig 3.1. We have mainly concentrated on ‘text’ field which contains actual textual tweet content.


```
{‘statuses':
[{'created_at': 'Thu Jun 22 16:45:22 +0000 2017',
 'contributors': None, 'favorited': False,
'is_quote_status': False,
 'user':
 {'followers_count': 134910,
 'screen  name': 'Lesism',
'time_zone': 'London',
 'favourites  count': 51141,
  'description': 'The greatest dreams are achieved with open eyes and a conscious mind.        GB❤',
 'listed  count': 2074,
 'geo_enabled': False,
'profile  image  url': 'http://pbs.twimg.com/profile  images/815550901008138240/tKiFU  bZ  normal.jpg', 'id  str': '163535738',
 'friends_count': 101203,
'name': 'Les Floyd',
'protected': False,
 'contributors  enabled': False,
 'entities':
{'created  at': 'Tue Jul 06 17:26:49 +0000 2010', 'lang': 'en', 'default  profile': False},
 'text': "@RMD1035 She's dead, Richard...\n\nPolice found the body of a girl in landfill... suggests murder or suicide... and I…
https://t.co/qhhNY8d9ts",
'geo': None,
 'source': '<a href="http://twitter.com" rel="nofollow">Twitter Web Client</a>',
 'in_reply_to_status_id': 877929912077737984,
 'truncated': True,
 'place': None,
 'id_str': '877930521795407875',
 'retweet  count': 0,
 'entities':
 {'user  mentions': [{'id  str': '39051086', 'id': 39051086, 'indices': [0, 8], 'name': 'Richard Dickinson', 'screen  name': 'RMD1035'}],
'urls': [{'expanded_url': 'https://twitter.com/i/web/status/877930521795407875', 'display_url': 'twitter.com/i/web/status/8…', 'indices':
[117, 140], 'url': 'https://t.co/qhhNY8d9ts'}], 'symbols': [], 'hashtags': []},
'metadata':
{'iso  language  code': 'en', 'result  type': 'recent'},
'lang': 'en',
'retweeted': False}]}
```


Fig 3.1 Tweet Format

3.3 Reddit Data Collection

To extract Reddit data, we used the Python Reddit API Wrapper (PRAW 4.3.0). We are using Script Application with PRAW (PRAW documentation). We wrote queries which will extract ‘title’ from all subreddits which contain terms in suicide vocabulary. Our focus is to extract data in English language only. Fig 3.2 shows format of Reddit post retrieved after execution of a search query.

domain: self.Prevent_Suicide
id: 5rptnr
author: Flyer127
score: 2
selftext: Personally I was going to attempt suicide at 4-6 years old, i held a pizza cutter to my throat and thought that it was sharp enough to take my life while my parents were laying on their bed assuming that it was a joke, I then held a knife to my throat and said I was going to do it, my parents then told me that it's forbidden and you will go to hell and all that stuff which ultimately stopped me from committing suicide, however if suicide wasn't forbidden in my religion I would have been in a grave right now
I would also love to hear your attempts and who stopped you.
title: Did you ever attempt suicide? How old were you at that time?
url: https://www.reddit.com/r/Prevent_Suicide/comments/5rptnr/did_you_ever_attempt_suicide_how_old_were_you_at/

Fig 3.2 Reddit post format

3.4 Sentiment Polarity Movie Review Data

We used Sentence Polarity Dataset v1.0 from Sentiment Polarity Movie Review Dataset from Cornell University. The dataset contains 5331 positive and 5331 negative sentences (Pang and Lee, Cornell University). We used this dataset to train our network to distinguish between positive and negative sentiments in tweets and posts.

3.4.1 Examples of Positive sentences from the dataset

- the result is something quite fresh and delightful.
- it's great escapist fun that recreates a place and time that will never happen again.
- this beautifully animated epic is never dull.
- while parker and co-writer catherine di napoli are faithful to melville's plotline, they and a fully engaged supporting cast . . . have made the old boy's characters more quick-witted than any english lit major would have thought possible.

- whereas oliver stone's conspiracy thriller jfk was long, intricate, star-studded and visually flashy, interview with the assassin draws its considerable power from simplicity.
- funny, sexy, devastating and incurably romantic.
- this humbling little film, fueled by the light comedic work of zhao benshan and the delicate ways of dong jie, is just the sort for those moviegoers who complain that 'they don't make movies like they used to anymore'.
- a tour de force of modern cinema.

3.4.2 Examples of Negative sentences from the dataset

- simplistic, silly and tedious.
- it's so laddish and juvenile, only teenage boys could possibly find it funny.
- a film that loses sight of its own story.
- trying to make head or tail of the story in the hip-hop indie snipes is enough to give you brain strain -- and the pay-off is negligible.
- the ending doesn't work . . . but most of the movie works so well i'm almost recommending it, anyway -- maybe not to everybody, but certainly to people with a curiosity about how a movie can go very right, and then step wrong .
- a mess. the screenplay does too much meandering, norton has to recite bland police procedural details, fiennes wanders around in an attempt to seem weird and distanced, hopkins looks like a drag queen.
- it's a bad sign when you're rooting for the film to hurry up and get to its subjects' deaths just so the documentary will be over, but it's indicative of how

uncompelling the movie is unless it happens to cover your particular area of interest.

- such a fine idea for a film, and such a stultifying, lifeless execution.
- enigma is well-made, but it's just too dry and too placid.

3.5 Positive and Negative words

We obtained an extensive dataset consisting of words labeled as positive and negative (Hu and Liu). The dataset consists of 2007 positive words and 4884 negative words. We added some words to the existing dataset, as we find them appearing in tweets and posts and lacking in the original dataset. Table 3.2 shows sample positive words and Table 3.3 shows sample negative words from our dataset.

Table 3.2 Sample Positive Word List

adroit	advanced	adventuresome	affability	affectionate
adroitly	advantage	adventurous	affable	affinity
adulate	advantageous	advocate	affably	affirm
adulation	advantageously	advocated	affectation	affirmation
adulatory	advantages	advocates	affection	affirmative

Table 3.3 Sample Negative Word List

abnormal	abomination	abrasive	absentee	absurdness
abolish	abort	abrupt	absent-minded	abuse
abominable	aborted	abruptly	absurd	abused
abominably	aborts	abscond	absurdity	abuses
abominate	abrade	absence	absurdly	abusive

Chapter 4

ARCHITECTURE AND IMPLEMENTATION

In this chapter, we present architecture of our system and implementation details to achieve automatic, real-time detection of suicidal content in social media data.

