Student name	Dac Phuc Ho
Student number	795603
Supervisor	Prof. Karin Verspoor
Total number of credit points	75
Type of project	Research Project
Subject code	COMP60002
Project title	Identifying individuals with high risk of developing suicide ideation on online forum

Identifying individuals with high risk of developing suicide ideation on online forum

Dac-Phuc Ho

Submitted in partial fulfilment of the requirements for the degree of

Master of Science

School of Computing and Information Systems
THE UNIVERSITY OF MELBOURNE

October 2017

Abstract

Over the past decade, the rise of social networking sites such as Twitter, Facebook, Reddit has changed the way people express their opinons. At the same time, the advancement of machine learning and natural language processing provides powerful tools to analyze massive amount of data users generated over time. Those tools are extremely beneficial to researchers in mental health field. The analysis of users' public posts help mental health professionals to arrange appropriate support for users with mental health issues, especially users who are vulnerable to suicidal thoughts.

Efforts have been made to utilize machine learning to predict suicidal individuals on social media platforms. However, classifiers built in those studies usually performed on a small randomly sampled subset of a big dataset obtained. Thus, the applicability of said classifiers in practice remains unknown. In this study, we try to build classifiers that produce state-of-the-art performance using a test set with small sample size. Then, we expand the test set to see whether the classifiers can maintain the performance on a more realistic scenario or not. We also report observations on language differences between labelled at-risk individuals and 'control' users.

Declaration

I certify that

1. this thesis does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any university; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person where due reference is not made in the text.

2. where necessary I have received clearance for this research from the University's Ethics Committee and have submitted all required data to the Department.

3. the thesis is approximately 18,000 words in length, exclusive of tables, maps, bibliographies and appendices.

Signed: Phuc HoDac

Date: 28/10/2017

Dac-Phuc Ho, October 2017

Acknowledgements

First and foremost, I would like to express my sincere appreciation to my supervisor Prof. Karin Verspoor for the guidance despite of her heavy workload plus other academic duties. She provided valuable suggestions and insightful feedback for my thesis. I am truly thankful to have such a friendly, humorous and knowledgeable Professor as my supervisor.

I am grateful to Loi Nguyen, Thao Phuong Nguyen. I would have been in a much more miserable state if not for their support during my time in Australia. I am also lucky to have a chance to collaborate with and learn so much from my postgraduate classmates Tri Chandra, Steven Spratley, Victor Dilks and Harvey Nguyen.

I would like to thank Nhat-Duc Nguyen, Hung Viet Cao, Son-Tung Hoang, Ha Bach, Minh Duc Tran and other undergraduate classmates. We had so much fun in college and they encouraged me to take on further study.

My affectionate thank goes to my relatives Thuy-Trang Nguyen, Van-Trang Ho, Lan-Huong Hoang, Duc Nguyen, Bich Nguyen, Nam-Huan Hoang, Van-Anh Nguyen for the emotional and esteem support.

I am greatly indebted to Duy-Khiem Nguyen, Tuan Pham, Binh Pham Van, mentors, staff and other colleagues at NashTech Vietnam who have taught me not only domain knowledge but also professional attitude in the first days of my career.

I would like to extend my thank to friends who have shared various parts of my life: Anh-Tu Tran, Quang-Duy Le, Son-Vu Le, Ngoc-Hiep Tran, Tri-Khang Le, Thanh-Dat Nguyen, Dung Tran and many many others to whom I am thankful. I feel guilty that I cannot list all their names here. They have made my life much more enjoyable than I could ever hope for.

Finally, I wish to express my heartfelt gratitude, love and deepest appreciation to my family. My parents, Boi-Huong Hoang and Hien Ho Dac, always support my choice even though it is a difficult path. My elder sister Quynh-Nga Ho Dac always warm my heart with her cheerful attitude and I appreciate the consideration from my brother-in-law Manh-Ha Nguyen. They have given me unconditional love and it became my motivation. Without them, I would have never made this far.

Contents

List of Figures

List of Tables

Chapter 1
Introduction

1.1 Internet and social networking sites

During the past decade, the number of Internet users has increased exponentially to nearly 3.2 billion users. A large portion of them use social media websites for a wide range of activities: from sharing personal posts, discussing a particular topic to advertising businesses. These applications are enabled by numerous advantages of social networking websites compared to traditional forms of communication namely convenient connectivity, easy to access and adaptability to its users. Among the leading social network sites, some can attract from hundreds of million up to billions of users such as Facebook[1] (2 billion monthly users [50]), Twitter[2] (328 million active users [3]) or Reddit[3] (234 million unique users [1]). Each of these websites has different main features and functions, but overall allows users to post their content publicly.

The rapid development of social media sites has changed the habits of perceiving information and express opinions of young generation. They are now more open to speak out their mind via computer-mediated communication channel. The shift to digital age has enabled researchers in mental health area to access data much more easily. It provides powerful tools to collect structured and unstructured data, opening directions in online public health such as mental health surveillance: predicting the number of suicide cases [51], mental health-related posting behaviors [5] or stress level of individuals monitoring [32].

[1] https://www.facebook.com/
[2] https://twitter.com/
[3] https://www.reddit.com/

1.2 Youth mental health and online therapies

Mental health disorders and poor mental state have detrimental effects on one's life such as decreasing enjoyment and quality of life, lowering productivity, making one vulnerable to abuse and ultimately susceptible to suicide. The causes of mental illness come from psychological, biological, and environmental factors of each individual. Some factors cannot be controlled like genetic, individuals have higher risk of developing mental illness if their family history show symptoms of it, such factor is called risk factor. Heredity of mental illness has been observed in previous studies [27]. On the other hand, factors from environment - for example, education and affection in family are protective factors - can be modified by people. Risk factors are factors occur before symptoms of mental health problems and contribute to the risk of developing mental illness [30]. In contrast, protective factors mitigate the effects of risk factor, preventing mental illness progressing [46].

For the past two decades, intervention programs that aim to promote youth mental health have been improved in many aspects. Before the Internet plays a vital role in information dissemination, these programs were implemented mostly in school settings considering that is the most effective ways to reach adolescents at school. However, the school-based approach requires a lot of human resources and intensive engagement with students. As a consequence, it is hard to sustain the program upon completion and there is high chance that funding withdrawal decision being issued when project was running halfway. Spence and Shortt [48] have raised their concerns regarding effectiveness and efficacy of school-based programs. It might not be justified for the objective being preventing student from developing depression in long term while the programs were brief and did not incorporate environmental change to evaluate resilience of students. Other than school-based programs, face-to-face intervention also take place at medical centers and clinics with mental health professionals directly give diagnosis and instruction to service users. Naturally, those means of intervention have similar limitations in sustain-

ability and implementation namely time-consuming, adequate training for instructors, cost for setting up and maintaining physical location.

The Internet evolution offers a new way of delivery for intervention programs. Now the programs can be delivered to more students without incurring fixed cost for each additional participant through communication channels provided by web servers. The number of Internet users under the age of 25 is increasing dramatically. The advantages of the Internet in self-learning, as a direct result of Information Retrieval, allow adolescents to access information more effectively without restrictions to space and time. Moreover, the cost for server maintenance is minimal when compared to costs of school-based setting programs, making it easier to sustain the program until its completion. Although the research approach using Internet-based therapies is pretty young, there is evidence indicating efficacy of Internet-based treatment programs is equal or better to programs with school setting.

There are several web-based intervention programs that was deployed to enhance awareness of adolescents about mental health namely beyondblue[4], MyHealthMagazine[5], WalkAlong[6], moodgym[7]. These websites' main feature is providing content and information related to mental health. They made the effort to present the information in an interactive and educational way because they expect visitors to be adolescents. Due to their short time of deployment, the effectiveness of most of those programs have not been assessed by research methodology, therefore we cannot draw conclusion about the exact weak points of Internet-based programs compared to school-based projects. However, if Internet intervention programs are proved to be as effective as one of traditional programs in term of preventing depression and psychological disorders, it is worth shifting programs from face-to-face delivery to Internet delivery because of mentioned advantages and characteristics of Internet-based programs [25].

[4]https://www.beyondblue.org.au/
[5]http://www.yoomagazine.net/
[6]https://www.walkalong.ca/
[7]https://moodgym.com.au/

1.3 Studies in mental health on the Internet

There are a variety of approaches and scopes of studies since the emergence of both large-scale data collection and data analysis and visualization method. At population level, observations using trend-based approach facilitated by big data platform or social network sites - for instance, Google Trend[8] allows user to visualize statistics of keyword(s) sorted by geographical area or time period [20] - are proved be useful to develop prediction models for general population.

At individual level, types of study diversify even more. Many papers utilize Natural Language Processing (NLP) techniques to analyze the language pattern of a large number of users and identify people are vulnerable to mental health illness even before the onset of symptoms [2, 12]. Another direction that aims to support individuals is intervention with involvement of technology. Numerous mobile and smartphone applications with features like personal reminder, providing relevant information for users have been developed to study the efficacy and feasibility of mobile intervention [43].

The increase in data storing capacity and data transfer speed also enhance the real-time interaction between people. This leaf in technology encourages computer-mediated communication and make online therapies possible. Web-based apps are deployed and recognized for their cost-effectiveness in delivering therapies within reasonable time frame via computers. Web-based therapies have significant advantages over traditional face-to-face therapies like cognitive behavioral therapy (CBT). In the past, people must meet professionals in person and it poses barriers such as transportation, scheduling for appointments and long waiting time lead to high cost and inconvenience. Whereas web-based apps do not encounter any of mentioned issues, they offer services to much more larger percentage of population. Several organizations and technology start-ups companies build their service to connect people in need of mental health counseling to trained volunteers or mental health professionals (e.g. 7cups[9], CrisisTextLine[10], TalkLife[11]). These sites utilize the synchronous text-based online interventions - mainly online chat and text

[8] https://trends.google.com/trends/
[9] https://www.7cups.com/
[10] https://www.crisistextline.org/
[11] https://talklife.co/

message - and generally obtain positive results in post-treatment reviews from users [22].

1.4 Potential sources of data

The limitations of traditional clinical treatment have caused reluctance of young people to seek help from mental health professionals in face-to-face setting. The main concerns are financial cost, accessibility of services, lack of information and knowledge and social stigma associated with mental illness [44]. Along with advancement of Internet, young people tend to adopt a habit of being online regularly. They find information about their mental health problems on the Internet and this has become a growing tendency among people experiencing depression and other psychological disorders. Furthermore, youngsters have particular interest for online help due to availability, high levels of anonymity, relatively low cost and less stigmatizing than conventional intervention. These advantages help young people to overcome barriers specified earlier when seeking suitable help for them.

We have reviewed some of Internet-based intervention programs, they fall into one of the three categories:

(1) Self-help: people seek knowledge regarding their symptoms.

(2) Professional treatment: patients connect to trained clinicians via websites designed for mental health topics.

(3) Peer-to-peer support: people connect to non-professionals for advices, personal stories and emotional support.

Among those three forms of intervention, peer-to-peer support is a prominent aspect both to Internet users and researchers. Previous study showed that millions of people visit support groups on daily basis. Peer-to-peer support gathers people with similar stressors or health problems with the aim of mutual support or assistance of experienced users to other users in need of help. Peer support takes place in pairs or groups via the Internet. There are many online peer support platforms, they are either asynchronous (e.g. discussion groups, forums) or synchronous (e.g. chat room).

Both paradigms are popular and widely used for intervention, however, they are different in nature. Chat-based intervention usually involves mental health client and a person or several people who act as counselors even though they are not necessarily mental health professionals. Conversations take place in real time via websites or mobile phone apps, helping listeners to engage with users and their stories in great details. Most of such services are very cautious about data of conversations due to confidentiality of users. During the chat, users often reveal personally identifiable information such as name, location or employment history. Each country has its own privacy law or act where leaking private information can lead to legal actions. Because of this sensitive matter, chat-based services always consider carefully whether to share data with researchers or not. Each application to data sharing request undergo strict assessment of staff in charge of research from the company behind the website, who usually have research background. As a result, access to data of chat-based services is limited to most researchers.

On the other hand, forums or discussion boards allow user to post content publicly, inviting other to comment the post with relevant advices and support. Since the posts are publicized on forums, any registered, subscribed users to that forums or even regular visitors can read the posts. While users of chat-based services feel more comfortable to share their stories along with many details, forums' users reach much wider variety of readers despite briefer version of the story. With growing number of people seeking help on online forums, researchers are able to collect data more easily and less worry about privacy issue than data from synchronous intervention.

1.5 Computer Science and suicide research

Suicide is one of the leading causes of death in many countries. In 2012, there are approximately 804,000 people committed suicide worldwide [52]. The number of suicide victims in United States is 41,149 people, costed the health care industry approximately $51 billion annually[9]. It becomes more of a concern as the number of suicidal adoles-

cents and young adults aged 15-34 increases every year. Given these statistics and the fact that the total number of suicide cases are increasing annually, it is no surprise that suicide prevention becomes top priority for health care industry.

Suicide prevention efforts mainly focus on early risk recognition and direct patients to appropriate support. A study has reported strong correlation between suicide and mental health illness; some of indicators included previous attempts of suicide and medical record of psychiatry [21]. Young people not only hold a large percentage of demographics on suicide but also generate massive amount online content in form of posts, videos and pictures on social networking sites. The sites can be used to discuss suicide-related methods, copycat suicide or express suicidal thoughts. It is infeasible to manually monitor all the posts to highlight posts with suicidal messages. Subsequently, there emerge the need to automatically spot such posts and inform stakeholders to decide pertinent intervention. Researchers hope to apply NLP and machine learning (ML) techniques, more specifically sentiment mining to pinpoint posts of suicidal people.

1.6 Thesis outline

This thesis is divided into five chapters.

Chapter 2 presents a detailed literature review on the topics of applications of Computer Science in suicide research. The chapter introduces types and process of suicide followed by the application of NLP and ML in emotion detection in suicide notes. The review compares several approaches and results of emotion classification task. Moreover, current trends of social media analysis related to suicide are summarized. They include predicting national suicide numbers based on dysphoria blog posts, a study to verify contagion effect of celebrity suicide and attempts to identifying people with high risk of developing suicidal thoughts. The research opportunities for this thesis are revealed based on these reviews.

Chapter 3 describes tool and resources used for the experiment. This chapter give brief introduction to a popular NLP framework, a ML package and employed NLP methods.

Chapter 4 gives overview of the source of data and research questions. Details of the

experiment conducted will be covered. Those steps are data collection, construction of user classes, feature selection, two schemes of the experiment, results and discussion of our finding.

Chapter 5 summarizes the details of this thesis and give some thoughts about future outlook of this field of study.

Chapter 2
Literature review

2.1 Types and process of suicide

Types of suicide

Suicide is defined as act of a person has self-awareness and intentionally engages in effort to end his or her life. In other word, suicidal behaviour is any active or passive act initiated by a person with expectation to cause self-inflicted death. Emile Durkheim, a French sociologist specified four different types of suicide [18], they are illustrated in Figure 2.1.

(1) **Egoistic suicide:** It is a type of suicide occurs when an individual has low level of social integration. It means one does not feel he or she belong to or being accepted in a community. If this happens for a long time, it can lead to feeling of emptiness, melancholy and chronic depression. Those people feel isolated and receive little social support when they undergo hardship in life, leading to higher chance of committing suicide. It is called egoistic suicide because it springs from excessive individualism. To take example of this, Durkheim reported that the suicide rate among unmarried men is higher than that of married man since the unmarried are less bound and not connected to social norms.