4.1 Architecture

In architecture, we show how the four datasets: Sentence Polarity Movie Review Dataset, Twitter Dataset, Reddit Dataset, and Positive-Negative words dataset come together to create the input required by the neural network. Fig 4.1 gives the overview of system's architecture.

We implemented this system using Python 3.5.2. For editing, running, and testing our code and visualizing intermediate results, we used Jupyter Notebooks. The Jupyter Notebook App is a server-client application that allows editing and running notebook documents (contain Python code and rich text elements) via a web browser (Jupyter Notebook documentation). Subsequent sections in this chapter describe implementation stages in detail.

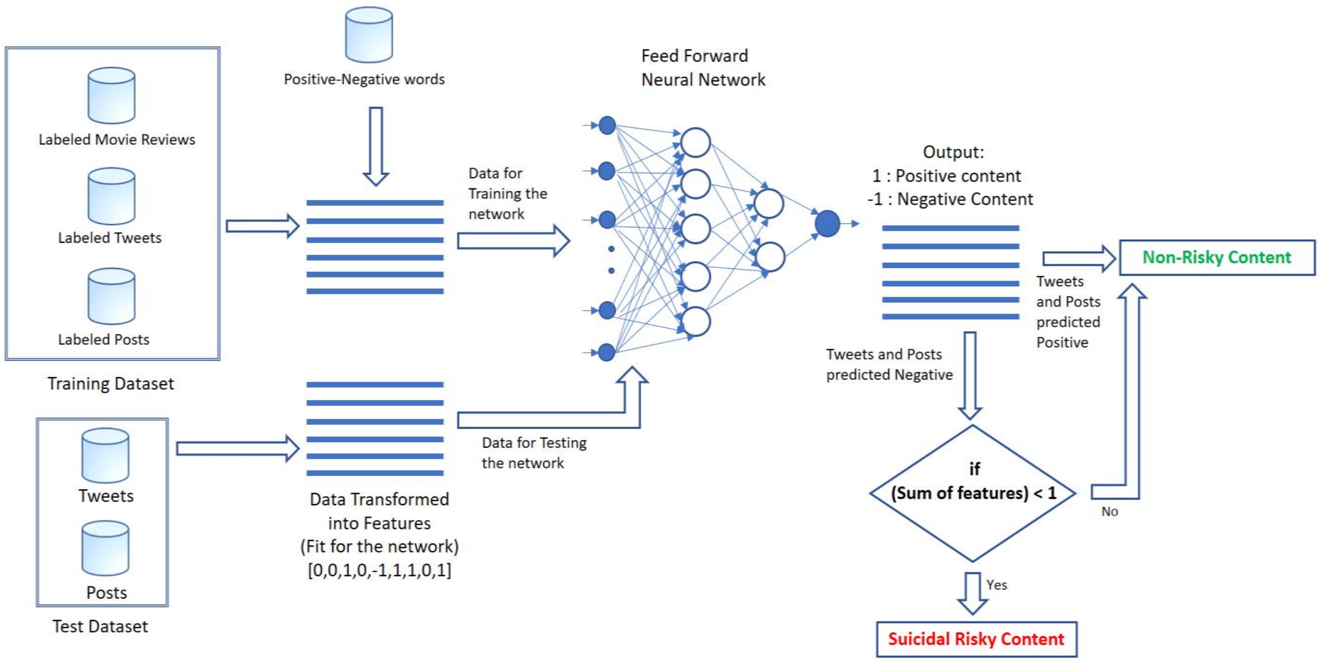


Fig 4.1 Architecture Overview

We used Scikit-Learn package (sklearn.neural_network) (scikit-learn 0.18.2 documentation) in Python for implementing the neural network. We used Multi-Layer Perceptron Classifier (sklearn.neural_network.MLPClassifier). MLPClassifier helps us define the parameters such as number of hidden layers and nodes in each hidden layer, solver algorithm, number of iterations and learning rate. The model we used is Multi-Layer Perceptron, with two hidden layers having 5 and 2 nodes, respectively. We used Rectifier Linear Unit (ReLU) as the activation function and a constant learning rate of 0.001. We used Stochastic Gradient Descent algorithm and iterated over the dataset 200 times. The function classifier.fit (sklearn function) is used to train the model. This comprised of supervised learning as we trained the neural network on the Labeled Movie Review, Twitter, and Reddit datasets.

The advantage of using a neural network over other classifiers is that there are many more parameters and combinations of weights due the presence of hidden layer. It is also capable of learning models in real-time (on-line learning) using function partial_fit (sklearn function).

4.2 Data Cleaning and Tokenization

We retrieved tweets using Twython API and posts using PRAW. Since, we are focusing on textual data for our experiments, we extracted ‘text’ field from tweets and ‘title’ field from posts and created separate files for tweet-texts and post-titles.

4.2.1 Data Cleaning

The data we collected in files contained foreign elements such as symbols, special characters, links, hashtags, etc. The next task was to remove these foreign elements from the collected data, to produce clean data for tokenization and feature generation.

4.2.1 Tokenization

Once the Twitter and Reddit datasets were cleaned, we tokenized them into individual words rather than sentences. We used Natural Language Toolkit (NLTK 3.2.4) for tokenization. NLTK is a platform used to build Python programs to work with human language data (NLTK 3.2.4 documentation). NLTK provides Twitter-aware tokenizer called TweetTokenizer. We used strip_handles and reduce_len parameters of TweetTokenizer, to remove user handles and reduce length of characters occurring multiple times. When we passed Reddit data through the same tokenizer, the one coded for Twitter dataset, tokenization results were accurate.

Hence, we used same tokenizer for Reddit dataset as well. Tokenizer divides the line into lists of words, converts all words to lower case letters, and stores tokenized data in comma separated files (one for Twitter data and one for Reddit data). Table 4.1 shows sample tweets and post data after tokenization.

Table 4.1 Sample tweets and posts after tokenization

Original Data (tweets and posts)	Tokenized Data
never felt so suicidal in my life :'(i cant take this	[never,felt,so,suicidal,in,my,life,i,cant,take,this]
@gukhwabrdg @dear_pjm I love death	[i,love,death]
RT @intosnight: I feel so attacked and suicidal rn :(https://t.co/PiyNEGFjjo	[i,feel,so,attacked,and,suicidal,rn]
WAY too much lately! [#depressed #ugly #fat #anxiety #depression #hate #tired #selfharm #suicide #suicidal #death... https://t.co/DroQozPgcH	[way,too,much,lately,depressed,ugly,fat,anxiety,depression,hate,tired,selfharm,suicide,suicidal,death]
RT @LR_MerseyCare: Identify quickly whether you are suffering from #postnatal #depression by reading our FREE guide: https://t.co/03vWh3ECA...	[identify,quickly,whether,you,are,suffering,from,postnatal,depression,by,reading,our,free,guide]
Before I die I want to _______?	[before,i,die,i,want,to]
SUICiDe cAUghT ON Tape!!	[suicide,caught,on,tape]

4.3 Creating data features

After tokenizing Twitter and Reddit datasets, next step is to transform data into features using Positive-Negative words. This transformation creates the Input Layer for the Neural Network. For each tweet, post, and review, if the token is present in Positive word list, 1 is appended to the feature data, if the token is present in Negative word list, -1 is appended to the feature data, otherwise 0 is appended. 0 signifies that the word is neutral and this information is preserved in training the neural network and therefore, it is not eliminated. We resize the feature data by zero padding the records having length less than the maximum record length. The output of this process serves as the input to the model, which consists of series of features comprising of 1s, -1s, and 0s.

One of the challenges we faced here was to check for substrings of words. For example, the word, ‘holiday’ may be present in Positive-Negative words dataset but the token in Twitter/Reddit/Movie Review datasets is ‘holidays’. Our system handles such cases by checking the word devoid of the trailing ‘s’ but it doesn’t handle cases where the word is in another form (ly, ed, es, etc). Since Positive-Negative words dataset is fairly extensive, containing varied forms of words (as can be seen in section 3.5 above), these cases don’t go unhandled. Also, we came across misspelled words, for example we found multiple tweets/posts (identified as suicidal risky) with word suicide, (mis)spelled as ‘sucide’ or ’suicid’. We added these misspelled versions of words suicide and suicidal to our negative word list. We also added misspelled version of some other words like depression (depresion, depresed, depresing) , kill (kil, kiling), obsessive-compulsive disorder (ocd), etc which we observed getting used, more often, in suicidal risky content online. But, other words which are misspelled or abbreviated will get replaced by 0 indicating neither positive nor negative sentiment, even though the correctly spelled word is present in Positive-Negative word list.

Fig 4.2 and 4.3 show how a tweet and a post get transformed from its original form to tokenized form to input feature. It can be seen from Fig 4.3 that word suicidal is written as ‘sucidal’ in a Reddit post. Since we have added this word to our negative word list, it is identified as a negative sentiment and got replaced by -1 (instead of 0) in input feature vector.