(2) **Altruistic suicide:** In contrast to egoistic suicide, altruistic suicide occurs in societies with high integration. When an individual is overwhelmed by a community's beliefs, he put his personal values lower than community's values as a whole and considers mutual interests of community is more important than his needs. Examples for this type of suicide are suicide bomber and Samurai from Japan. Suicide bombers believe their death contribute to progress of his group to a common goal. In Samurai's code of feudal Japan,

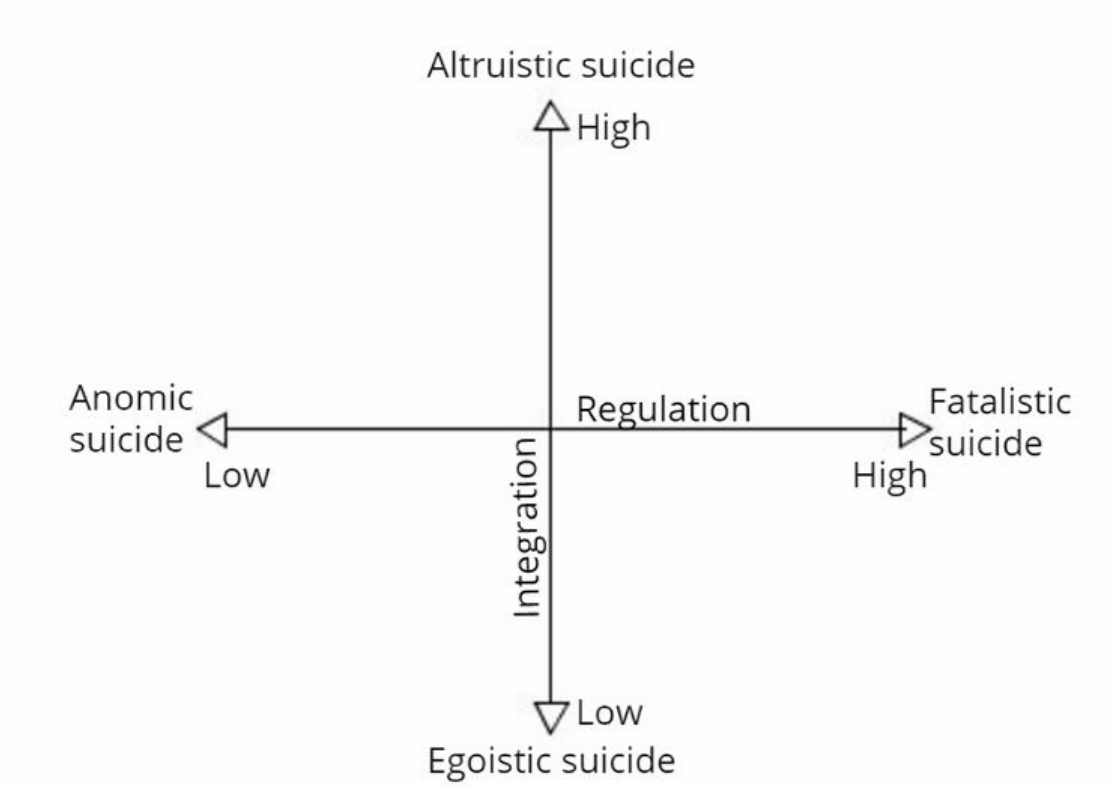

Figure 2.1: Types of suicide according to Émile Durkheim

a Samurai who fail to complete an assigned task or lose in battle is seen as disgrace thus a samurai commits suicide when he feels humiliation and dishonored.

(3) **Anomic suicide:** Anomic suicide takes place in society with low degree of regulation. In a situation of sudden economic and social turmoil, an individual may question his morality due to lacking of social direction and deregulation of social ethic. People get frustrated when they fail to realize their desires and are being disappointed by it constantly. An example of this is suicide rate surges when there is a economic upheaval, crash of stock market or bankruptcy.

(4) **Fatalistic suicide:** Opposite to anomic suicide, fatalistic suicide happens when people are kept under tight regulation. An overregulated society has extreme rules and discipline, making its inhabitants feel suffocated by high standard and expectation. These people would prefer death to living within such harsh environment. Typical example of

this kind of suicide is students under pressure of academic performance set by parents or their teachers.

Process of suicide

Suicidal ideation refers to an individual's rumination about conducting suicidal behaviour. Suicide plan includes idea about time, location and method of suicide. Suicide attempt is a self-initiated, self-inflicted and injurious act following suicide plan performed in order to end one's life.

Figure 2.2 shows the process of suicide. Suicidal ideation is first phase, where thought of suicide appears with possibly a statement or threat of suicide. Suicide plan precedes actual suicide attempt, it is characterized by suicide rehearsal which is behavioral enactment of a lethal method for suicide.

It is important to identify suicidal ideation because taking action early may stop progression of suicide process. Therefore, spotting suicidal ideation is vital and should be top priority in suicide prevention.

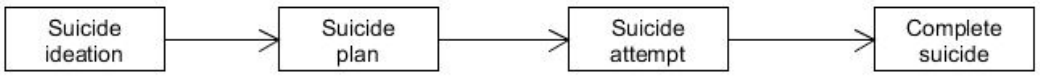

Figure 2.2: Flow diagram of suicide process

2.2 Emotion classification in suicide notes

Although the research of suicidal text started in 1960s with linguistic analysis of suicide notes to distinguish between genuine and stimulated notes, the application of NLP and ML in the same task appear in 2010 [16]. Pestian et al [39] applied ML algorithms to differentiate the notes of suicide completer and notes written by healthy control group and compared the accuracy of classifiers to accuracy of mental health providers. The best algorithm obtained the accuracy of 78%, exceeding human experts accuracy by 13%. This

promising result laid the foundation of NLP techniques on suicide notes. It also suggests the potential possibility of computer assist human in assessing suicide risk. However, the sample size for the study is only 66 which is usually considered very small in machine learning context.

In 2011, the competition i2b2 NLP Challenge[1] attracted 106 scientists from 24 teams to solve a shared task [40]. Track 2 of the challenge require participants to classify sentences in suicide notes in one or more of 15 categories which are 13 emotions: love, guilt, fear, hopelessness, forgiveness, sorrow, anger, abuse, blame, happiness, peacefulness, thankfulness, pride and 2 non-emotion classes namely instruction and information. The training set is 600 actual suicide notes collected by Dr. Edwin Shneidman and Cincinnati Children Hospital Medical Center and 300 other notes were use as test set and released after the challenge ended.

The notes are labelled by different volunteers recruited online. Unlike crowdsourcing, these volunteers are not chosen randomly but were purposely contacted directly via email or indirectly by announcement on Facebook groups. Therefore, they have some emotional connections to suicide topic. The researchers sent e-mail to approximately 1500 members of some online suicide support groups and announced the study on Facebook suicide bereavement pages. Respondents must meet some criteria: over 21 years of age, primary language is English and willing to annotate 50 notes. After that, they will be asked for general information like relationship to the lost person, how long it has been, and presence of symptoms of mental illness before suicide. They gained access to website to do the annotation task online. Only annotators that have 50% agreement rate or more with gold standard after the first 10 training notes can continue to label 50 notes more. The characteristics of annotator is summarized in Table 2.1.

Each team is allowed to submit at most three systems thus some teams have similar strategies: develop two classifiers and "combine" them into a hybrid classifier. Wang et al [49] developed a ML classifier using support vector machine (SVM) along with one-against-one approach for multi-label classification. Before training the data, they pre-processed data to aim for betting parser accuracy and give similar meaning to syntacti-

[1] https://www.i2b2.org/NLP/Coreference/

Respondents who completed the training task	169
Respondents who completed the full task	64
Male	10%
Female	90%
Average age (standard deviation)	47.3 (11.2)
High school diploma	26
Associates degree	13
Bachelor's degree	23
Master's degree	34
Doctoral degree	4

Table 2.1: Characteristics of annotators

cally different expression. For example, they correct misspellings, normalize symbol and number ('+' to 'and', '$100' to '$MONEY$'). Their second system is a rule-based classifier that create set of patterns from training data automatically and using χ^2 test and choose a fixed number of pattern sets to reduce pattern space. A notion called g-measure was defined, it is a conditional probability of sentence that contain a specific pattern belong to a specific class. The highest g-measure is chosen for that categories to be labelled for a sentence if it is higher than a specific threshold. The team has experimented to find threshold that produce highest F-score and choose it to be fixed threshold for hybrid system. A simple combination rule is used to form a hybrid classifier: if a sentence is not labelled by SVM classifier, then result of rule-based classifier is assigned to that sentence. The threshold approach gives better precision (often higher than 0.6) than recall (lower than 0.4) since there are some sentences with various emotions in them and the threshold overlook low occurrence emotion.

Similarly, Sohn et al [47] proposed a ML system and a rule emphasis system. They used Weka[2] (Waikato Environment for Knowledge Analysis) and its multinomial Naive Bayes feature to create the ML system. Some techniques are applied in the process namely token normalization (e.g. "can't", "ca n't" convert to "cannot"), classifier ensemble (using different values for a parameter) and corpus-reannotation. Their rule-based system simply utilizes Perl regular expressions to match defined patterns. Again, hybrid system combines the result of the two systems to get the final label. They have more balanced

[2]https://www.cs.waikato.ac.nz/ml/weka/

result between precision and recall with both are in range 0.5 to 0.6. They also have data pre-processing with popular methods like symbol normalization using regular expression and post-processing phase. A shortcoming of this paper is lack of deep syntactic analysis making some rare classes (sorrow, forgiveness, abuse) did not occur at all. Another improvement can be made by analyze the effect of combining two systems, why it only improves the F-score just a tiny bit compared to other systems.
Luyckx et al [33] approached the problem with a new way as they pre-process the data then carry out two processing phases: calibration and thresholding. In the pre-processing step, they tried to divided multi-labelled sentences into single-labelled fragments. If the sentence cannot be divided, then it would be discarded from training data. The calibration phase use SVM classifier with ten-fold cross-validation scheme on output of pre-processing step and evaluate lexical features, context features and lexicon-based features. The lexical features include some frequently occurred word that associated with a specific emotion (e.g. ”can't”, ”tired” for hopelessness). The context features consider the label of preceding and following sentences while lexicon-based features make use of a vocabulary of emotion. The thresholding phase applied various threshold to LibSVM (a learning package) probability estimates. In term of F-score, out of these three group, Union system of Sohn's group attained highest score of 0.5640 followed by 0.5038 and 0.5018 of Wang's team and Luyckx's team respectively. The mean performance of all team is 0.4875 with the median of 0.5027. The team with highest score achieve F-measure of 0.6139. The differences between these systems are summarized in Table 2.2.

Although all three teams have done some pre-processing steps but only Sohn's team do post-processing part to eliminate misclassified cases such as salutation being labelled instruction. They also used different learning algorithm from other teams (Naive Bayes vs SVM). This can be one of the reasons why they have smaller difference between precision and recall compared to their counterpart. Naive Bayes is usually work well on small text collection. Unfortunately, Luyckx's team could have achieved better result if they did not re-annotate the data. They chose to re-annotate because it shows promising result in the development phase but it turn out not good as they expect with F-score of 0.5018

	Micro F_1	Multi-system	Learning algorithm	Feature engineering
Sohn's team	0.5640	Yes (ML, rule-based and hybrid)	Multinomial naive Bayes	Named Entity Recognition, n-gram tokenization, token normalization
Wang's team	0.5038	Yes (ML, rule-based and hybrid)	one-against-one SVM	n-gram tokenization, POS tagging, sentiment analysis
Luyckx's team	0.5018	No	one-vs-all SVM	Multi-label training sentences re-annotated into single-label instances, unigram tokenization

Table 2.2: Comparison between teams: each team take different approaches, learning algorithm and feature engineering method

while intact data could give them 0.5230 with roughly the same precision and recall (0.53 and 0.5 respectively). In these papers, the teams use standard, well-known methods in NLP and off-the-shelf packages to implement their systems, and it is not surprise since this challenge aim is to apply techniques to problem, not investigating pure theory. The challenges and workshops like i2b2 NLP Challenge inaugurate a specific task and provide approaches to research community, promoting good solutions to be considered for similar tasks.

2.3 Analysis of suicide with social media data

Another research trend is using social media data to relate blog posts and suicide cases. Huang et al [24] collected blog entries from MySpace.com and use dictionaries containing suicide-related keywords to detect blogs with suicidal intention. The study is one of the first to use social media data to find suicidal thoughts and that, in turn, set the new goal for NLP in the mental research: identifying mental suffering victims actively using social network and are at brink of suicide. Unfortunately, MySpace lost its popularity lead to researchers switching to other social networking sites for their study and finding

the keyword is a primitive approach for a NLP task.

2.3.1 Predicting national suicide numbers with social blog posts

South Korea has the second highest suicide rate in the world (29.1 per 100,000), that make it the highest suicide rate in OCED countries (Organisation for Economic Co-operation and Development) [53].Won et al [51] build a model to predict to suicide number in South Korea. The research team collected suicide data from January 1st 2008 to December 31st 2010. Their model considers social media data in form of suicide weblog posts and dysphoria weblog posts (blog entries containing specific words relate to suicide or dissatisfaction with one's life), economic and meteorological data which are Korea Composite Stock Price Index (KOSPI), consumer price index (CPI), unemployment rate, temperature and sunlight hours of Seoul. Another interesting feature is celebrity suicides, they also have quite strong correlation with number of suicide cases in corresponding month.
The dataset is split into two sets: training set with data in two years 2008-2009 and validation set with data of 2010. They calculate numbers for 3-day epochs instead of taking daily numbers for each feature. For example, suicide count feature is total number of suicide cases nationally over the time period of three days while the temperature features are average temperature over the course of three days. The researchers defined celebrity suicide for this paper as suicides emerge on three pre-eminent national television channels (SBS, KBS and MBC). There are 6 cases match the definition, 5 out of 6 people were actors or actresses and the remaining case was former President. Univariate linear regression model is used on each candidate predictors mentioned above to check whether they are statistically significant (P-value <0.05) or not. The analysis excludes two variables namely CPI and unemployment rate. It is surprise that unemployment rate is ruled out since unemployment is thought as one of causes of depression, poverty and suicide.
The final model attained the adjusted R-squared value of 0.66, prediction accuracy of 0.79 and correlation of 0.74 on validation set (96/121 epochs). A second sensitivity analysis without celebrity suicide without the celebrity suicide variable was performed as the variable only fit short term trend. The accuracy increases to 0.83 but the correlation

decreases by small amount 0.02. Won and his colleagues notice the dysphoria weblog count is a powerful predictor for actual long term number of suicide deaths. This study has confirmed association between digital social blog posts and national suicide data and pointed out possibility of using predictive model for suicide research on a large scale.

Later in the USA, Jashinsky et al [28] investigated the suicide risk factors via Twitter data. They collected tweets (small blog entry with 140 characters restriction) with official Twitter application programming interface (API). The API allows users to retrieve tweet in a specific period of time with some available criteria such as keyword. They created a list of filter term to identify high concerning entries and attempted to remove the entries with joke, sarcastic intention. The result shows there is correlation between observed Twitter data with data obtained from real world suicides in states with highest suicide rate such as Alaska, New Mexico and Idaho. Both studies focus on public blogs containing keywords and actual number of suicidal deaths but there are differences between them: sample data duration range (3 months vs 3 years), geolocation (each states vs whole nation), number of variables. This research suggests it is a good idea to incorporate social media data into model that inspect suicide trend across geographic regions but more filter must be developed to get messages of real victim rather than relying on keywords in blog.

2.3.2 Confirming Werther effect on social media

Kumar et al [31] investigated Werther effect of online forum. The term ”Werther effect” refers to the phenomenom of increase in number of completed suicides and suicide attempt after reported case of celebrity suicides. The term is derived from the main character in the novel *Die Leiden des jungen Werthers* (The sorrows of young Werther) by the famous German writer Johann Wolfgang von Goethe. Imitative suicidal behaviour follows well-known celebrity suicide is ”copycat suicide” where lethal means and methods of suicide in a celebrity suicide case announced publicly on media are being used in suicide attempts. Kumar and his colleagues collected data on Reddit, a social media platform

encompasses many subforums, each of those corresponds to a particular topic. Reddit provides an official API just like Twitter with various features including collecting posts, comments and metadata accompanied with them.

To collect data related to suicide, the researchers used API to collect posts from the subforum r/SuicideWatch (SW - a suicide support forum) in time period October 16th, 2014 - December 19th, 2014. A total of 66,059 posts from 19,159 unique users was crawled. The way of identifying celebrity suicide in this study was different from that of the introduced work above [51]. They take suicide from a Wikipedia page listing celebrity suicides and get 10 suicides. To make sure those suicides are 'important' enough to cause Werther effect, a comparison of Wikipedia page view count of corresponding celebrity between two periods (2 weeks before and after) was conducted. The measure was converted to z-score, 9 out of 10 suicides show positive change in z-score thus being incorporated in the model.