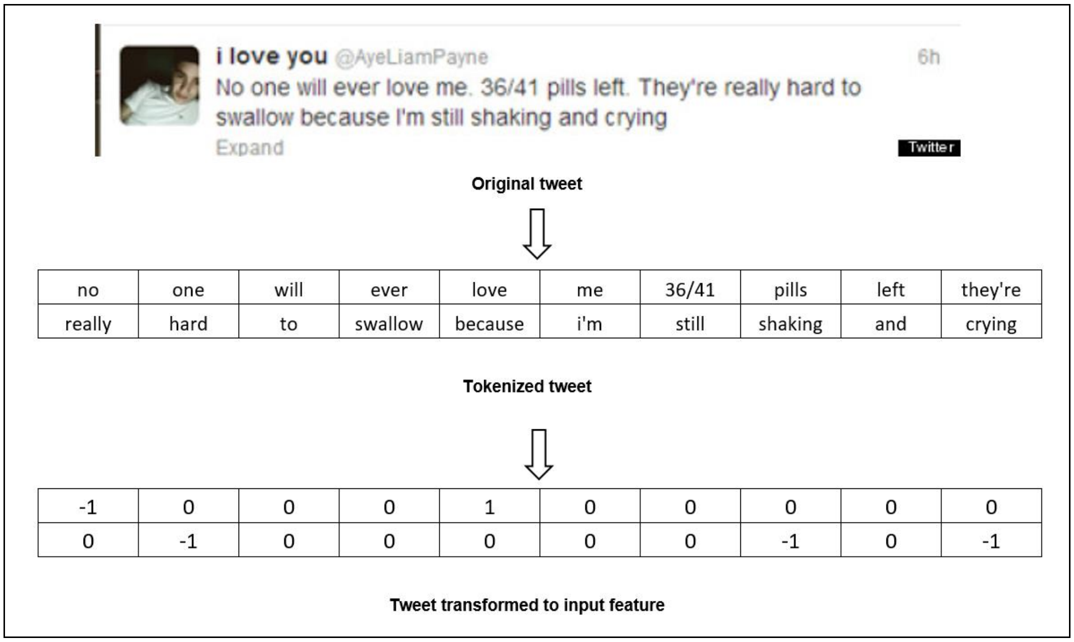


Fig 4.2 Sample Twitter data transformation

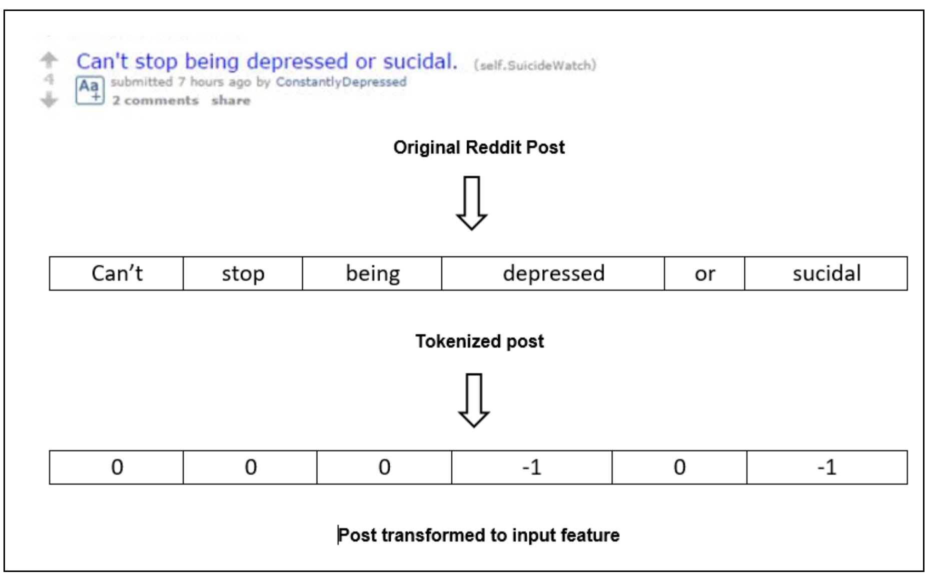


Fig 4.3 Sample Reddit data transformation

4.4 Data processing through Trained Neural Network

The input data features are then passed through a trained Neural Network. The Neural Network predicts whether the data content is positive or negative. If the data is predicted as positive content it is marked as (suicidal)Non-Risky content, but if data is predicted as negative content we check for feature sum, if sum of features is less than 0, system marks the data sample as Suicidal Risky content otherwise (suicidal)Non-Risky content. We added this extra condition for the data predicted negative because we observed that there are lot of sentences which give some information regarding suicide or related topic but is not necessarily a suicidal ideation. For example, 'My brother committed suicide this morning, hug your family', this post gives some information about suicide, but it is not considered as suicidal ideation. When we pass this data through our system, the neural network predicts it as a negative sentiment, but the extra layer of condition puts this data in (suicidal)Non-Risky class. We then save this Suicidal Risky data and Non-risky data in separate comma separated files for further analysis.

Chapter 5

EXPERIMENTS AND RESULTS

In previous chapter we described detailed implementation of our system to classify social media data as suicidal risky content and non-risky content. This chapter will discuss the experiments conducted to evaluate the system and the evaluation results.

5.1 Training the Neural Network

As mentioned in section 4.1 above, we are using Multi-Layer Perceptron Classifier model with two hidden layers having 5 and 2 nodes respectively. We trained this model using Movie Review Dataset, Twitter data set and Reddit dataset. The Movie Review dataset is labeled as Positive review and Negative review, but getting tweets and posts labeled was a challenging task. To the best of our knowledge, neither tweets nor posts datasets exist which are labeled for suicidal(negative) and non-suicidal(positive) content. Therefore, we need to create our own labeled Twitter and Reddit datasets to train and evaluate our system.

5.1.1 Labeling Twitter and Reddit data

To identify whether a particular social media post indicates suicidal intent or not, is challenging and requires domain knowledge. We presented a list of tweets and posts, extracted by our system, to psychologists and psychiatrists, who work for suicide prevention, and requested them to classify the data as suicidal content and non-suicidal content. Due to the time required for this manual classification, the number of tweets and posts classified by them was limited (60 tweets and 40 posts). When we compared their inputs, we found that they agree with each other for 89% of the time. Since we needed more data to train and test the system, we decided to request some more individuals to classify the data, this time individuals with a little or no domain knowledge (based on general understanding of the context). We found that input from these individuals agree with experts' input for 86-90% of the time. Hence, we continued this manual classification and we got 1000 tweets and 1000 posts labeled as suicidal content and non-suicidal content. Also, we found some tweets and posts by social media users who attempted suicide after posting these messages from news reports and articles. We added these tweets and posts to our dataset as suicidal tweets and posts. Table 5.1 shows number of tweets and posts identified per label.

Table 5.1 Labeled Twitter and Reddit data statistics

Data	Positive(Non-suicidal)	Negative (Suicidal)
Tweets	615	385
Posts	640	360

5.1.2 Creating training datasets

To evaluate learning rate of our model we created two sets of training data, Dataset A and Dataset B. Table 5.2 gives the number of reviews, tweets and posts used in each of Dataset A and Dataset B. Dataset A is created by dividing the labeled datasets into training and test datasets as per 70:30 percentage proportion and Dataset B is created by dividing the labeled datasets into training and test datasets as per 80:20 percentage proportion.

Table 5.2 Training datasets

Data	Number of data samples in **Dataset A**	Number of data samples in **Dataset B**
Positive Reviews	3000	4800
Negative Reviews	3000	4800
Positive tweets	430	492
Negative tweets	270	308
Positive posts	448	512
Negative posts	252	288

5.1.3 Comparing training and cross validation scores for training datasets

We trained our model first with data from Set A and calculated training score and cross validation score using 10-fold cross validation. Fig 5.1 shows learning curve of our model for Dataset A. We decided to check the validation score by training the model with more samples. Therefore, we trained our model with Dataset B. Fig 5.2 shows learning curve of our model for Dataset B. We observed that our model learned better with Dataset B, so, we used Dataset B to train our model.

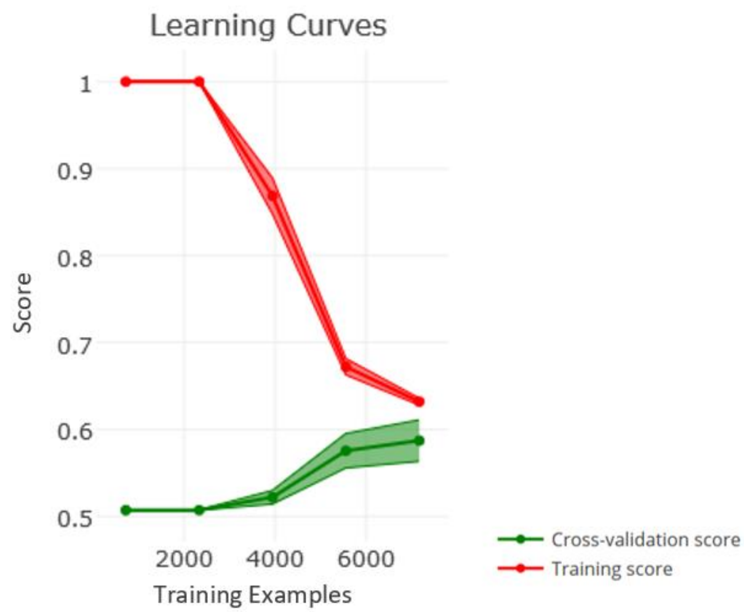


Fig 5.1 Learning curve of the model for Dataset A

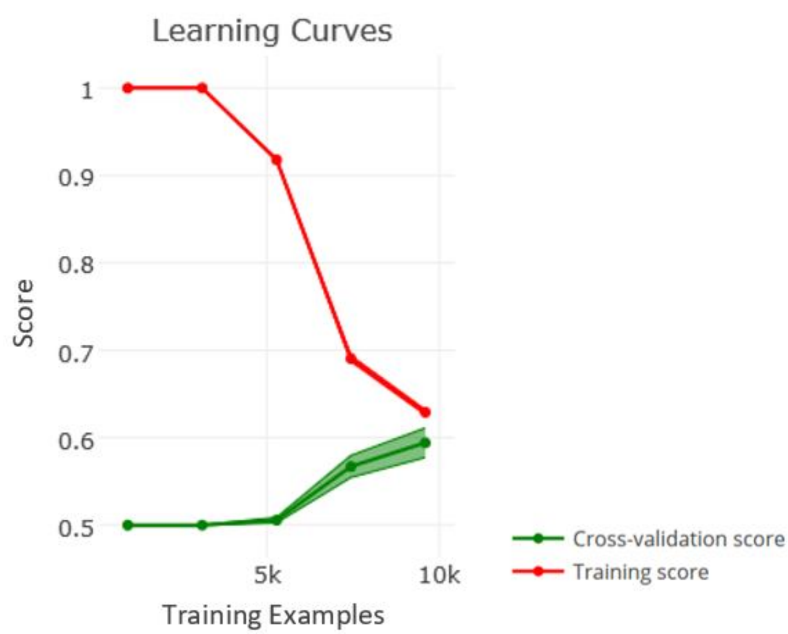


Fig 5.2 Learning curve of the model for Dataset B

5.2 Prediction Results for test datasets

We created test datasets from our labeled Twitter and Reddit datasets. Table 5.3 shows test dataset statistics. We passed this data through our system and collected prediction results for Twitter and Reddit data in separate files. The module classifier.predict (scikit-learn documentation) is used to predict the labels for test dataset.

Table 5.3 Test Dataset

Data	Number of data samples in test dataset
Positive(Non-suicidal) tweets	123
Negative (Suicidal) tweets	77
Positive (Non-suicidal) posts	128
Negative (Suicidal) posts	72

5.2.1 Performance Evaluation

Confusion matrix describe the performance of our classification model on a set of test data for which true values are known. Fig 5.3 and 5.4 show confusion matrix for Twitter and Reddit datasets respectively. We evaluated performance of our model by calculating accuracy, precision, recall and F1 score.

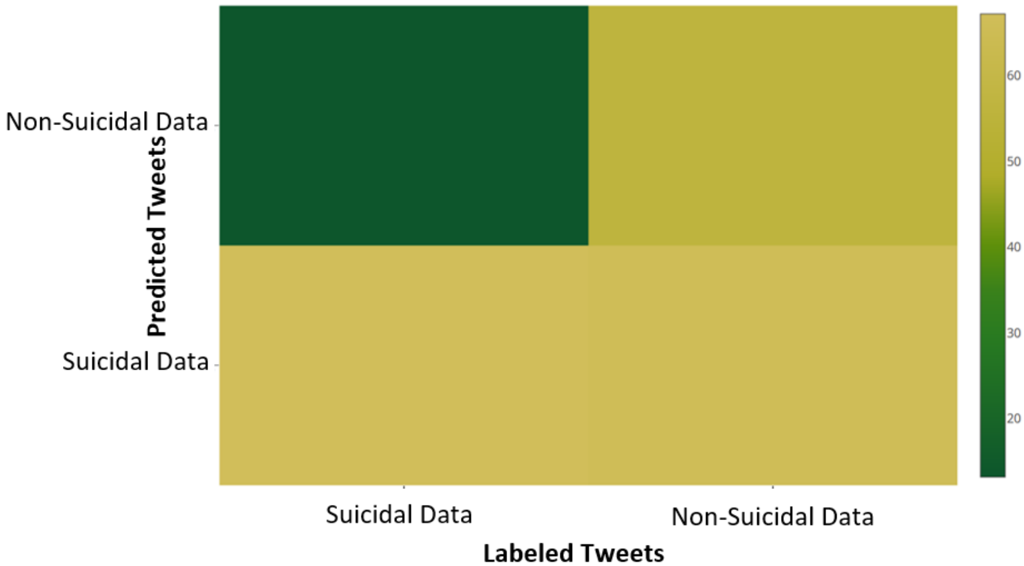


Fig 5.3 Confusion Matrix for Twitter test dataset

Table 5.4 Confusion Matrix in numeric form

	Predicted Suicidal	Predicted Non-suicidal
Suicidal Data	64	13
Non-suicidal data	63	60