The hypothesis to be tested in this study is whether or not there exists a surge in the number of posts in SW after a celebrity suicide. A baseline was established by choosing 20 consecutive two-week time periods as pairs. These pairs must have no celebrity suicide in its time and starts in day of week. A set of control group comprises of 21 mental health (MH) subreddits is listed. The purpose is to determine the increase in posts of SW is caused by celebrity suicide but not by mental health problems. 32,509 posts from 23,807 unique users of MH subreddits were obtained. Content analysis that includes linguistic measures, n-gram analysis and topic model analysis was carried out.

The research team categorize linguistic features: affective attributes, cognitive attribute, linguistic style attributes and social attributes. The first and second category are measures derived from pschycholinguistic lexicon LIWC[3] (Linguistic Inquiry and Word Count) such as positive, negative affect, synonyms for set of cognitive and perceptive words like "see", "hear", "feel", "death". The third features are lexical density (words that have POS tag verb, noun, adjective, adverb), temporal reference (post, present, future tenses), social/personal concerns (words in topics of family, friend, work, home), interpersonal awareness (first person singular/plural, second and third person pronouns). Social at-

[3]http://liwc.wpengine.com/

tributes include metadata for a post: length of post, number of upvotes, downvotes and comments, average comment length. N-gram analysis involves log likelihood ratio of unigram, bigram and trigram between prior and post celebrity suicide. Topic model analysis runs Latent Dirichlet Allocation (LDA) [7] to retrieve 50 topics. After that, the model calculated posterior probability of each topic for both items in each pair time period. The objective of this analysis is to measure the increase in topic by computing the difference between posterior probability of pre- and post-suicide.

The result of comparison between number of posts in two periods is no surprise as mean change in z-score of number of posts after suicide is higher than that of baseline and control group (3.64 compare to 1.95, 0.45). The results of statistical testing on linguistic features give insight of content between pre- and post-suicide time period. Post-suicide posts show more negativity in emotional expression, cognitive biases and lower lexical density. They also reveal SW users is more lingering to the past than interest in future. Besides they show little concerns over society but focus on their own self by using more first person singular. The n-gram analysis report more symptoms of depression, anxiety after celebrity suicide. Negative n-gram like “i hate it”, “i gave up”, “tired of living” rise in frequency post-suicide.

The work verified the Werther effect in posting activity of an online forum. The increase of posting activity in suicide support forum after a celebrity suicide is not coincidence but follows a consistent pattern. However, we should be cautious when drawing inference from this study as we do not have sufficient evidence on how posting activity related to actual number of completed suicides or suicide attempt.

2.3.3 Identifying at-risk individuals and suicidal thoughts on social media

Analysis of Twitter posts before suicide attempters

Coppersmith et al [11] explore the Twitter posts of suicide attempters to see whether or not their linguistic style is different after the attempt. The researcher crawled the Twitter data to identify people who have publicly disclose their suicide attempt and enough

information for reader to infer the time of it. 554 users are found to had declared their suicide attempt, only 163 of them indicated the exact date and 125 users have data prior to their suicide attempts. To avoid situation in which users did not actually attempt suicide or they was joking with their friends, a human examined tweets to ensure three conditions: the attempts seems to be true as stated; users are talking about their own attempts, not attempts of other people; and time of the attempts can be determined.

Demographics of users gives information about age group and gender of suicide attempters. In the data, more women are present than men and almost all of attempts are in the age range from 15 to 29. This distribution does not follow Twitter demographics as middle age and adult users account for 12% and 37% of total number of users respectively. It is expected given the nature of social network usage in youth where youngsters are more active and open to share opinion online.

The linguistic characteristics of suicide attempters is compared to neurotypical group (control group). Token is the unit for comparison, token here refer to a single word, emoticon or symbol. An emoticon is defined as visual representation of a facial expression using punctuation marks, numbers and letters, usually written to express a person's feelings or mood. For example, ":D" denotes a big smile. Similarly, an emoji is a graphic symbol that represent an object or an facial expressions but offer more versatility than emoticon due to larger range of expression including weathers, animals, places and its integration in website/application interface. Emoticons and emojis are very useful features to capture users' emotion at the time of writing. Notable pattern is observed: control group use more emoticons and emojis, users often talk about suicide after the attempt rather than before it, users' language is self-attentional prior to an attempt.

Preprocessing phase is quite straightforward, all usernames and website addresses are replaced by single tokens "@" and "*". For example "Check out this awesome site: https://theawesomesite.com powered by @username123 :)" would be "Check out this awesome site * powered by @ ! :)". Character n-gram model with $n = 1, 2..., 5$ is used together with logistic regression to separate suicide attempters from matched users in control group. Figure 2.3 taken from the paper [11] shows the ROC (receiver operating characteristic) curve of this classification task. A ROC curve illustrates the tradeoff be-

tween true positive rate (recall) and false positive rate to aid the decision making process for optimal model. At a single point, the classifier achieves 70% recall with only 10% false alarms.

A report [29] has shown statistics of suicide attempt in youth and we can expect 4-8% of users aged 15-29 will or have tried to take their life at some time. That makes the number of neurotypicals is ten times more than the number of at-risk youngsters. Assuming a population of 1000 people within age range 15-29, we expect 40-80 people will attempt suicide but taking 60 as a single number for simplicity. If the classifier maintains its performance in the experiment, we can identify 42 (70% of 60) of them and misclassify 94 (10% of 940) attempters as neurotypicals. This false negative need more screening for better assessment. Although this study showed some promising result in identifying at-risk individuals, the sample population is limited to female youngsters aged 15-29. This subpopulation of all age and gender group of Internet users is by no mean small. However, generalizability of this method is not guarantee given characteristic of data in this paper.

Classifying suicide communication content on Twitter

While most studies are concerned about binary classification of suicidal intent, Burnap et al [8] explores the robustness of machine classification in a multiclass labelling task of Twitter posts where all classes are related to suicide but only one class shows suicidal intent. To define a list of terms to represent suicidal language, four websites with suicidal themes were selected for crawling. These websites either have section[4, 5] or are entirely dedicated to suicidal discussion[6, 7]. 200 posts from each of these websites plus 1000 posts tagged with the word 'suicide' from popular microblogging site Tumblr[8] were labelled whether this post is of suicidal person or not by human on crowdsourcing online service Crowdflower[9]. The labelled corpus was analyzed by applying Term Frequency/ Inverse

[4]http://www.experienceproject.com
[5]http://www.enotalone.com
[6]http://www.takethislife.com
[7]http://www.recoveryourlife.com
[8]https://www.tumblr.com
[9]http://www.crowdflower.com

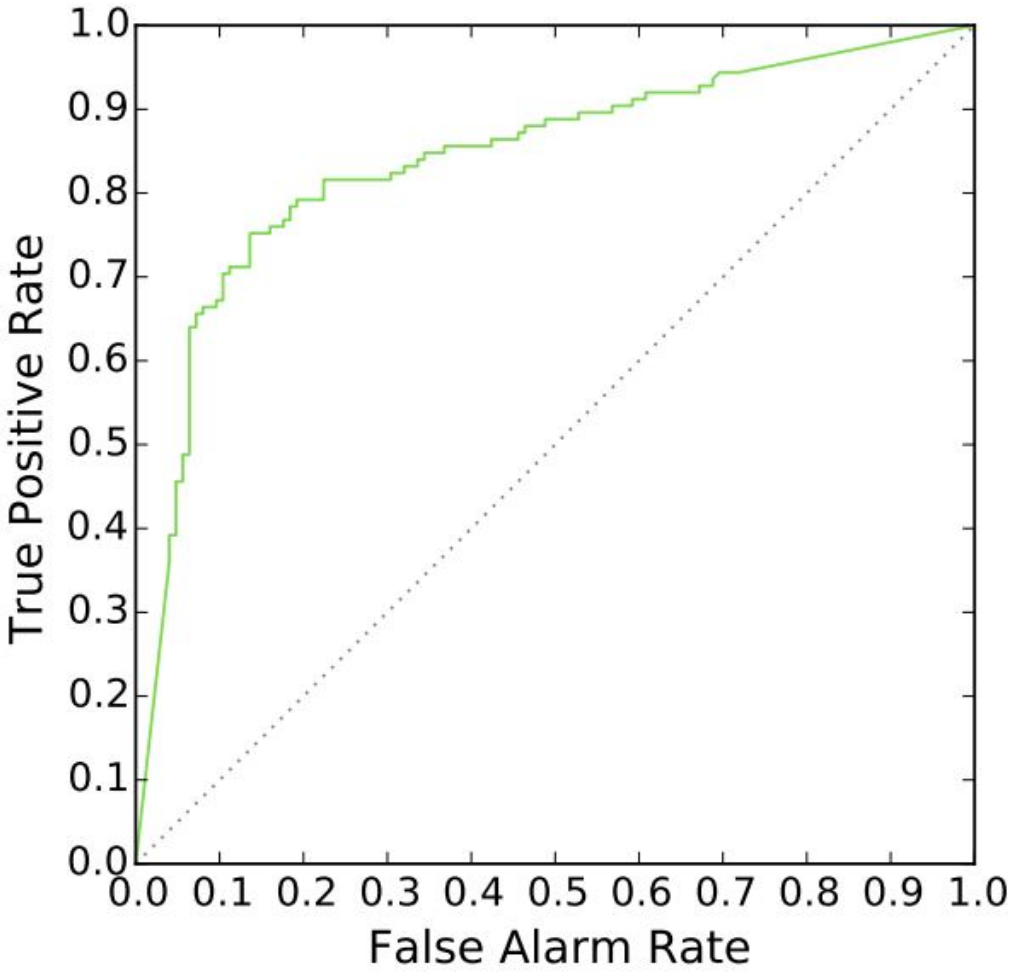

Figure 2.3: ROC curve for separating users who attempted to take their life from matched neurotypicals [11].

Class	Description	% of dataset
C1	Evidence of possible suicidal intent	13
C2	Campaigning (i.e. petitions etc.)	5
C3	Flippant reference to suicide	30
C4	Information or support	6
C5	Memorial or condolence	5
C6	Reporting of suicide (exclude bombing)	15
C7	None of the above	26

Table 2.3: Types of Suicidal communication with relative % proportion in dataset [8]

Document Frequency (TF.IDF) process to *n*-gram language model (n = 1, 2,..., 5) at word level. Two experienced suicide researchers examined the top 500 terms from analyzing task to finally produce the list of 62 keywords that suggest possible suicidal ideation. Some examples of words in this list are ”never wake”, ”kill myself”, ”don't want to exist”. Over 4,000,000 tweets were crawled after applying suicided-related search terms in the time period February 1st, 2014 - March 15th, 2014. Another set of search terms is proposed using name and surname of reported people die by suicide in England over the same period. The researchers sampled the two datasets to get random 1000 posts, 800 from the first dataset and 200 from 'names' dataset. Again, the same crowdsourcing service was used to classify posts into seven classes. Table 2.3 taken from the paper [8] gives description of classes. After removing tweets with low inter-rater agreement score computed by CrowdFlower (less than 75%), 816 tweets remain to be used for ML classifier. There are three feature sets in the experiment. Feature set 1 representing lexical characteristics of the sentences contain the following:

- *Parts of speech*: frequency of each POS tag as a feature.
- *Structural features*: inclusion of negations in the sentence, external communication link (URL, reply), usage of first person pronoun.
- *General lexical domains*: using Domains labels from Princeton WordNet®[10] (a large lexical database for English where words are grouped into sets of synonyms [19], hereafter referred to as WordNet) to generate features represent categories such as home, psychology, etc.

[10]https://wordnet.princeton.edu/

- *Affective lexical domains*: include categories representing moods, emotional responses such as anger, sadness, love, hate, joy, etc.
- *Sentiments Score*: SentiWordNet[11] (a lexical resource for opinion mining based on WordNet) gives each word a score in interval [0,1] for negativity and positivity. The total score of all words in a tweet were used as features.
- *Frequent words*: The most used of unigrams, bigrams and trigrams in the training set.
- *Keyword list*: list of 62 keyword described above.

Feature set 2 represents sentiment and emotional features. This set of features was compiled using LIWC to extract more labels related to affective emotions in a tweet. These include term that have connection to topics causing distress like death, money, health, achievement, job. LIWC also have three categories frequently used in mental health research namely 'cognitive mechanisms', 'affect' and 'social words'. These mentioned features encompass feature set 2. Feature set 3 consists of regular expressions and pattern matching rules to match misspelled/shortened words. It is an attempt to recognize important phrases in a noisy and informal text environment due to number of character limit on Twitter. Some example phrases are "suicidal / cutting / bad / these ... thoughts / feelings", "want/wanted/wanting to die", "call/offer for/of help", "kill/killing/hate myself", "talk/speak to someone/somebody", "took/taken his/her own life". Three set of features was combined into one called combined set. Principal Component Analysis (PCA) was applied to the set to reduce dimension and retain sets of uncorrelated features (principal components). The text of tweets is transformed into word vector of uni-, bi- and trigrams and 500 words retained as features. The total number of feature is 1444.
Baseline was established using SVM, rule-based (Decision Tree - DT) and Naive Bayes (NB). All these three classifiers were used in an ensemble classifier called Rotation Forest (RF). RF divides the feature set into smaller sets and run PCA on each set to generate principal components. These components are used to build classifiers. Ensemble meta classifiers produce out from multiple classifiers' output by using a voting mechanism.

[11] http://sentiwordnet.isti.cnr.it/

Class	C1	C2	C3	C4	C5	C6	C7
C1	**58**	0	15	0	0	0	5
C2	0	**18**	1	4	0	4	1
C3	11	0	**143**	0	1	5	17
C4	0	4	5	**18**	0	2	6
C5	1	1	1	0	**31**	1	1
C6	0	6	9	7	2	**76**	4
C7	20	0	23	0	2	4	**94**

Table 2.4: Confusion matrix for the best performing classification model [8]

The voting rule here is choosing output with highest probability across all base classifiers. After some rounds of experiment running RF on the three baseline classifiers, the team dropped DT from the ensemble classifier, kept only SVM and NB in it.

Overall, the RF meta classifier using Maximum Likelihood voting mechanism with combined dataset achieved the best performance: precision = 0.732, recall = 0.729 and F-measure = 0.728. The key class of interest - suicidal ideation is inspected one more time and the RF model once again achieved highest recall (0.744) and F-measure (0.690) but not precision (0.644 while the highest being 0.657). Misclassification mostly occurred in class 1 (suicidal ideation) and class 3 (flippant reference to suicide) indicating challenge posed by sarcasm and irony. Table 4.5 taken from the paper [8] illustrates the confusion matrix of the best performing classification model on all seven classes. This research explored the approach of using ensemble meta classifier to identify suicide-related communication on social media. If previous introduced works only interest in binary classification of suicidal and non suicidal posts, this one expand the scope the task. Multiclass labelling task can be useful in a broader context when other suicide-related messages may help automatically recognize news, reports, campaigns. Incorporating such classifier is beneficial for general-purpose crawling system. Although this study used some techniques specifically for Twitter, the main idea is likely applicable to other social media sites as well.

Identifying individuals who develop suicidal ideation on Reddit

Another prominent social media site is Reddit[12]. Reddit is one of the best sites for researchers because all the data is publicly available and development team even provides official API with many functionalities including collecting posts and comments. De Choudhury et al [14] presented a statistical methodology to identify individual who likely to discuss suicidal ideation given disclosure of their mental health earlier on Reddit. The research team collected posts and comments data from suicide support subreddit r/SuicideWatch (SW) and 14 mental health (MH) subreddits. All the posts and comments are in the time period February 11th, 2014 - November 11th, 2014. The size of content posted is considerable: 63,485 posts, 209,766 comments from 35,038 users of MHs and 16,348 posts from 9,224 users of SW. Two user classes are constructed for the classification task:

- **MH**: users who only posted on MHs in the first six months but never posted on SW in the last three months.
- **MH→SW**: users who only posted on MHs in the first six months and posted on SW in the last three months

The process yielded 440 MH→SW users and 28,831 MH users. To balance two classes, 440 users from MH were sampled randomly. The data reduced in sized accordingly: 4,731 posts, 46,949 comments from MH→SW and 8,318 posts, 54,086 comments from MH. The prediction model adopted five sets of features: *Linguistic structure*, *Interpersonal awareness*, *Interaction*, *Content* and *Full*. Automated readability index; percentage of nouns, verbs and adverbs; linguistic accommodation encompassed *linguistic structure*. *Interpersonal awreness* includes percentage of first person plural/singular, second and third person pronouns. *Interaction* measures number of posts and comments each user had written, number and length of comments received, mean score of posts and duration between post submission and the first comment to post. *Content* features are all the unigrams and bigrams from posts and comments of both MH and MH→SW. The *Full* set is combined set of all four mentioned sets of features.