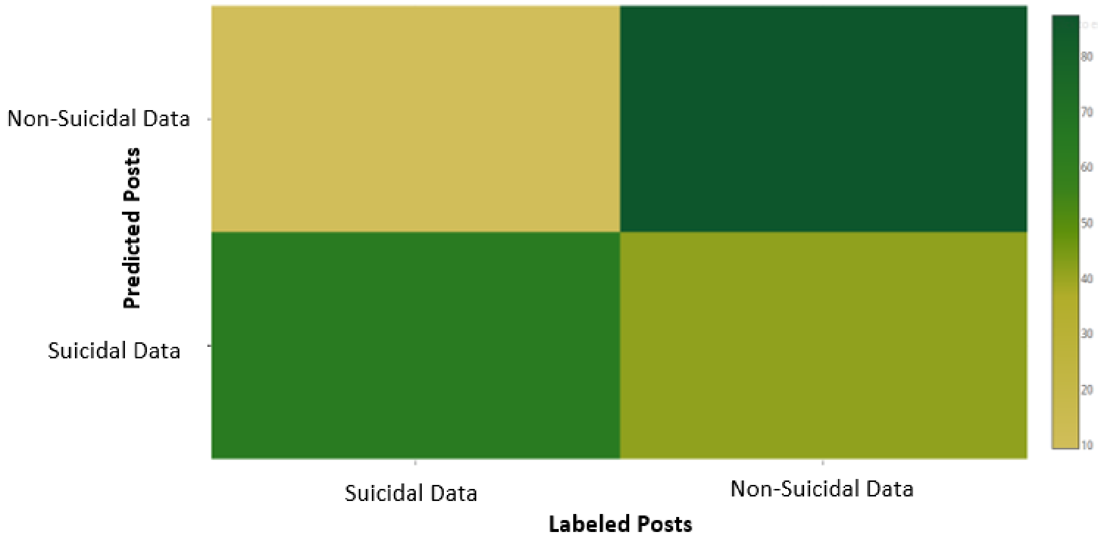
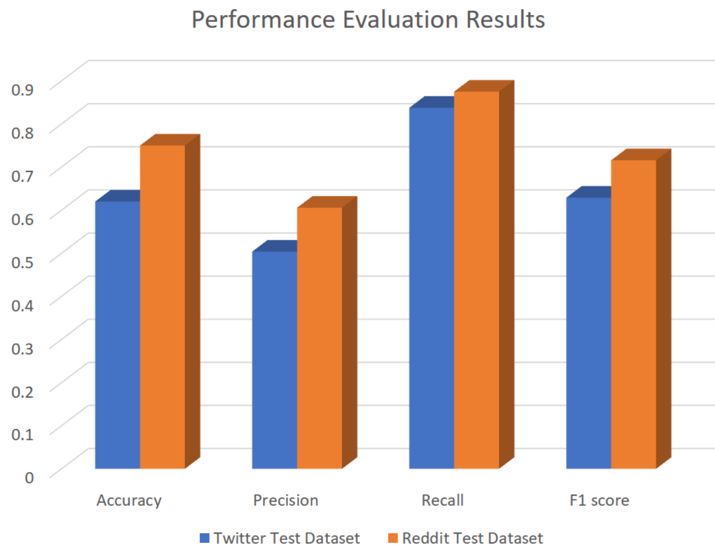


Fig 5.4 Confusion Matrix for Reddit test dataset

Table 5.5 Confusion Matrix in numeric form

	Predicted Suicidal	Predicted Non-suicidal
Suicidal Data	63	9
Non-suicidal data	41	87

Accuracy is the most intuitive performance measure and it is simply a ratio of number of correctly predicted data samples to total number of data samples. We got 0.62 accuracy for Twitter dataset and 0.75 accuracy for Reddit dataset, which means our model is approximately 68% accurate. Since our datasets are not symmetric, we must consider other parameters as well to evaluate performance of our model. Fig 5.5 gives values of accuracy, precision, recall and F1 score calculated for Twitter and Reddit test datasets.



Performance Measure	Twitter Test Dataset	Reddit Test Dataset
Accuracy	0.62	0.75
Precision	0.5039	0.6057
Recall	0.8311	0.875
F1 score	0.6274	0.7158

Fig 5.5 Performance Evaluation Results

5.3 Analysis of Results

It can be seen from above fig 5.5 that precision of our system for Twitter dataset is 0.5039 which is less. Precision answers the question that of all tweets that are labeled as suicidal by our system, how many are actually suicidal. As we can see from fig 5.3, confusion matrix for Twitter test dataset, number of non-suicidal tweets getting labeled as suicidal is high. We analyzed output data to gain insights for this issue. It is seen from the data that there are tweets which give some information or facts about suicide or related topics such as depression, mental health issues, death related issues, etc. for example,

‘RT Self harm or suicide are never the solution whenever you feel lonely or in pain please speak up and share with your friends’, ‘RT sorry but I have to say something lying about suicide is NEVER okay, people were severely affected by the thought of losing loved ones’,

‘Depression is weird soon as you get one little inkling of success you just want to sabotage urself’,

‘Depression is a mental health issue while you're sleeping you're not thinking, you're dreaming’,

these tweets get classified as Suicidal Risky content by our system. To get better results, we need to train our model with more data samples.

Recall is the ratio of correctly predicted positive class labels to all data samples in actual positive class. It answers the question, of all tweets/posts that are suicidal, how many did our system label as suicidal. Recall values of our system for both the test datasets are above 0.8, which are impressive. This means that our system can catch approximately 86% of suicidal content online. We further analyzed suicidal tweets/posts labeled as non-suicidal by our system. We found tweets/posts containing suicidal content written in sarcastic way or using words having positive connotation, for example,

‘am i suicidal or just starving hmm’,

‘RT im back better more suicidal than ever’,

‘Am I wrong for thinking that suicide should be a fundamental human right’

We need to look for more such data samples in order to train the network.

F1 score is the weighted average of precision and recall. For our system, where datasets are not symmetrical, F1 score is more useful in evaluating the performance of the system. F1 score of our system for Twitter Dataset is 0.6274 and for Reddit test dataset is 0.7158, which are pretty good.

We also analyzed output data against the data we extracted from Twitter and Reddit. As our extraction process extracts only those tweets/posts which contains word/s from our suicide vocabulary (suicide terms), as mentioned in section 3.1, we evaluated labeled data samples for some high frequency suicide terms from suicide vocabulary. For example, out of total number of tweets extracted due to presence of word suicide are labeled as suicidal tweets by our system. It is found that there are total 16 tweets in test dataset containing word suicide and 6 out of these 16 tweets got classified as suicidal.

Fig 5.6 shows percentage of labeled suicidal and non-suicidal tweets for Twitter test dataset by subcategory of suicide vocabulary

Fig 5.7 shows percentage of labeled suicidal and non-suicidal posts for Reddit test dataset by subcategory of suicide vocabulary

This analysis helps in understanding vocabulary use, choice of words, frequently occurring words and their relevance in suicidal ideation online.