[12]https://www.reddit.com/

Some observations on difference between characteristic of MH and MH→SW give general impression about two cohorts. Firstly, MH→SW users appear to display poorer linguistic structure and lower readability index. Secondly, MH→SW users is more self-focused and show little social concern by using first person pronoun more often than control group. Finally, MH→SW users tend to have fewer posts but their posts are usually greater in length, indicating high degree of social isolation.

Content analysis entails causal inference. The authors chose propensity score matching [45] to identify tokens that increase the likelihood of posting on SW. The tokens obtained make sense intuitively and increase the probability by 30-53%. Some examples of these tokens are "useless", "anxiety", "no friends", "have nothing", "to cry". In contrast, some tokens decrease the likelihood on SW: "counseling", "intimate", "hope it", "and enjoy" significantly reduce the chance by 50-57%. Besides identifying treatment tokens, De Choudhury and her colleagues proceeded to link these tokens to specific themes and normalized spectral clustering algorithm was used. The algorithm maps the original space of similarity values to eigenvectors of a Laplacian matrix and applies standard clustering method (e.g. k-means). Six dominant themes corresponding to the first 6 eigenvalues of the matrix. Two researchers with experience on mental health content of social networking sites coded these 6 clusters of tokens and achieved high agreement rate (Cohen's κ = 0.74). The six themes and some example tokens are:

- *Hopelessness*: "have nothing", "no real", "kill myself", "abandoned", "die".
- *Anxiety*: "anxiety", "panic", "to cry".
- *Impulsiveness*: "ending", "freaking out"
- *Self-esteem*: "hate it", "giving up".
- *Loneliness*: "my friends", "my parents", "alone".
- *Severe or stigmatized illness*: "depression", "psychosis", "disorder".

The main point of the study is the supervised learning task of predicting and classifying MH→SW users from MH users. The train/test ratio is 80/20 and user classes in each set are balanced (704 users for training set plus 176 users for held-out validation set). The ML

Actual/Predicted	Class 0	Class 1	Total
Class 0	73	15	88
Class 1	20	68	88
Accuracy	83.5%	77.5%	80% (mean)
Precision	0.79	0.82	0.81 (mean)
Recall	0.83	0.78	0.81 (mean)
F-1	0.81	0.80	0.80(mean)

Table 2.5: Classifier performance distinguishing MH→SW and MH [14].

classifier used is regularized logistic regression. 10-fold cross validation was performed on the training set to tune parameters of all five sets of features. To reduce randomness of cross validation, the team run the models 10 times and reported the result that is most fitted to the data. Performance of the best model (*Full*) on held-out set is reported in Table 2.5.

There are several limitations in this study. First of all, there may exists self-selection biases as Reddit does not enforce one account per user policy thus a user can have throwaway account, where user use an account to post in particular topics then abandon it, or register multiple accounts. Therefore, a person who intrinsically belongs to MH→SW may get excluded if he or she uses another account other than the main one to post in SW. Furthermore, we see imbalance between size of MH and MH→SW and the authors decided to sample MH. This method shows robustness of the classifier in theoretical scenario but cannot ensure effectiveness in a larger population.

Analysis of commentary related to suicidal ideation on Reddit

Following the previous study, De Choudhury and Kiciman [15] continue to assess the impact of online social support in suicidal ideation. Data collection method is the same but now the focus shifts to comments received instead of posts from at-risk individuals. The dataset contained 62,024 comments from 32,362 users for 440 MH→SW users and 41,894 comments written by 21,358 users for 440 MH users. All comments are tokenized, stopword removed and timestamp attached, into n-gram tokens (n = 2). Comment-tokens in a user's timeline is considered *treatment* to that user. Post- and comment-tokens occur before a *treatment* are *covariates*. According to the terminology, a *treatment group* is a

group consists of users who received *treatment* and a *control group* characterized by users who have not encountered *treatment*. Stratified propensity score matching is a statistical method to estimate the effect of a treatment from confounding covariates. The treatment group and the control group are treated as one big group and are divided into strata in which covariates of treatment subgroup and covariates of control subgroup are homogenous statistically. In this case, the dataset is stratified based on propensity score of receiving a particular treatment. Estimated propensity score is likelihood of receiving a treatment given covariates and it is calculated by a machine-learned function. To learn a propensity function, researchers applied averaged perceptron algorithm on vector $H = h_1, h_2..., h_n$ where h_i is a binary variable that flags 1 if token i appears before the treatment token and flags 0 otherwise. The treatment tokens are tokens that appear in more than 10 users' timeline of MH subreddits, the total number of treatment tokens is 11,278. The dataset is divided into 10 strata.

To ensure the strata is balanced, meaning users of both treatment and control groups are having similar effect when receiving comments, two human raters are employed. One is an expert in mental health content on social media sites and the other is a mental health professional. 150 treatment tokens with highest and lowest z-score were chosen. Examples of negative treatment effect tokens (increase likelihood of being in MH) are: ″lucky″, ″a reasonable″, ″enjoys″, ″gently″, ″heart and″; positive treatment effect tokens ((increase likelihood of being in MH→SW): ″pain and″, ″not easy″, ″struggled″, ″hating″. The human raters' task is to mark the similarity (0 or1 where 1 indicates high similarity) of 300 post pairs where a post pair consist of one post from treatment group user and one post from control group user. The agreement rate between two raters was high (Cohen's $\kappa = 0.81$). This coding frame pinpointed which strata is balanced.

Table 2.6 taken from the paper [15] shows 40 comment tokens with highest and lowest z-score. The first half displays negative treatment tokens, the second half display positive treatment token. Coverage is percentage of users who belong to unclipped strata. PMI is the pointwise mutual information between the treatment token and the SW outcome (a user belongs to MH→SW or not). Some tokens decrease likelihood of participating in MH→SW in the future by a significant percentage namely ″gently″, ″sure of″, ″be

Feature	Count	Coverage	Effect	z-val	χ^2	PMI
gently	43	0.3	-0.31	-2.18	2.55	0.02
sure of	43	0.49	-0.22	-2.04	2.31	0.15
is helpful	37	0.49	-0.12	-1.45	1.86	0.15
be tough	39	0.51	-0.25	-1.44	0.82	0.01
fight the	34	0.51	-0.23	-1.22	1.53	0.16
enjoyed it	46	0.3	-0.18	-1.1	0.71	0
be ready	39	0.49	-0.04	-1.06	1.41	0.18
nice i	54	0.39	-0.06	-1.06	1.01	0.13
really fun	35	0.2	-0.06	-1.01	1.23	0.13
totally agree	37	0.49	-0.05	-0.93	0.98	0.17
completed	54	0.4	-0.09	-0.91	0.25	0
enjoys	32	0.51	-0.08	-0.85	1.23	0.18
defeat	46	0.4	-0.28	-0.83	0.55	0
to defend	44	0.2	-0.08	-0.79	1.18	0.14
was nice	37	0.49	-0.12	-0.77	0.97	0.1
really liked	40	0.51	-0.1	-0.61	0.88	0.08
be super	42	0.6	-0.03	-0.54	1.38	0.17
instructions	54	0.39	-0.17	-0.53	0.32	0
your home	33	0.49	-0.11	-0.45	0.55	0.08
kindness	42	0.4	-0.11	-0.37	3.42	0.19
proud	127	0.6	0.31	5.35	4.14	0.55
a hobby	35	0.49	0.53	4.87	4.57	0.76
am sorry	34	0.49	0.53	4.77	4.69	0.77
suicide	123	0.49	0.28	4.67	4.21	0.49
you wish	32	0.49	0.55	4.54	4	0.8
together with	32	0.49	0.51	4.54	4.16	0.72
medication	114	0.49	0.35	4.51	4.13	0.56
friend you	32	0.49	0.52	4.43	3.8	0.74
your opinion	33	0.4	0.5	4.35	4.44	0.69
to respond	40	0.49	0.49	4.34	3.41	0.69
i care	40	0.4	0.55	4.31	4.57	0.81
depressed	187	0.4	0.3	4.28	5.01	0.53
seek	132	0.39	0.27	4.26	4.41	0.47
pain and	51	0.3	0.58	4.24	6.05	0.87
do well	44	0.4	0.56	4.11	4.32	0.82
stay strong	48	0.4	0.52	4.09	4.16	0.74
medical	133	0.49	0.19	4.05	2.67	0.37
vent	84	0.49	0.32	4.02	3.59	0.54
hating	39	0.49	0.52	3.99	3.02	0.74
misery	34	0.49	0.5	3.96	3.13	0.7

Table 2.6: Comment tokens given by propensity score matching that contribute to increased or decreased change in likelihood of being in MH→SW or MH respectively [15].

tough", "fight the", "defeat". Whereas some tokens show very high positive treatment effect such as "suicide", "hating", "pain and", "am sorry". To grasp the context of treatment token, 200 comments that contained the 40 tokens (100 for negative treatment tokens plus 100 for positive treatment tokens) in Table 2.6 were sampled randomly. The same human raters first coded the first 50 comments and developed a shared vocabulary to come up with a support coding scheme. Using that scheme, they continued to code the remaining comments and reached a high inter-rater agreement rate Cohen's $\kappa = 0.86$. The comments are coded into one of six social support categories: emotional support, esteem support, information, instrumental support, network support and acknowledgements. De Choudhury and Kiciman delved into distribution of social support type and found high number of esteem boosting comments (31%), network support (23%) and emotional support (16%) in coded comments received by MH. While informational support and acknowledgements only occupied a small proportion of comments contained negative treatment effect, they actually were dominant social support types in comments associated with MH→SW (40% and 23% respectively). Table 2.7 taken from the paper [15] gives example of comments from each types of support.
This research provided insights of the connection between commentary representing social support and future risk of developing suicidal ideation. Different types of social support were investigated and some of them show different effects from others, i.e. some specific types are associated with MH who do not display behaviour of posting about suicide ideation while MH→SW received many comments categorized as informational support and acknowledgements. This implication can help to build a guideline to comment in sensitive forums like mental health and suicide support discussion forums. However, this method cannot infer true causality as it can only gather online data but not other potential offline resources that affect users' behaviour such as medical records or offline activities.

In summary, we introduced concepts, types and process of suicide followed by researches utilizing NLP and ML to build automated systems for suicide related prediction. Firstly, we present the task of assigning emotion labels to each sentence of a suicide note.

Higher likelihood of being in MH	Higher likelihood of being in MH→SW
Emotional Support	
i *totally agree*. It is hard. I have been there and it is not easy to handle the financial stress, buying a house, girlfriend being eight months pregnant, car issues, job issues, family issues. (↓5%)	I've recently lost friends whom I've known for 10 years, due to me being 'insensitive'. So yes sadly it does happen, I get you and what you are doing through. you are *not alone* (↑16%)
Esteem Support	
cheers mate, *fight the* stigma, you can do it! (↓23%)	You have great potentials in self-actualizing your own situation and ending your *misery*. (↑50%)
Informational Support	
Ever thought of trying to find professional care? I suggest you do that. You need to give life a second chance it may surprise you a lot. I know it can *be tough*, but worth it (↓25%)	If your issue is with the taking of *medication*, talk to them about taking it, discuss your issues with it. Like the guy above said, it may help and could be worth a try, but it is good to discuss concerns about that sort of thing with the person prescribing it. (↑35%)
Instrumental Support	
Start by going for meditation. it can *gently* help you break habitual negative thought patterns, and might also help you get a little bit of that "distance" from yourself that you are looking for (↓31%)	Bro, eat healthy, run, keep your room clean, actively suppress negative thoughts, force yourself to do something productive, even if it's just pursuing *a hobby*. (↑53%)
Network Support	
There is no reason to be nervous and yet everyone here understands and have been precisely at the same place you are in your brave post. [...] i hope some of this discussion *is helpful* to you. (↓12%)	Thats not true at all. everyone in this community really wants to hear your story. They would want *to respond*. Everyones story is worth a listen don't you think? (↑49%)
Acknowledgements	
Exactly this. i would *be super* frustrated too. Anxiety is debilitating and very difficult to cope with (↓3%)	I understand you are *depressed*. Depression is the annihilation of motivation. So it's no wonder u quit the job (↑30%)

Table 2.7: Example (slightly paraphrased) comment excerpts containing one of the tokens identified to significantly decrease or increase likelihood of being in MH→SW or MH. We show the specific tokens in italics, and their treatment effects inside brackets [15].

The participating teams usually develop two separate classifiers and combine those two into one using voting mechanism of ensemble meta classifier. The challenge promoted sentiment analysis in suicide research with high quality dataset curated by human. The trend of using social media data in suicide prediction started around 2007 when a study investigated MySpace websites to find suicidal posts using a set of particular keywords. Following that direction, there are papers examining the accuracy of machine prediction for national suicide numbers in South Korea and USA using social blogposts. After that, efforts to utilize ML for analyzing large amount of data on social networking sites continue to expand. With Reddit and Twitter as source of data, researchers attempted to predict people who have high risk of suicidal thoughts and behaviour. The main task in their studies was usually binary classification of deciding whether a user is at risk of developing suicidal thoughts or not given their posting history [11, 15, 15] . There is one research that attempted multiclass classification [8].

These studies gave promising results with accuracy being more than 70%. However, the researchers sampled only a small proportion of the datasets obtained for testing classifiers. We believe that this is a research opportunity and plan to apply ML classifiers on a bigger dataset.

Chapter 3
Tools and resources

3.1 Framework and libraries

3.1.1 NLTK

NLTK[1] (**N**atural **L**anguage **T**ool**K**it) is a popular framework for working with human language. It is developed by Steven Bird[2], Ewan Klein[3] and Edward Loper[4]. The project was started in 2001 and still continue to update and release at the time of writing. The platform provides a suite of processing libraries for text classification, stemming, tokenization, parsing, tagging and semantic reasoning functionalities. It also serve as wrappers for other NLP libraries such as Stanford CoreNLP [35]. We can easily access more than 50 corpora and other lexical resources such as WordNet® through interfaces of Python programming language.

Authors of NLTK also published the book *Natural Language Processing with Python* [6] to give practical introduction and hands-on guide to programming for NLP. The book covers many topics of computational linguistic, their examples by graphical demonstrations and sample data. Along with comprehensive online documentation, readers learn the fundamentals of writing Python scripts for NLP tasks like categorizing text, working with corpora and analyzing linguistic structure of them. Both NLTK and the book are free of charge and publicly available online. As a result, NLTK is used as a teaching tool for NLP education in many universities around the world.

[1] http://www.nltk.org/
[2] http://www.stevenbird.net/
[3] http://homepages.inf.ed.ac.uk/ewan/
[4] http://ed.loper.org/

3.1.2 Scikit-learn

Scikit-learn[5] is a free software package for machine learning and data analysis written in Python [38]. The project was originally started in 2007 by David Cournapeau[6] and still under active development as of 2017 by many different researchers and developers. Like NLTK, scikit-learn is designed for Python and compatible with other numerical and scientific Python libraries like NumPy and SciPy. Scikit-learn features machine learning algorithms for classification, clustering and regression tasks such as linear regression, SVM, k-means, decision tree and boosting. It is very well-maintained with stable release and online documentation on its website. Furthermore, each algorithm and its variant are illustrated by practical examples including brief explanation with citation, figures and sample code with comments to enhance readers' understanding of the problem.

3.2 Employed NLP methods

3.2.1 Tokenization

Tokenization is usually a crucial step appearing in early stage of the NLP workflow to facilitate processing with more complex analytical techniques. Tokenization is the task of divide a character sequence into defined document units referred as tokens. A token is defined as "an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing" [34]. In theory, tokens can take form of any textual elements of chosen granularity such as punctuation or fomulae but in practice, tokens are usually words.

As shown in Figure 3.1, tokenization have dropped punctuation of the sentence and each word is getting full meaning. However, tokenization phase may encounters difficult choice in a broader context. For example, take a look at the sentence "*Tokenization*

[5]http://scikit-learn.org/
[6]https://github.com/cournape

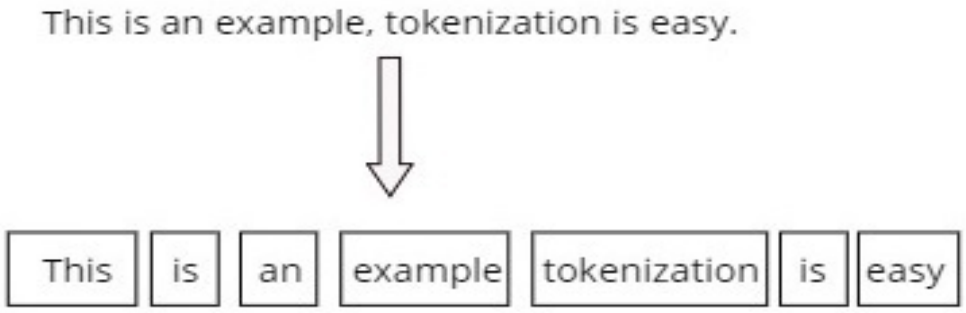

Figure 3.1: Example of tokenization. Tokenizer splits a sentence into words.

isn't easy". Two words Tokenization and easy are pretty straightforward since they are splitted by whitespace. In contrast, *isn't* can be tokenized by different rules: isn't, isnt, is n't, isn t. Furthermore, each language and topic contain jargon and words refer to specific entities but are not in standard dictionary. For instance, AK-47 is name of a semi-automatic rifle, programming language C# and so on. Computer era brings new types of technology terminology, which appear frequently on online document, such as website addresses (https://www.google.com/), email addresses (johndoe@gmail.com), decimal IP addresses (123.172.2.1). These character sequences should be recognized as one entity but in some cases, it would cause unnecessary index to be included in the extracted dictionary. In the experiment of this thesis, NLTK Tokenizer package, which is improved TreebankWordTokenizer [7] along with PunktSentenceTokenizer[8] for the specified language, was used.

3.2.2 Stop words removal

Stop words are words that contribute very little in term of semantic, thus being considered low value in IR systems. These words usually appear with high frequency in collection. To remove stop words, high use words must be manually assessed for semantic content.

[7] http://www.nltk.org/api/nltk.tokenize.html#nltk.tokenize.treebank.TreebankWordTokenizer
[8] http://www.nltk.org/api/nltk.tokenize.html#nltk.tokenize.punkt.PunktSentenceTokenizer

If a word is identify as a stop word, it is added to a list called *stop list*. The length of a stop list varies between less than twenty (15-18 terms) to a few hundred words (200-300 terms). Depending on purpose and scale of a system, a stop list can be very general or specifically relate to a topic.
Using a stop list can potentially reduce query time due to smaller index of a system. However, web search engines in practice do not use stop lists partially because of immense server clusters behind them generating enormous computing power.
Common words belong to world classes article, conjunction, preposition and frequent words like *now* or *very* are included in stop lists. General-purpose stop lists are available online and can be easily incorporated into IR systems. Sometimes removing stop words can be detrimental as vital information may be missed after going through stop list filter (e.g. *"To be or not to be"* is a well-known verse that may get taken out of indexed vocabulary) so utilization of these lists must be carefully examined. In the experiment of this thesis, Stopwords corpus of NLTK was used. It is a corpus containing 2400 stopwords for 11 languages. Stop list for English contain 153 words.

3.2.3 Part-of-speech tagging

A Part-of-speech (POS) tag is category of words assigned to a word which display similar (primarily syntactic) grammatical properties of it in the sentence. Words that share the same POS tag typically play similar roles in structure of sentences. Example of some POS tags are NOUN (noun), ADJ (adjective), VERB (verb), PRON (pronoun). In practice, there are many sets of POS tag (tagset) such as Universal[9], Treebank. Some tagsets usually comprise fine-grained tag like "singular proper noun" for more complex analyzing task. In early years of building automatic POS tagger, handwritten rules distinguish parts of speech much like decision tree. These rules may work in a particular type of document or some set of rules can be applied widely to many text corpora of a language.
When machine learning techniques are utilized for POS tagging task, there are three main approaches: statistical tagging, rule-based tagging and hybrid tagging. Statistical tagging

[9]http://universaldependencies.org/u/pos/

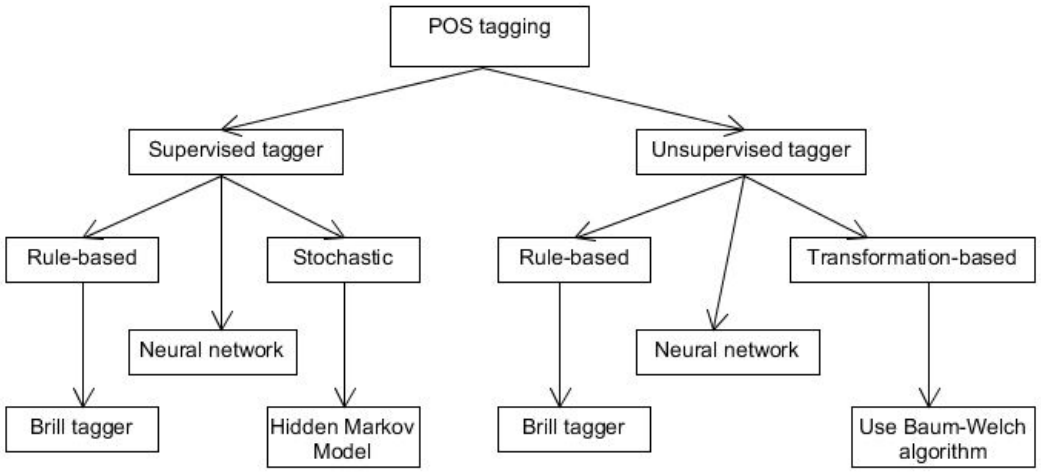

Figure 3.2: POS tagger classification

relied on corpora annotated by human as training data to compute probability of a specific token belong to a POS tag category. The drawback of statistical tagger is some tagged sequences are not correct due to low usage in a language. In other word, it is susceptible to rare case of word in an unfamiliar context and most of the time misclassify to usual tag of that word. Hybrid approach require both hand coded rules and probabilistic features of words, but it is not widely used in practice since balancing weight between these two features is not an easy task for a general-purpose tagger.

Figure 3.2 shows classification of POS tagger. Neural network, rule-based tagger can be either supervised or unsupervised learning paradigm. POS tagging plays an important role in the experiment of this thesis as we need to extract tags like nouns, verbs, adverbs. In the experiment, we used NLTK POS tagger which is a neural network tagger trained on Penn TreeBank corpus.

Sentiment analysis

Sentiment analysis is the process of identifying, quantifying the sentiment content of a text using NLP or statistical methods. The general approach is to enumerate words that express emotions, opinions, attitudes and assign scores to them. There are several promi-

nent lexicons for sentiment analysis such as LIWC, SentiWordNet, SenticNet[10], VADER[11]. The lexicon we chose is VADER (**V**alence **A**ware **D**ictionary and s**E**ntiment **R**easoner). VADER is a simple rule-based model built on lexical features which are attuned to sentiments expressed in social media context [26]. VADER returned four types of scores for each sentence: positive, negative, neutral and compound. The first three scores are proportions of text that fall in each corresponding category thus they are add up to 1. The compound score is in interval [-1, 1] with -1 indicates extreme negative and 1 indicates extreme positive.

3.3 Employed ML models

We use three off-the-shelf models offered in scikit-learn for the experiment of this thesis.

3.3.1 Naïve Bayes

Naïve Bayes is a probabilistic model based on Bayes' theorem. It assumes features are independent from each other. Given a feature vector $\mathbf{x} = (x_1, x_2, ..., x_2)$ representing n independent variables and a target class variable y, the conditional probability of instance $\mathbf{x}$ belong to class y can be expressed as:

$$P(y|x_1, x_2, ..., x_n) = \frac{P(y)P(x_1, x_2, ..., x_n|y)}{P(x_1, x_2, ..., x_n)}$$

The "naïve" assumption states that each feature is conditionally independent from every other features:

$$P(x_i|y, x_1, x_2, ..., x_{i-1}, x_{i+1}, ..., x_n) = P(x_i|y)$$

[10] http://sentic.net/
[11] https://github.com/cjhutto/vaderSentiment

Now, the original probability is simplified:

$$P(y|x_1, x_2, ..., x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1, x_2, ..., x_n)}$$

The classification rule is:

$$P(y|x_1, x_2, ..., x_n) \propto P(y)\prod_{i=1}^{n} P(x_i|y)$$
$$\Rightarrow \hat{y} = \underset{y}{\operatorname{argmax}}\, P(y)\prod_{i=1}^{n} P(x_i|y)$$

where $\hat{y}$ denotes the prediction outcome.

3.3.2 Logistic regression

Logistic regression is regression model that is widely used in binary classification task (the dependent variable takes only two values). Logistic regression is one of linear models where the response (dependent) variable is estimated using linear function:

$$y = w_0 + w_1x_1 + ... + w_nx_n$$

Weights and features can be represented in vector form $\mathbf{w} = (w_1, w_2, ..., w_n)$ and $\mathbf{x} = (x_1, x_2, ..., x_n)$. The probability function estimating likelihood of y belong to positive class ($y = 1$) is:

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + exp(-\mathbf{w}^\intercal\mathbf{x})} = \sigma(\mathbf{w}^\intercal\mathbf{x})$$

where $\sigma(z) = \frac{1}{1+exp(-z)}$ is called *sigmoid* or *logistic* function.

3.3.3 Support vector machines

Support vector machines is a set of supervised learning models used for classification and regression task. In general, SVM try to establish one or more hyperplane to sepa-

rate data points and assign them to target classes. Suppose we have a training dataset $(\vec{x}_1, y_1), (\vec{x}_2, y_2), ..., (\vec{x}_n, y_n)$ where y either take values -1 or 1 (binary case) and $\vec{x}_i$ is a feature vector. A hyperplane can be represented as:

$$\vec{w}.\vec{x} - b = 0$$

where $\vec{w}$ is the normal vector to the hyperplane. If data points are linearly separable, we want to maximum distance between two classes. We define two hyperplanes $\vec{w}.\vec{x} - b = 1$ and $\vec{w}.\vec{x} - b = -1$. The distance between them is $\frac{2}{||\mathbf{w}||}$ where $||\mathbf{w}|| = \sqrt{w_1, w_2, ..., w_2}$.
Our objective is to maximize

$$\min_{i=1,...,n} = \frac{y_i(\vec{\mathbf{w}}.\vec{\mathbf{x}} + b)}{||\mathbf{w}||}$$

as a function of **w** and b.

Chapter 4
Methodology

4.1 Introduction

4.1.1 Research question

Research objective is to explore the feasibility of using NLP and ML to identify at-risk individuals who might develop suicidal ideation in Internet environment. More concretely, the research questions are as follows:

- How can we improve efficiency of existing work on identifying people with mental health problems who actually develop suicidal ideation on online forum using natural language processing and machine learning?
- What machine learning techniques are appropriate for such classification task on a large scale?

The forum we choose for this experiment is Reddit. In literature, existing work [14] already examined suicide support subforum of Reddit and built classifier to pinpoint at-risk users. Although the classifier performed quite well in the research, it performed only on a small sample of entire dataset thus robustness of the classifier in realistic scenario is unknown. We are going to borrow the approach of [14] to investigate feasibility of such classifier on a larger population. We will build classifiers with comparable performance and apply it to large dataset collected.

4.1.2 Overview of Reddit

Reddit is a prominent social networking sites with an extremely large user base. Founded in 2005, Reddit has 234 million active users and attracts 542 million visitors per month as of 2017. In term of website ranking based on number of visitor, Reddit ranked 4^{th} in U.S and 9^{th} in the world [1]. Reddit aggregates news from many sources including itself, allowing users to discuss by commenting on the original post which is hosted on the sites as a thread. Users can interact with a submission by voting the post/comment up or down to express whether they like or dislike the post/comment. The score of a post (i.e. difference between the number of upvote and downvote for the post) determine the position of the post on the page. The more score a post got, the higher position of the post. This feature allows high scored posts to be highly visible to visitors due to being on top of the page. Figure 4.1 shows a screenshot of Reddit front page at the time of writing. On front page, posts have gathered high number of upvote in short period of time which can be up to tens of thousands upvotes in several hours. Reddit is organized by areas of interest. Each area of interest or topic has each own space called ”subreddit” for posting and discussion, much like a subforum. Convention for typing name of a subreddit is adding ”r/” before the name of subreddit to indicate direct link to that subreddit (e.g. *r/movies* refers to Movies subreddit - https://www.reddit.com/r/movies/). Every subreddit is moderated by moderator to ensure posts and comments follow guideline and do not digress from the topic. As moderation entails substantial monitoring, a moderator is either a human or an automated bot. To create a new subreddit, a Reddit user (also known as redditor) must meet specific criteria: his account must be at least 30 days old and his account must accumulate enough post ”karma”. ”Karma” reflects how much a redditor contribute to Reddit community, ”karma” can be earned by getting upvotes for external link posted or upvotes for comments. The exact number of karma required for creating new subreddit is unknown to normal users, only admins know. Reddit is very diverse as subreddit topics include various fields ranging from general themes such as science, sport, news, movies, food to specific themes namely r/Moviesinthemaking, r/harrypotter, r/PoliticalHumor.

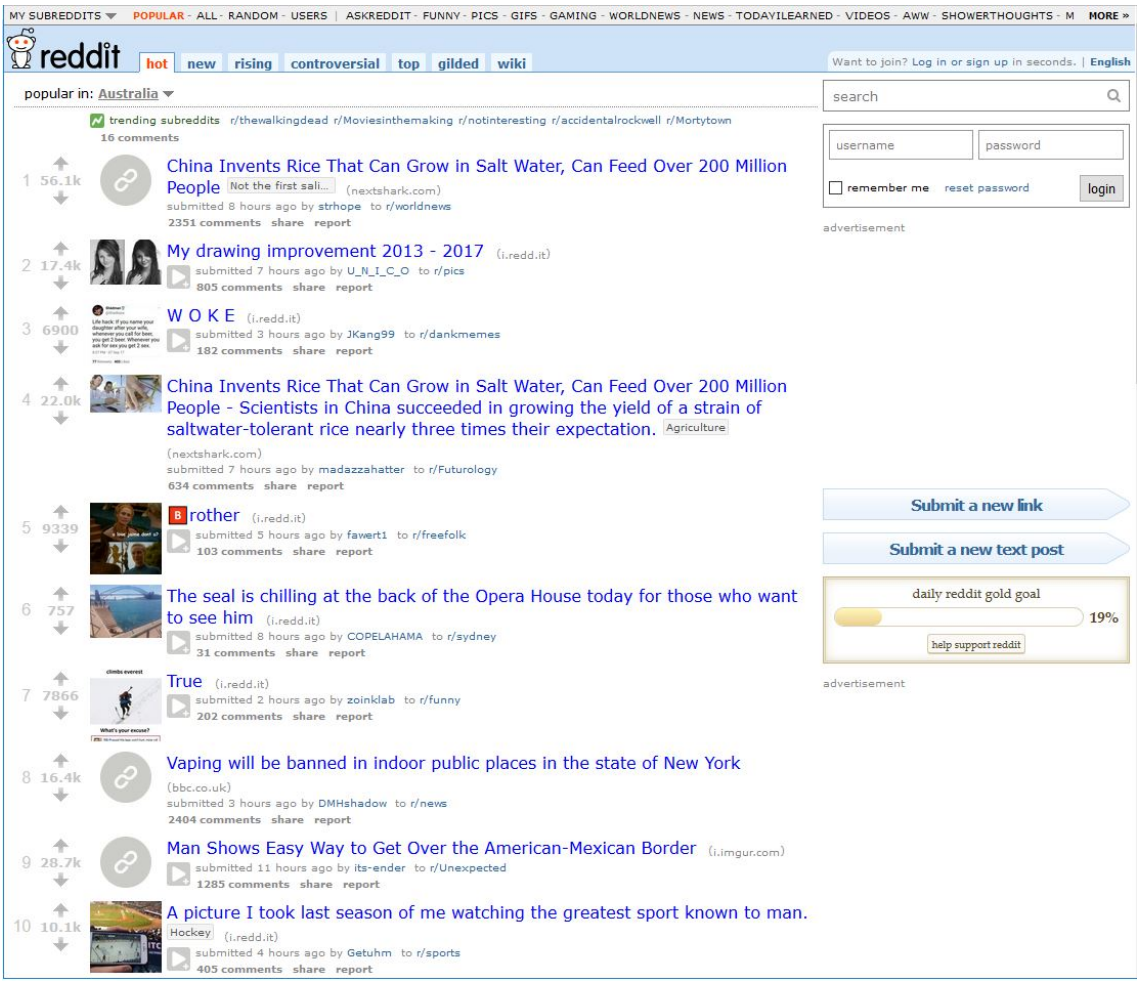

Figure 4.1: User interface of Reddit front page.