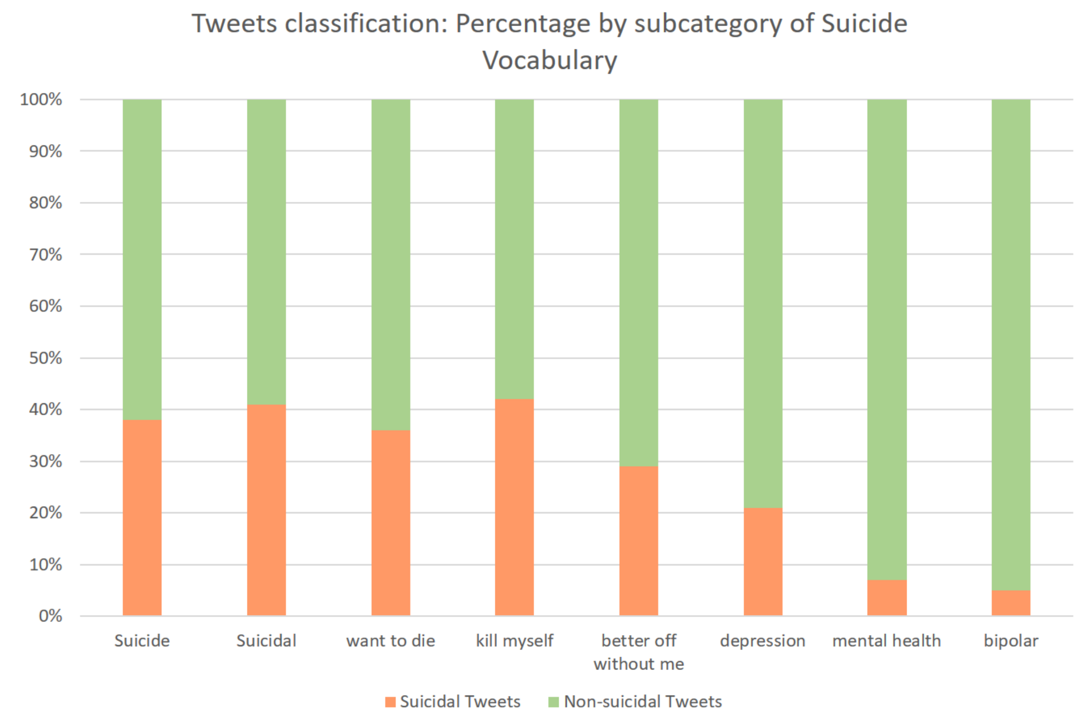


Fig 5.6 Tweets classification: Percentage by subcategory of Suicide Vocabulary

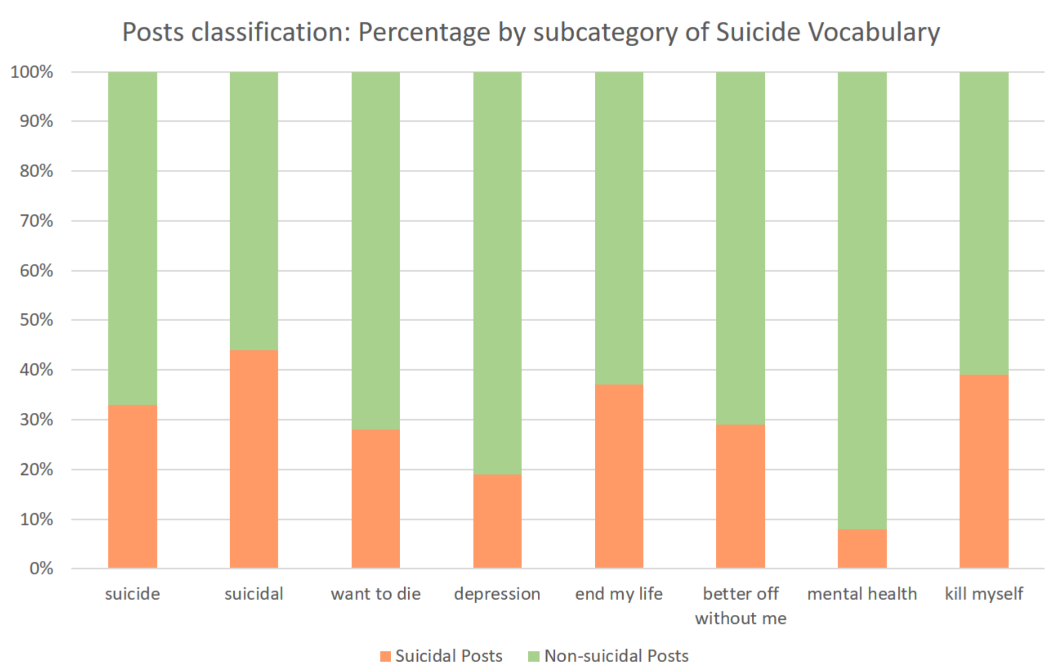


Fig 5.7 Posts classification: Percentage by subcategory of Suicide Vocabulary

Chapter 6

CONCLUSION AND FUTURE WORK

In previous chapters we described about our classification model, experiments, and analysis of classification results. In this chapter, we will discuss about limitations, conclusion, and future work.

6.1 Limitations

For our research, we were interested in gathering data from various social media sites. We found that Facebook, Twitter, Tumblr, Reddit are majorly used social media sites to broadcast textual messages. We tried extracting data from all the four social media sites, but due privacy policies we could not extract general data from Facebook and Tumblr. We could extract data which was limited to our network only. Therefore, we had to limit our scope for data retrieval to Twitter and Reddit.

There are limitations on number of tweets and posts that can be extracted using APIs (Twython and PRAW, respectively) at a time and in a day. Our system can extract around 600 tweets and 850 posts at a time. We try to get most of the relevant tweets and posts by searching data based on keywords used frequently and commonly in suicidal ideation and we update our suicide vocabulary accordingly.

For our research, we focused on textual messages broadcasted on Twitter and Reddit. We extracted data based on suicide vocabulary and we used Positive-

Negative word dataset to create data features as explained in chapters 3 and 4. Both these datasets contain list of individual words. It is found that social media users use a lot of abbreviations for original words while writing textual messages online. For example, they use ‘m’ for ‘I am’, ‘sry’ for sorry, ‘lol’ for laugh out loud, etc. Also, they use words with their shortened spellings, for example, ‘depresion’ for ‘depression’, ‘kil’ for ‘killing’, ‘u’ for ‘you’, ‘gues’ for ‘guess’, etc. We observed one more common pattern of writing words with spaces or dashes or dots between letters to convey the intensity of feelings behind those words, for example, ‘goood-for-noothing’, ‘E v e r y o n e’, ‘sui….ci…de…sui..ci..d..e’, ‘kil…kil…’, ‘---die---‘, etc. We tried handling most of these cases, but there are a lot of challenges involved in understanding these patterns and extracting meaningful information from them.

We can extract age and gender information neither from Twitter nor from Reddit as this information is optional for Twitter users and they can hide it from getting extracted using privacy settings and Reddit also has policies on similar lines. This limits generalizability of the results.

6.2 Conclusion

As rightly said by Joseph Franklin, a psychologist at Florida State University who studies suicide risk, ‘There is just a tiny predictive signal’ (indicating future suicide attempt), in this thesis we tried to catch that signal using social media data. We researched for developing a system to automatically detect social media data for suicidal ideation. Our system exhibited comparable performance when tested for datasets from two different social media sites Twitter and Reddit. Our results show that it is possible to detect suicidal ideation from online data using trained neural network and there can be a single system in place which can process data from multiple sources to predict suicidal content with analogous prediction results. Having said this, we know that there is a lot more to understand and predict in a textual message broadcasted online. People can use the words 'suicide', 'kill myself' or 'want to die' for multiple reasons and from various perspectives. Also, differentiating users in immediate danger of committing suicide from users with passive suicidal ideation is something that our system is unable to do. Overall, the results of this thesis are encouraging and suggest that future work in the right direction will yield promising results in identifying suicidal intents from social media data and potentially responding to vulnerable users in real-time.

6.3 Future Work

Extracting meaningful information from social media data is a challenging job. Language used in social media conversation plays a critical role in this process. It often changes with inventions of numerous spin-off versions of words. For example, when Instagram blocked a hashtag #selfharm, people came up with new hashtags #selfharmm or #selfinjuryy or slang hashtags indicating self-harm. This continuous evolution of online language needs to be incorporated in training datasets so that classifiers will also learn this language and these data samples will not go unhandled.

We also observed that, not just text but people indicate their suicidal intent using emoticons and images. Our system currently focuses on textual part of online posts, but it is necessary to upgrade our system to extract information from emoticons and images as well.

We can extract information like time and geographical locations from tweets and posts to generate location based results. Also, it is found that suicide rate increases during holiday seasons like Christmas, Thanks Giving, New Year, et al. The time of tweet/post will be helpful in generating these results.

As we are progressing in devising various methods to automatically detect suicidal intent from social media data, social media users are coming up with new ways to express their suicidal thoughts. Recently, a few social media users live streamed their suicide attempt! Thankfully viewers reported these incidents and lives of those vulnerable users were saved. But now there a need to automatically detect such live streaming and respond to the situation in real time, within few minutes.

REFERENCES

[1] World Health Organization http://www.who.int/mediacentre/factsheets/fs398/en/

[2] American Foundation for Suicide Prevention https://afsp.org/about-suicide/suicide-statistics/

[3] O'dea, Wan, Batterham, Calear, Paris, Christensen 2015. Detecting Suicidility on Twitter.

[4] Greenwood, Perrin, Duggan. Social Media Update 2016. http://www.pewinternet.org/2016/11/11/social-media-update-2016/

[5] Speaking of Psychology: Preventing Suicide, Episode 14. https://apa.org/research/action/speaking-of-psychology/preventing-suicide.aspx

[6] Dr. Robert E. Accordino. http://www.dodlive.mil/2017/01/19/social-media-offers-tools-to-improve-mental-health-reduce-suicide/

[7] Suicide Vocabulary Word List. https://myvocabulary.com/word-list/suicide-vocabulary/

[8] Why scientists think your social media posts can help prevent suicide. https://mashable.com/2016/06/26/suicide-prevention-social-media/

[9] Hootsuite Insights. https://hootsuite.com/products/insights/

[10] Semantria. https://www.lexalytics.com/semantria/excel

[11] Rapidminer. https://rapidminer.com/solutions/text-mining/

[12] Godbole, Srinivasaiah, Skiena 2007. Large-Scale Sentiment Analysis for News and Blogs.

[13] Kouloumpis, Wilson, Moore 2011. Twitter Sentiment Analysis: The Good the Bad and the OMG!.

[14] Abboute, Boudjeriou, Entringer, Aze, Bringay, Poncelet 2014. Mining Twitter for Suicide Prevention.

[15] Colombo, Burnapa, Hodoroga, Scourfield 2015. Analysing the connectivity and communication of suicidal users on twitter.

[16] Liu, Cheng, Homan, Silenzio 2017. Learning from various labeling strategies for suicide-related messages on social media: An experimental study.

[17] Birjali, Beni-Hssan, Erritali 2016. Prediction of Suicidal Ideation in Twitter Data using Machine Learning algorithms.

[18] Jashinsky, Burton, Hanson, West, Giraud-Carrier, Barnes, Argyle 2013. Tracking suicide risk factors through Twitter in the US.

[19] Pang and Lee, Cornell University.
http://www.cs.cornell.edu/people/pabo/movie-review-data/

[20] Twython 3.4.0 documentation. https://twython.readthedocs.io/en/latest/

[21] PRAW documentation. https://praw.readthedocs.io/en/latest/

[22] Hu and Liu 2004. Mining and Summarizing Customer Reviews.

[23] Jupyter Notebook documentation. http://jupyter.org/

[24] NLTK 3.2.4 documentation. http://www.nltk.org/

[25] scikit-learn 0.18.2 documentation. http://scikit-learn.org