As mentioned in the previous section, Reddit provides a free API[1] for any developers who want to build applications that need to crawl posts and comments on Reddit. There are a variety of scripts that users developed using Reddit API: auto reply bot, auto crawler, auto unit converter, image resizing bot, external link thumbnail displaying bot, message reminder, etc. Most of these bots operate by the principle: getting posts and comments in real time, processing according to programmed procedure, producing output which usually is a comment. For example, a grammar correcting bot crawl all posts/comments in real time, whenever it detects a grammar mistake (say "could of") in a post, it replies to that post explaining and correcting to "could have". Another use-

[1]https://www.reddit.com/dev/api/

ful example is unit converter which convert unit from imperial system to metric system since most posts from American use imperial unit while most non-U.S readers are familiar with metric unit.

4.2 Experimental design

4.2.1 Data collection

In this experiment, we are using Google BigQuery[2] as an alternative to Reddit API for crawling posts and comments. BigQuery is an IaaS (Infrastructure as a Service) site that interoperate with Google Cloud Storage[3] to provide interactive analysis of big data warehouses. Its features include managing data, query data, integration to other systems and access control. On BigQuery, data is managed by CRUD operations (Create, Read, Update, Delete) or is imported from Google Storage. The main language used for querying tables is standard SQL or its dialect. Result of a query can be downloaded in CSV (comma-separated values) or JSON (JavaScript Object Notation) format. Owner of the dataset is able to choose options for sharing data with a selected user, with groups or publicly.

A dataset of all posts and comments from Reddit is uploaded and update monthly on BigQuery. The tables contain posts and comments of Reddit from 2006 to the latest month can be found at `https://bigquery.cloud.google.com/dataset/fh-bigquery:reddit_posts`.

We queried the dataset to get posts and comments of 14 MH subreddits and SW subreddit (r/SuicideWatch) from February 11th, 2014 to November 11th, 2014. The list of MH subreddits includes r/mentalhealth, r/bipolarreddit, r/depression, r/ptsd, r/hardshipmates, r/survivorsofabuse, r/StopSelfHarm, r/BPD, r/psychoticreddit, r/traumatoolbox, r/EatingDisorders, r/rapecounseling. The content in these MH subreddits have been examined and verified to be relevant to mental health discussion theme [37]. The data is of JSON format which

[2]https://cloud.google.com/bigquery/
[3]https://cloud.google.com/storage/

	MH subreddits	SW subreddit
Number of posts	60556	20025
Number of comments	278100	131330
Average length of posts	251	250.3
Median length of posts	164	165
Average length of comments	66.7	57.2
Median length of comments	36	26

Table 4.1: Overview of the dataset.

is identical to built-in data structure *dictionary* of Python. The dataset on BigQuery separate posts and comments into different tables, so we have two JSON files as a result: one consists of posts and the other consists of comments. The total number of posts and comments are 80581 and 409430 respectively. There are 34,549 unique authors who have posted in all these 15 subreddits. Table 4.1 shows breakdown statistics of the dataset.

Table 4.2 shows common words used among both MHs and SW. Most of those words are common in both dataset and relatively equal in rank such as ”like”, ”people”, ”think”, ”know”, ”one”. Those kinds of words are pretty general and do not reveal much about content of the text. However, several words characterize the groups namely ”depression”, ”help”, ”years”. The word ”depression” occurs at high frequency in MHs indicate discussion among community while ”help”, ”years” may infer calls for help of at-risk individuals or struggle over a long period.

4.2.2 Constructing user classes

The time span of nine months in our data is split into two periods: the first six months (February 11^{th}, 2014 - August 11^{th}, 2014) and the last three months (August 12^{th}, 2014 - November 11^{th}, 2014). The 'control' group in this experiment is the group of users who do not have suicidal thoughts even though they may encounter problems with mental health. On the other hand, at-risk individuals are defined as people who are challenged by mental disorders and later develop suicidal ideation. In the context of online forum, we can only observe posting activity so we identify users who only posted on MH subreddits during the first six months but never posted on SW subreddit during the last three

MH subreddits		SW subreddit	
Word	Count	Word	Count
like	87222	like	25290
feel	75385	want	22485
know	64496	know	21755
get	55209	life	21213
want	52596	feel	20429
time	49249	get	17066
really	48359	even	15100
life	47545	would	14854
people	45693	time	14826
even	43755	people	14607
would	40328	one	14156
one	39295	really	14015
friends	33830	going	11658
think	32685	never	11482
go	32661	think	11233
depression	32379	friends	11120
going	32089	go	11009
things	31335	much	10039
never	31227	years	9995
much	28790	help	9810

Table 4.2: Top 20 common words in each dataset (stopwords excluded).

months as 'control' group (hereafter referred to as MH). Similarly, users who only posted on MH subreddits during the first six months and posted on SW during the last three months are defined as at-risk individual (hereafter referred to as MH→SW). By "posted" we mean submit a post, that translates to only users that start a discussion are included. This criterion excludes users who only gave commentary and were not subject of the discussion.

The procedure returned 465 users of MH→SW an 24,761 users of MH. To create two balanced classes, we randomly sampled 465 users from 24,761 users. After that, we extracted posts and comments from these users. We discarded all the posts and comments that do not contain text (indicating external links, images or videos) or the content is deleted. Each entry of post gathered contained the title, text, score, number of comments of that post. We had 1,484 posts and 6,115 comments from 465 MH→SW user and 584 posts, 1,988 comments from 465 MH users. There are some cases we may miss at-risk

individuals: users who posted in MH subreddits before February 11th, 2014 or posted in SW subreddit after November 11th, 2014. We assumed the number of such cases is quite small and did not consider to expand the period. Another point to note is all the posts and comments are from 15 mentioned subreddits, users may have posted somewhere else outside these subforums.

4.2.3 Feature selection

In this section, we present the process of features selection for prediction framework. Our aim is to build a classifier that performs comparable to the classifier built in [14]. We chose some features like existing work but some features get removed because it did not improve performance or we cannot recreate such features due to lack of low-level implementation details. There are four sets of features namely *interpersonal awareness, linguistic structure, interaction, sentiment*. Each set has several different variables related to the corresponding category. All these measures are averaged across all posts and comments of a user.

- *Interpersonal awareness:* this set includes percentage of first person plural (i, me, my, mine), first person singular (we, us, our, ours), second person pronouns (you, your, yours) and third person pronouns (he, she, it, him, her, his, hers, its, they, them, their, theirs). Past work [12] used this measure to quantify self-awareness, collective attention and social interaction of an individual.

- *Linguistic structure:* this measure comprises percentage of nouns, verbs, adverbs and pronouns in submissions; number of 'difficult' words; automated readability index; Fleisch reading ease. The 'difficult' words here are words that have more than three syllables. Fleisch reading ease is a scale to gauge the difficulty of documents in term of understandability. This Fleisch reading ease score is on scale 0-100. A high score indicates the text is easy to read and vice versa. Automated readability index has similar use like Fleisch reading ease but is computed by different formula and the scaled score is reverse to it: the lower score, the easier to read. The nouns, verbs, adverbs and pronouns are labelled by NLTK POS tagger with "Universal"

tagset, the remaining variables are calculated using textstat[4] Python package. These variables quantify overall writing structure of a user, suggesting his cognitive state.

- *Interaction:* variables in this set is associated with metadata of posts and comments namely length of posts/comments submitted, length of titles of posts authored, score for posts/comments submitted, number of comments to post submitted. These kinds of variables show how active a user is on Reddit and how much support that user received.

- *Sentiment:* this set utilizes sentiment analysis on each post or comment. Each post was assigned four types of polarity score: pos (positive), neg (negative), neu (neutral) and compound score. The scores are calculated using VADER [26] introduced in previous section.

- *Content:* this set includes 90 tokens (unigrams or bigrams) reported in [14] that distinguished between MH→SW. Among these tokens, 70 tokens associated with increase in the likelihood of posting in SW in the future. Whereas the remaining 20 tokens decrease the likelihood of being in MH→SW. All reported tokens either have highest or lowest treatment effect. This features have binary values: if a user used a token, variables corresponding to that token is valued 1 or 0 otherwise.

- *Combined*: this set is combination of all five sets above.

Table 4.3 and Table 4.4 derived from [14] show 90 tokens mentioned above. The number of token users, population coverage and χ^2 statistics have been omitted. All these tokens are statistically significant as p-value <0.001, ordered in descending order of z-score (positive treatment effect tokens) or treatment effect (negative treatment effect tokens). This statistic is acquired by using stratified propensity score matching analysis on tokens used by more than 10 users of MHs group which are more than 11,000 tokens.
As can be seen in the tables, the token significantly increases the likelihood of posting suicidal ideation in the future by large percentage 17-54%. Some tokens did not reveal informational content (e.g. "can't", "friends") and have lower treatment effect, below 20%.

[4]https://pypi.python.org/pypi/textstat/

token	treat effect	z	token	treat effect	z
depression	0.3	8.04	money_i	0.52	5.89
useless	0.51	7.05	out_as	0.53	5.89
suicide	0.32	6.66	this_happened	0.51	5.89
anxiety	0.24	6.56	this_world	0.5	5.88
suicidal	0.34	6.56	over_i	0.51	5.86
i_almost	0.52	6.44	still_a	0.51	5.85
and_an	0.51	6.4	off_a	0.51	5.85
medicine	0.52	6.38	loneliness	0.5	5.84
unless_i	0.53	6.36	class_and	0.52	5.84
hug	0.52	6.36	alone_i	0.34	5.84
they_didn	0.51	6.33	am_the	0.54	5.82
take_me	0.52	6.32	care_i	0.52	5.79
and_give	0.51	6.23	giving_me	0.51	5.79
shirt	0.53	6.22	they_get	0.51	5.79
happy_i	0.52	6.21	capable	0.49	5.79
i_talk	0.51	6.2	keep_in	0.52	5.77
locked	0.51	6.17	the_amount	0.52	5.76
can_t	0.22	6.14	hate_it	0.48	5.76
people_on	0.5	6.12	socially	0.51	5.75
do_for	0.52	6.11	increase	0.51	5.75
problems_i	0.51	6.08	t_keep	0.52	5.75
anyone_i	0.51	6.07	just_in	0.51	5.73
thoughts_and	0.53	6.07	picked_up	0.5	5.73
ve_started	0.52	6.04	t_help	0.28	5.71
stuck_in	0.5	6	no_real	0.5	5.71
no_friends	0.51	6	alone	0.19	5.71
but_only	0.51	5.98	existing	0.49	5.71
have_nothing	0.51	5.98	an_idiot	0.51	5.71
require	0.52	5.97	just_trying	0.52	5.7
would_get	0.5	5.96	t_deserve	0.51	5.7
but_can	0.52	5.95	depressive	0.52	5.69
been_there	0.51	5.95	can_give	0.51	5.69
who_don	0.51	5.92	friends	0.17	5.69
world_of	0.52	5.92	end_i	0.51	5.69
kills	0.53	5.9	existence	0.5	5.68

Table 4.3: (Statistically significant) treatment tokens obtained via propensity score matching that contribute to *increased* change in likelihood of posting in SW [14].

In contrast, many tokens gave cues about users' stories and their mental state (e.g. "suicidal", "useless", "depressive", "medicine", "no friends", "locked"). In general, extensive use of such tokens indicate underlying psychological distress, pessimistic perspective,

token	treat effect	z	token	treat effect	z
captain	-0.6	-4	straight_up	-0.56	-3.82
difference	-0.59	-4.47	preferred	-0.56	-3.71
the_trip	-0.57	-3.76	awesome_i	-0.56	-3.68
intimate	-0.57	-3.73	s_at	-0.55	-4.83
to_in	-0.56	-4.92	stated	-0.55	-4.8
too_hard	-0.56	-4.4	slight	-0.55	-4.61
suspect	-0.56	-4.4	and_enjoy	-0.55	-4.44
always_a	-0.56	-4.15	gotten_to	-0.55	-4.35
be_working	-0.56	-4.12	it_work	-0.55	-4.22
keep_your	-0.56	-3.82	came_from	-0.55	-4.21

Table 4.4: (Statistically significant) treatment tokens obtained via propensity score matching that contribute to *decreased* change in likelihood of posting in SW [14].

possible severe illness that need medication of users.

On the other hand, some tokens decrease the likelihood of posting in SW subreddit. Tokens such as ”awesome i”, ”and enjoy”, ”the trip”, ”it work” show significant more positive attitude than tokens mentioned above. Use of tokens like that signals sign of relief, recovery from hardship, enjoyment in life from users. Treatment effect of these tokens is very noteworthy, around -55% to -60%.

According to [15], stratified propensity score matching may encounter problems of imbalance in strata. That means the use of a token may have different effects on different users. Some users saw significant change in probability of joining MH→SW while some do not. There are tokens that both increase and decrease the likelihood of posting in SW so we do not incorporate such token into our features.

4.2.4 Experimental settings

We devise two settings for this experiment:

(1) **Theoretical setting:** in this setting, we runs classifier in two balanced classes with equal sample size. That mean there are 465 MH users and 465 MH→SW users in the dataset. The purpose of this setting is to compare with classification framework in [14]. If the performance is comparable, we can apply the same classifier in the other setting.

		True condition	
		MH→SW	MH
Prediction outcome	MH→SW	True Positive (TP)	False Positive (FP)
	MH	False Negative (FN)	True Negative (TN)

Table 4.5: Confusion matrix for the classification task.

(2) **Realistic setting:** in reality, the number of people challenged by mental health problems but do not consider suicide is different from people with similar problem and eventually develop suicidal ideation. As we can see in our data, MH users outnumbers MH→SW users by a tremendous amount. The total number of MH→SW users is only 1.84% of total users in final dataset. The final dataset for this setting contained 465 MH→SW users and 24,761 MH users. We are going to classify MH→SW users from MH users by using the same classification framework in theoretical setting.
The ML algorithms used in this experiments are Logistic regression, Naïve Bayes and Support Vector Machine. We performed 10-fold cross validation on the final dataset and run the model 10 times to mitigate randomness of cross validation. The evaluation metrics will be the standard metrics in most of supervised ML tasks: *precision, recall* and *F-1 measures*. Table 4.5 shows confusion matrix for this task.

Precision measures the rate of our correct prediction of MH→SW users over the total users we have predicted to be MH→SW.

$$Precision = \frac{TP}{TP + FP}$$

Recall measures the rate of correct prediction of MH→SW over the total number of MH→SW users.

$$Recall = \frac{TP}{TP + FN}$$

F-1 measure is the harmonic mean of precision and recall.

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

4.3 Result

4.3.1 Differences in linguistic, interpersonal, interaction and sentiment features

	MH	MH→SW	z	p
Interaction				
Post length	267.6	265.9	0.09	-
Title length	7.88	8.31	-1.10	-
Post score	6.0	5.7	0.24	-
# Comments received	3.99	5.49	-2.15	*
# Comments submitted	3.96	13.1	-6.34	*
Comment length	46.5	63.4	-2.83	*
Comment score	1.07	1.51	-3.86	*
Linguistic structure				
# Difficult words	55.0	52.6	0.69	-
Fleisch reading ease	70.7	75.5	-2.71	*
Automated reading index	7.04	7.14	-0.2	-
% Pronouns	0.07	0.08	-2.05	*
% Nouns	0.21	0.19	2.96	*
% Verbs	0.22	0.25	-4.72	*
% Adverbs	0.08	0.08	-0.91	-
Interpersonal awareness				
% 1st person singular	0.09	0.10	-5.88	*
% 1st person plural	0.0026	0.0024	0.65	-
% 2nd person pronouns	0.005	0.004	1.23	-
% 3rd person pronouns	0.032	0.031	0.41	-
Sentiment analysis				
Negative score	0.13	0.15	-5.1	*
Neutral score	0.68	0.69	-0.65	-
Positive score	0.115	0.114	-0.31	-
Compound score	-0.19	-0.36	3.98	*

Table 4.6: Theoretical setting: summary of feature sets for MH users class and MH→SW users class.

Table 4.6 show means of measures of two user classes. Half of these measures characterize differences between MH and MH→SW. We performed z-test on two samples and show statistical significance at $p = 0.05$ level. An asterisk "*" denotes the difference between two populations is statistically significant and a dash "-" denotes otherwise. Each of the categories namely structure of language, social awareness, sentiment analysis and

forum interaction have variables that are statistically significant. Thus, we can observe some differences between two classes and they may suggest some psychological phenomena.

In the set of features related to forum interaction, we see differences in term of comments between MH users and MH→SW users. MH→SW users submitted far more comments than MH users (z = -6.34). MH→SW users also received more comments from other in their posts (z = -2.15). Note that all posts and comments are in MH and SW subreddit so variables that quantify commentary of MH→SW users may be indicative of active engagement in mental health communities of MH→SW users. Both length of comments authored by MH→SW users and score of those comments are higher than that of MH users. Such differences may imply the concerns of MH→SW in mental health area. Prior work [14] show the opposite that MH→SW display symptoms of social isolation. This may be true in general context where dataset in previous work also included posts from different subreddits. However, this observation suggests that MH→SW users are usually active in mental health related communities, not other areas of interest.

Linguistic structure features show differences in readability; use of pronouns, nouns, and verbs. Automated reading index could not capture discrepancy of understandability of users from two classes while Fleisch reading ease reflected lower coherence in text from MH→SW users (z = -2.71). Lower use of nouns (z = 2.96) reveals little concerns for topics [10]and higher use of verbs (z = -4.72) show focus on actions which may have connection to self-disclosure on social media [23]. Surprisingly, the number of 'difficult' words (words have more than 3 syllables) shows no difference despite poorer linguistic structure of MH→SW users.

First person singular pronouns are used more by MH→SW users (z = -5.88). This reflects higher self-focus and have tendency to express personal thoughts more than MH users. This observation is consistent with suggestion that MH→SW users may involve in self-disclosure posting activity more than normal.

Sentiment analysis show more negative attitude and pessimistic outlook of MH→SW users. Both negative polarity score and compound score of MH→SW users' posts and comments are lower than MH users' submissions (z = -5.1 and z = 3.98 respectively).

	MH	MH→SW	z	p
Interaction				
Post length	282.5	265.9	1.11	-
Title length	8.18	8.31	-0.46	-
Post score	6.2	5.7	0.4	-
# Comments received	4.18	5.49	-3.4	*
# Comments submitted	4.3	13.1	-11.5	*
Comment length	41.9	63.4	-6.9	*
Comment score	1.07	1.51	-5.19	*
Linguistic structure				
# Difficult words	57.1	52.6	1.67	-
Fleisch reading ease	69.7	75.5	-2.76	*
Automated reading index	7.55	7.14	0.66	-
% Pronouns	0.07	0.08	-2.17	*
% Nouns	0.21	0.19	2.74	*
% Verbs	0.23	0.25	-5.26	*
% Adverbs	0.084	0.082	-1.4	-
Interpersonal awareness				
% 1st person singular	0.09	0.10	-6.22	*
% 1st person plural	0.0025	0.0024	0.38	-
% 2nd person pronouns	0.005	0.004	1.39	-
% 3rd person pronouns	0.031	0.031	-0.37	-
Sentiment analysis				
Negative score	0.13	0.15	-7.8	*
Neutral score	0.68	0.69	-1.24	-
Positive score	0.114	0.115	-1.29	-
Compound score	-0.18	-0.36	5.2	*

Table 4.7: Realistic setting : summary of feature sets for MH users class and MH→SW users class.

This may link to emotional distress or hardship confrontation.

Table 4.7 shows results of z-test between means of variables of the two user classes in realistic setting. The column for MH→SW users is identical to that of the previous table. There is no change in column of p as well, meaning all statisticcally significant variables in theoretical setting still hold their validity in this z-test. This table confirmed representability of MH random sample in this experiment.

4.3.2 Classification results

Results of the classifiers in theoretical setting

	Logistic Regression			Naïve Bayes			SVM		
	P	R	F1	P	R	F1	P	R	F1
IA	0.56	0.65	0.60	0.54	0.82	0.65	0.54	0.69	0.60
LS	0.53	0.70	0.60	0.51	0.92	0.66	0.59	0.54	0.56
I	0.66	0.61	0.63	0.57	0.78	0.63	0.52	**0.94**	0.67
S	0.55	0.61	0.58	0.58	0.82	0.68	0.65	0.60	0.62
C	0.71	0.60	0.65	**0.77**	0.32	0.45	0.70	0.66	0.67
Com	0.76	0.66	**0.70**	0.70	0.46	0.56	0.56	0.79	0.65

Table 4.8: Theoretical setting: classification results of SVM, Naïve Bayes, Logistic regression on six sets of features.

We report classification results of balanced classes setting in Table 4.8. There are abbreviations in the table: IA = Interpersonal Awareness, LS = Linguistic Structure, I = Interaction, S = Sentiment analysis, C = Content features, Com = Combined features. All the results reported are averaged after performing 10-fold cross validation. Overall, the classifier logistic regression on combined set of features give the best performance (F1 = 0.70). In term of precision, Naïve Bayes give the highest precision on content features (P = 0.77) followed by logistic regression performed on combined features (P = 0.76). The highest recall among all classifier is achieved by SVM on interaction features (R = 0.94), slightly higher than Naïve Bayes with linguistic features. The Naïve Bayes classifier displayed the nature of tradeoff in machine learning clearly: when it classifies using the first four sets of features, average recall obtained is very high (0.78- 0.92) while average precision remains relatively low (0.51 - 0.58); the other two sets produced relatively high precision (0.77, 0.70) but recall results are below average (0.32, 0.46). The performance of logistic regression is quite consistent and comparable to one in [14]. Our best results for logistic regression with combined dataset in one run are P = 0.87, R = 0.78 and F1 = 0.82. We are going to proceed to use these classifiers in the realistic setting.

Results of the classifiers in realistic setting

Predicted / Actual	MH→SW	MH
MH→SW	0	0
MH	47	2476
Accuracy	98.1%	
Precision	0	
Recall	0	
F-1	0	

Table 4.9: Realistic setting: SVM classifier performance

Predicted / Actual	MH→SW	MH
MH→SW	1	4
MH	46	2473
Accuracy	98%	
Precision	0.2	
Recall	0.02	
F-1	0.03	

Table 4.10: Realistic setting: Logistic regression classifier performance

Predicted / Actual	MH→SW	MH
MH→SW	29	564
MH	18	1912
Accuracy	76.9%	
Precision	0.04	
Recall	0.61	
F-1	0.07	

Table 4.11: Realistic setting: Naïve Bayes classifier performance

Table 4.9, Table 4.10 and Table 4.11 show the performance of the three classifiers in three single runs. Our objective is to correctly identify MH→SW users in the population so we chose the runs that have the highest numbers of true positive outcomes. Overall, the prediction outcomes are skewed heavily toward the MH class. SVM overlooked all MH→SW instances thus failed to classify any MH→SW users. Since the MH class overwhelms MH→SW class by enormous amount (47 MH→SW vs 2476 MH), the classifier still achieved high accuracy (98.1%). The logistic regression classifier returned 5

MH→SW instances but only 1 was correct, attained precision = 0.2, recall = 0.02, F1 = 0.03 and accuracy = 98%. Among three classifiers, Naïve Bayes gave the highest number of true positive instances (TP = 29) thus achieved impressive recall = 0.61. However, the number of false positive cases is far too high (FP = 564) compared to the other classifiers. As a result, precision is dragged down to 0.04, giving F1 = 0.07 which is higher than logistic regression but by no mean can be considered reasonable in ML context.

Both results of SVM and logistic regression classifiers presented above are representative of overall performance on all sets of features. That means no matter what set of features is chosen for the classification task, the result is pretty much the same for SVM and logistic regression classifiers. Nevertheless, this is not the case for Naïve Bayes. It gave similar results for the first four sets (linguistic structure, forum interaction, sentiment scores, interpersonal awareness) with TP = 0 and TN = 2476. The 'content' set and the combined set returned more than 17 True Positive cases each run. Features in combined set are largely from the 'content' set (90/112 features). This contribution may be attributed to similarity of performance of the classifiers on the two sets of features.

4.4 Discussion

Recently, prior works [11, 14] made attempts to predict which online social media users have high risk to have suicidal thoughts or suicide attempts. We borrowed the approach of using Reddit data to identify such users. As compared to classifier from [14] in the same scheme, our logistic regression classifiers gave comparable performance. Note that our dataset is smaller as we only focus on posts and comments in 15 subreddits while their dataset contained all posting history of a user regardless of topics discussed. We also have fewer features for the classification task. The methods of feature selection and dataset filtering reduce computation required for the task. This reduction is meaningful in practical context where computing resources are valuable as volume of data generated by Internet users is gigantic.

In the feature selection step, polarity scores based on sentiment analysis of each posts and

comments were incorporated into the model. Hypothesis testing show statistical significance of mean differences of negative scores as compound scores between the MH group and MH→SW group. These scores reflect negative attitude, decreased mental well-being of individuals who belong to at-risk group. We also saw difference between number of comments authored, average length of comment and score for those comments are higher for MH→SW users. This difference indicates more active engagement in mental health related communities of at-risk individuals which is counter-intuitive as existing literature [10] suggested detachment from social interaction of at-risk individuals. Perhaps topics of concerns of those people are narrowed down when they face emotional distress, anxiety or other psychological disorders.

We explored the feasibility of applying those ML classifiers in realistic context where true at-risk individuals only occupy a small proportion in entire population (1.8% in the experiment). Performance of the classifiers dropped completely in this setting as both SVM and logistic regression failed to predict true positive instances which are 0 in most runs. Although Naïve Bayes returned a reasonable number of true positive cases, false positive became tradeoff with over 550 false positives in the result. Overfitting rendered SVM and logistic regression less pragmatic in tasks that have huge imbalance between classes in term of raw count. If a classifier with similar performance of Naïve Bayes is adopted for such task, additional screening is certainly needed to reduce problems caused by high number of false positive instances.

The experiment addressed the research questions. Firstly, we built three classifiers and sets of features that contained fewer features than that of the existing work. De Choudhury and her colleagues [14] incorporated all unigrams, bigrams of the tens of thousands of posts/comments and their relative frequencies as 'content' features. The dataset used in our experiment is also smaller in size due to posts and comments limited in particular subreddits. Nevertheless, the performance did not decrease, suggesting high efficiency in step of feature selection. Secondly, the experiment setting without class balance in dataset examined robustness of said classifiers. It shows that generalizability of this approach is quite poor if problem of overfitting is not addressed, possibly by adding more features, applying more regularization techniques. Naïve Bayes classifier did not suffer

from overfitting but its precision is not desirable in practice.

There are several limitations and caveats in this study. The number of subreddits examined is quite small considering Reddit has many other prominent mental health communities such as eating disorder, addiction, schizophrenia, etc. These kinds of mental illness are possibly responsible for many cases of suicide in general. Since Reddit does not restrict user to own only one account, our dataset may be contaminated by people who posted under multiple accounts or throwaway account. The experiment only utilized three mainstream classifiers while ML offers many more algorithms or approaches such as Decision tree, Ensemble meta classifier, Neural networks, etc. Furthermore, only posting activity is covered in this study as predictor variables but mental well-being of one person affected by many other factors like financial stability, relationship, physical illness. Such factors and other elements outside of Reddit cannot be observed with this approach. In some unusual cases, the original poster may not be the one that has suicidal thoughts. Instead, they do ask for advices from community but for their acquaintances, their friends or their loved one. That cases and similar ones may flag false predisposition to suicidal thoughts and need to exclude from analysis.

Chapter 5
Conclusion

5.1 Summary of chapters

Chapter 1 gives readers overview of emergence of social networking sites and how it opens new research directions in mental health field. An examples of conventional mental health intervention programs shifting to online therapies is use of school-based programs has decreased since web-based applications offer more flexible choices in providing information without demanding substantial resources like programs in school setting.

Brief overview of recent studies in online mental health are presented. From monitoring mental well-being of individuals to analyzing search queries data of an entire geographical region, many studies with different scales facilitate the development of mental health support apps. Several potential sources of data for research, which are chat-based interventions (mobile apps, text message service, web chat) and online discussion forums, are listed in the following section.

Computer Science have paved way for automated large-scale data analysis, reducing human workload in many tasks. The interdisciplinary field of study between Computer Science and Psychology mainly focused on utilizing NLP and ML to solve problems that are traditionally done by human or to observe behaviour in large population. The main theme of this thesis is suicide and we tried to delve into attempts to analyze suicidal people in several aspects. The main purpose of this field is to use machine prediction to aid suicide prevention programs.

Chapter 2 presents a detailed literature review on application of Computer Science in suicide research. Firstly, the chapter informs readers of four types and process of suicide. Secondly, this chapter introduces the 2011 i2b2 NLP Challenge which asked the participants to assign emotions to each sentence of suicide notes. There are several participating teams published papers explaining their approaches. We summarized and compared these approaches plus reported results of each team.
In the third place, we present the main trends of analyzing online social network posts for predicting suicide numbers and spotting at-risk individuals. The reported number of suicide cases in South Korea is matched with machine prediction based on dysphoria blog posts, economical and meteorological variables with a high accuracy of 0.83. Another study observed correlation between number of suicidal Twitter posts and real numbers of suicide reported.
Contagion effect of celebrity suicide, a well-known effect in psychology literature, has been inspected on Reddit and relevant evidences was confirmed. Researchers made efforts to develop systems that aid human in suicide prevention. Most of these attempts is to try to pinpoint at-risk individuals on online social media platforms. Twitter and Reddit are two prominent sites that have been frequently serve as sources of data in these attempts. Four studies using data from the two websites are introduced and one of them established the approach used in this thesis.

Chapter 3 describes tools and methods used for the experiment of this thesis. We made use of NLTK, scikit-learn, VADER. Besides, we listed some popular NLP techniques namely tokenization, stopwords removal and part of speech tagging.

Chapter 4 presents the research questions and the experiment trying to address those questions. We chose Reddit as our source of data and introduced general features of it which are "karma" (serve as contribution point), API, post voting. Organizational structure of Reddit allows redditors to match their interest to various subreddits and it facilitates data collection for research. Researchers can access data for free by using the official Reddit's API to crawl posts and comments corresponding to topics of interest. An

alternative way to collect data is using Google BigQuery as all posts and comments of Reddit are shared on it monthly plus it is faster and easier to use than the API due to power of cloud computing platform of Google and comprehensive standard SQL queries interface.

Next, we elaborate how data is filter through criteria to form a big dataset containing posts and comments of relevant mental health subreddits over two specific periods of time. To prepare for the classification task, we have to construct two classes of users: redditors who do not post in the suicide support subreddit SW and redditors who do so after previously posted in mental health discussion subreddits. Feature selection is one of the step distinguishing our work from the prior work on which this approach based. We introduced a new set of features contained polarity score from sentiment analysis of the posts and comments. Moreover, we dropped the idea of using all unigrams, bigrams in posts as features. Instead, a set of tokens which is verified to have high treatment effect (decrease/increase the likelihood of posting in SW) is used as substitute.

We specified two settings of the experiment: one is theoretical setting where two user classes are balanced (by sampling MH group) and the other is realistic setting with huge class imbalance. Evaluation metrics for this task is standard Precision, Recall, F1 because our binary classification task is typical supervised learning task. ML algorithms for this task are SVM, Naïve Bayes and Logistic regression, performing 10-fold cross validation. We reported results of statistical testing of mean differences between two populations and noticed several patterns in posting activity of MH→SW users. In general, MH→SW users show poor linguistic ability reflected by readability index, high self-focus language just like works in literature suggested. In addition, we also realized that MH→SW users are posting in mental health subreddit as well as expressing more negative, pessimistic language.

The results of the classification task are reported. In theoretical setting, Logistic regression classifier gave the best result overall with highest averaged F1 = 0.70; the precision and recall are high as well (P = 0.76, R = 0.66). SVM classifier generally produced low precision (0.54 - 0.70) but high recall (0.60 - 0.94). Naïve Bayes classifier performed similar to SVM with the first four sets of features but reverse around on 'content' and combined

sets. In realistic setting, Logistic regression and SVM classifiers failed to predict true positive users as all prediction outcomes went to MH users. In contrast, Naïve Bayes classifier returned a good number of true positives but even greater number of false positive cases. Finally, we discussed the results obtained, their meaning in suicide research context and pointed out limitations of this study.

5.2 Contributions of the experiment

There are three main contributions of this experiment:

- We built ML classifiers that used fewer features than existing work in literature but still produced state-of-the-art performance in the task of identifying people challenged by mental health problems who actually develop suicidal ideation on online forum.
- We applied the same ML classifier in a more realistic scenario to see whether these classifiers are robust enough to act as a preliminary screening step for suicide risk in online forums or not. Out of three classifiers, Naïve Bayes seemed to be the most applicable but too many false positive instances remain a problem to be solved in the future.
- We observed that although at-risk people are less socially active online, they seemed to have active engagement in mental health communities. Furthermore, their language express more negative, pessimistic outlook and that may imply cognitive markers of suicidal ideation.

5.3 Future directions

The application of NLP and ML to suicide prevention is still in its infancy. We believe that there are many ways to improve current approaches of predicting suicidal behaviour

among online communities. For instance, semi-supervised learning can be used to generate more training examples. The idea of incorporating more sentiment features seem promising as well. We only have polarity scores as features in our model but sentiment analysis is advancing very fast and might give intensity scores of words/sentences with high accuracy in the future. Additionally, fine-grained emotion detection is also a hot topic and certainly give useful features to similar classification tasks. If more data from other sources such as text-based or chat-based services can be retrieved easier, it will be beneficial to researchers as well considering general principle of ML "more data, more accurate". In addition, more efforts are needed to remedy the problem of misjudging posts with sarcasm or flippant references to suicide. Those kinds of posts make data noisy and as a consequence, false positives arise from such pitfall.

Provided that some screening tools are reliable enough to be implemented in online forums, there are two aspects of design to aid the intervention. Human moderators of those forums should receive notifications in real time should any user is flagged as at high risk of suicidal behaviour to take actions them deemed appropriate, be it check out users' posting activity or connect to mental health professionals calling for further support. Another aspect to consider is promoting self-help of users. Those who seek helps on webchat services usually are given psychometrical questionnaires to auto assess their mental well-being and main challenge in life, thus giving listener more information to adjust their language of support. We can design similar features in forums, showing relevant external links and information related to suicide support to flagged users. This adaptive intervention may reduce the time and effort of help seekers in times of hardship.

Bibliography

[1] "Reddit.com Site Info," 2017. [Online]. Available: https://www.alexa.com/siteinfo/reddit.com

[2] H. Almeida, A. Briand, and M.-j. Meurs, "Detecting Early Risk of Depression from Social Media User-generated Content."

[3] S. Aslam, "Twitter by the Numbers: Stats, Demographics & Fun Facts," 2017. [Online]. Available: https://www.omnicoreagency.com/twitter-statistics/

[4] S. Balani and M. De Choudhury, "Detecting and Characterizing Mental Health Related Self-Disclosure in Social Media," *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '15*, pp. 1373–1378, 2015. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2702613.2732733

[5] N. Berry, F. Lobban, M. Belousov, R. Emsley, G. Nenadic, and S. Bucci, "#WhyWeTweetMH: Understanding Why People Use Twitter to Discuss Mental Health Problems," *Journal of Medical Internet Research*, vol. 19, no. 4, p. e107, 2017. [Online]. Available: http://www.jmir.org/2017/4/e107/

[6] S. Bird and E. Loper, *Natural Language Processing with Python*, 2009. [Online]. Available: http://www.nltk.org/book/

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *The Journal of Machine Learning Research*, vol. 3, no. 3, pp. 993–1022, 2003.

[8] P. Burnap, W. Colombo, and J. Scourfield, "Machine Classification and Analysis of Suicide-Related Communication on Twitter," *Proceedings of the 26th ACM Conference on Hypertext & Social Media - HT '15*, pp. 75–84, 2015. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2700171.2791023

[9] Centers for Disease Control and Prevention (CDC), "Web-based Injury Statistics Query and Reporting System (WISQARS)," 2013. [Online]. Available: http://www.cdc.gov/injury/wisqars/index.html

[10] G. Coppersmith, M. Dredze, and C. Harman, "Quantifying Mental Health Signals in Twitter," *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pp. 51–60, 2014. [Online]. Available: http://www.aclweb.org/anthology/W/W14/W14-3207

[11] G. Coppersmith, K. Ngo, R. Leary, and A. Wood, "Exploratory Analysis of Social Media Prior to a Suicide Attempt," *Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology*, pp. 106–117, 2016. [Online]. Available: http://www.aclweb.org/anthology/W16-0311

[12] M. De Choudhury, S. Counts, and E. Horvitz, "Predicting Postpartum Changes in Emotion and Behavior via Social Media," apr 2013. [Online]. Available: https://www.microsoft.com/en-us/research/publication/predicting-postpartum-changes-emotion-behavior-via-social-media/

[13] M. De Choudhury and S. De, "Mental Health Discourse on reddit: Self-Disclosure, Social Support, and Anonymity," *Proceedings of the Eight International AAAI Conference on Weblogs and Social Media*, pp. 71–80, 2014.

[14] M. De Choudhury, E. Kiciman, M. Dredze, G. Coppersmith, and M. Kumar, "Discovering Shifts to Suicidal Ideation from Mental Health Content in Social Media," *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, pp. 2098–2110, 2016. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2858036.2858207

[15] M. De Choudhury, S. S. Sharma, T. Logar, W. Eekhout, and R. C. Nielsen, "Gender and Cross-Cultural Differences in Social Media Disclosures of Mental Illness," *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17*, pp. 353–369, 2017. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2998181.2998220

[16] B. Desmet and V. Hoste, "Emotion detection in suicide notes," *Expert Systems with Applications*, vol. 40, no. 16, pp. 6351–6358, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2013.05.050

[17] R. O. Duda, P. E. Hart, and J. Miley, "Pattern Classification," no. January 2001, 2001.

[18] É. Durkheim, *On Suicide*. New York, NY, USA: Free press, 1897.

[19] C. Fellbaum, *WordNet: An Electronic Lexical Database*, 2nd ed. Cambride, MA: MIT Press, 1998.

[20] G. Fond, A. Gaman, L. Brunel, E. Haffen, and P. M. Llorca, "Google Trends®: Ready for real-time suicide prevention or just a Zeta-Jones effect? An exploratory study," *Psychiatry Research*, vol. 228, no. 3, pp. 913–917, 2015. [Online]. Available: http://dx.doi.org/10.1016/j.psychres.2015.04.022

[21] E. C. Harris and B. Barraclough, "Suicide as an outcome for mental disorders. A meta-analysis," *British Journal of Psychiatry*, vol. 170, no. MAR., pp. 205–228, 1997.

[22] S. Hoermann, K. L. McCabe, D. N. Milne, and R. A. Calvo, "Application of Synchronous Text-Based Dialogue Systems in Mental Health Interventions: Systematic Review," *Journal of Medical Internet Research*, vol. 19, no. 8, p. e267, 2017. [Online]. Available: http://www.jmir.org/2017/8/e267/

[23] D. J. Houghton and A. N. Joinson, "Linguistic markers of secrets and sensitive self-disclosure in Twitter," *Proceedings of the Annual Hawaii International Conference on System Sciences*, pp. 3480–3489, 2012.

[24] Y.-P. Huang, T. Goh, and C. L. Liew, "Hunting Suicide Notes in Web 2.0 - Preliminary Findings," pp. 517–521, 2008.

[25] J. M. Y. Huen, E. S. Y. Lai, A. K. Y. Shum, S. W. K. So, M. K. Y. Chan, P. W. C. Wong, Y. W. Law, and P. S. F. Yip, "Evaluation of a Digital Game-Based Learning Program for Enhancing Youth Mental Health: A Structural Equation Modeling of the Program Effectiveness," *JMIR Mental Health*, vol. 3, no. 4, p. e46, 2016. [Online]. Available: http://mental.jmir.org/2016/4/e46/

[26] C. J. J. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," *Eighth International AAAI Conference on Weblogs and ...*, pp. 216–225, 2014. [Online]. Available: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8109{%}5Cnhttp://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf

[27] S. E. Hyman, "The genetics of mental illness implications for practice," *Bulletin of the World Health Organization*, vol. 78, no. 4, pp. 455–463, 2000.

[28] J. Jashinsky, S. H. Burton, C. L. Hanson, J. West, C. Giraud-Carrier, M. D. Barnes, and T. Argyle, "Tracking suicide risk factors through Twitter in the US," *Crisis*, vol. 35, no. 1, pp. 51–59, 2014.

[29] L. Kann, T. McManus, W. A. Harris, S. L. Shanklin, K. H. Flint, J. Hawkins, B. Queen, R. Lowry, E. O. Olsen, D. Chyen, L. Whittle, J. Thornton, C. Lim, Y. Yamakawa, N. Brener, and S. Zaza, "Youth Risk Behavior Surveillance United States, 2015," *MMWR. Surveillance Summaries*, vol. 65, no. 6, pp. 1–174, jun 2016. [Online]. Available: http://www.cdc.gov/mmwr/volumes/65/ss/ss6506a1.htm

[30] A. E. Kazdin, H. C. Kraemer, R. C. Kessler, D. J. Kupfer, and D. R. Offord, "Contributions of risk-factor research to developmental psychopathology," *Clinical Psychology Review*, vol. 17, no. 4, pp. 375–406, 1997. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0272735897000123

[31] M. Kumar, M. Dredze, G. Coppersmith, and M. De Choudhury, "Detecting Changes in Suicide Content Manifested in Social Media Following Celebrity Suicides," in *Proceedings of the 26th ACM Conference on Hypertext & Social Media - HT '15*, no. 2.

New York, New York, USA: ACM Press, 2015, pp. 85–94. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2700171.2791026

[32] S. Liu, M. Zhu, D. J. Yu, A. Rasin, and S. D. Young, "Using Real-Time Social Media Technologies to Monitor Levels of Perceived Stress and Emotional State in College Students: A Web-Based Questionnaire Study," *JMIR Mental Health*, vol. 4, no. 1, p. e2, 2017. [Online]. Available: http://mental.jmir.org/2017/1/e2/

[33] Luyckx, Luyckx, Frederik Vaassen, Claudia Peersman, and Walter Daelemans, "Fine-Grained Emotion Detection in Suicide Notes: A Thresholding Approach to Multi-Label Classification," *Biomedical Informatics Insights*, vol. 5, p. 61, 2012. [Online]. Available: http://www.la-press.com/fine-grained-emotion-detection-in-suicide-notes-a-thresholding-approac-article-a3021

[34] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008.

[35] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The {Stanford} {CoreNLP} Natural Language Processing Toolkit," in *Association for Computational Linguistics (ACL) System Demonstrations*, 2014, pp. 55–60. [Online]. Available: http://www.aclweb.org/anthology/P/P14/P14-5010

[36] B. O'Dea, S. Wan, P. J. Batterham, A. L. Calear, C. Paris, and H. Christensen, "Detecting suicidality on twitter," *Internet Interventions*, vol. 2, no. 2, pp. 183–188, 2015. [Online]. Available: http://dx.doi.org/10.1016/j.invent.2015.03.005

[37] U. Pavalanathan and M. De Choudhury, "Identity Management and Mental Health Discourse in Social Media," *Proceedings of the 24th International Conference on World Wide Web Companion*, pp. 315–321, 2015.

[38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[39] J. Pestian, H. Nasrallah, P. Matykiewicz, A. Bennett, and A. Leenaars, "Suicide Note Classification Using Natural Language Processing: A Content Analysis." *Biomedical informatics insights*, vol. 2010, no. 3, pp. 19–28, 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21643548{%}5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3107011

[40] J. Pestian, J. Pestian, Pawel Matykiewicz, Brett South, Ozlem Uzuner, and John Hurdle, "Sentiment Analysis of Suicide Notes: A Shared Task," *Biomedical Informatics Insights*, vol. 5, p. 3, 2012. [Online]. Available: http://www.la-press.com/sentiment-analysis-of-suicide-notes-a-shared-task-article-a3016

[41] Pew Research Center, "Social Media and Mobile Internet Use among Teens and Young Adults," *PEW Research Center*, vol. 01, pp. 1–51, 2010. [Online]. Available: http://pewinternet.org/Reports/2010/Social-Media-and-Young-Adults.aspx

[42] G. Rakesh, "Suicide Prediction With Machine Learning," *The American Journal of Psychiatry Resident's Journal*, vol. 12, no. 1, pp. 15–17, 2017.

[43] A. L. Rathbone and J. Prescott, "The Use of Mobile Apps and SMS Messaging as Physical and Mental Health Interventions: Systematic Review," *Journal of Medical Internet Research*, vol. 19, no. 8, p. e295, 2017. [Online]. Available: http://www.jmir.org/2017/8/e295/

[44] D. Rickwood, F. P. Deane, C. J. Wilson, and J. Ciarrochi, "Young people ' s help-seeking for mental health problems," *Health San Francisco*, vol. 4, no. 3, pp. 1–34, 2005. [Online]. Available: http://www.atypon-link.com/EMP/doi/abs/10.5172/jamh.4.3.218

[45] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.

[46] A. L. Shortt and S. H. Spence, "Risk and Protective Factors for Depression in Youth," *Behaviour Change*, vol. 23, no. 1, pp. 1–30, 2006.

[47] Sohn, Sohn, Manabu Torii, Dingcheng Li, Stephen Wu, Hongfang Liu, and avishwar Wagholikar, "A Hybrid Approach to Sentiment Sentence Classification in Suicide Notes," *Biomedical Informatics Insights*, vol. 5, p. 43, 2012.

[48] S. H. Spence and A. L. Shortt, "Research Review: Can we justify the widespread dissemination of universal, school-based interventions for the prevention of depression among children and adolescents?" *Journal of Child Psychology and Psychiatry*, vol. 48, no. 6, pp. 526–542, 2007. [Online]. Available: http://dx.doi.org/10.1111/j.1469-7610.2007.01738.x

[49] W. Wang, L. Chen, M. Tan, S. Wang, and A. P. Sheth, "Discovering Fine-grained Sentiment in Suicide Notes." *Biomedical informatics insights*, vol. 5, no. Suppl. 1, pp. 137–145, 2012. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3409482{&}tool=pmcentrez{&}rendertype=abstracthttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3409482{%}7B{&}{%}7Dtool=pmcentrez{%}7B{&}{%}7Drendertype=abstract

[50] C. Welch, "Facebook crosses 2 billion monthly users," 2017. [Online]. Available: https://www.theverge.com/2017/6/27/15880494/facebook-2-billion-monthly-users-announced

[51] H. H. Won, W. Myung, G. Y. Song, W. H. Lee, J. W. Kim, B. J. Carroll, and D. K. Kim, "Predicting National Suicide Numbers with Social Media Data," *PLoS ONE*, vol. 8, no. 4, pp. 1–6, 2013.

[52] World Health Organization, "Preventing suicide: A global imperative," Tech. Rep., 2014. [Online]. Available: http://www.who.int/mental{_}health/suicide-prevention/world{_}report{_}2014/en/

[53] M.-s. Yoon, "South Korea still has top OECD suicide rate," 2015. [Online]. Available: http://www.koreaherald.com/view.php?ud=20150830000310