THE UNIVERSITY *of* EDINBURGH

Monitoring Depressive Symptoms using Social Media Data

Lucia Lushi Chen

Doctor of Philosophy

Institute for Language, Cognition and Computation
School of Informatics
University of Edinburgh

2021

Abstract

Social media data contains rich information about one's emotions and daily life experiences. In the recent decade, researchers have found links between people's behavior on social media platforms and their mental health status. However, little effort has been spent on mapping social media behaviors to the psychological processes underlying the psychopathological symptoms. Identifying these links may allow researchers to observe the trajectory of the illness through social media behaviors.

The psychological processes examined in this thesis include affective patterns, distorted cognitive thinking and topics relevant to mental health status. In the first part of the thesis, we conducted two studies to explore methods to extract affective patterns from social media text. We demonstrated that mood fluctuations and mood transitions extracted from social media text reflect an individual's depressive symptom level. In another study, we demonstrated that the affect from content not written by social media users themselves, such as quotes and lyrics, also reflects depressive symptoms, but the implications from these are different from content written by the users themselves.

In the second part of the thesis, we identified distorted thinking from social media text. We found that these thinking patterns have a higher association with users' self-reported depressive symptom levels than affect extracted from users' text. In the last part of the thesis, we manually compiled topic dictionaries related to suicidal ideations according to the psychopathology literature. We found that users' suicidal risk levels can be estimated by using these topics. The estimation can be improved by combining these topics with results from a language model.

The data-driven empirical studies in this thesis demonstrated that we can characterize the social media signals in a way that impacts our understanding of mental disorder symptoms. We blended data-driven methods such as machine learning, natural language processing and data science with theoretical insights from psychology.

Lay Summary

In the recent decade, researchers have started to use social media data to infer users' mental health status. The current technology often focuses on designing algorithms to identify whether a social media user has depression or not. However, an algorithm cannot diagnose depression from social media data alone. The clinical assessment uses many different information sources to diagnose, but this information is simply missing in the social media context. Our work aims to address these limitations by interpreting social media posts in a meaningful way to impact our understanding of mental disorders.

We first infer a social media users' mood pattern using the emotional words in their posts. *Mood* is an experience of feeling that runs in the background. Mood pattern reflects symptoms of mood-related psychiatric disorders. Our work demonstrated that we could infer one's mood fluctuations and identify thoughts that are not true to reality from social media posts. These signals can provide researchers more information about an individual's depressive symptom level. Nevertheless, we showed that content not written by social media users themselves, such as quotes and lyrics, also reflects depressive symptoms. We also demonstrated that using topics associated with risky behaviors can estimate a social media users' suicidal risk level.

Acknowledgements

I can't believe I am at the end stage of my Ph.D. journey. Growing up, I was one of the poorest students in the class, my parents had to work 15 hours a day to make ends meet. I had never imagined I would do a Ph.D. and I cannot express how grateful I am to have had this incredible opportunity. There are many people I need to thank for their part in this journey, but I am afraid I will fail to mention every one of them who played a big role in my support network.

I am deeply grateful to my supervisor, Maria Wolters, who introduced me to this Ph.D. program; thank you for your incredible support and guidance. I enjoyed being introduced to this amazing interdisciplinary research between psychology, AI and HCI. I had very little knowledge of AI and HCI at the beginning of my Ph.D. Thank you for your patience and guidance in exploring topics and research directions that I'm passionate about. Even though some of them were not within your expertise, you connected me with people who could help me. You introduced me to Walid, who is part of what turned out to be an extraordinary supervisory team.

I am also very grateful for all the lessons I have learned under Walid Magdy's supervision. Walid, you taught me to communicate my ideas clearly and effectively. My experience of being a tutor, lab demonstrator and marker on your course helped me understand the topic in a more in-depth and practical way. Most importantly, you taught me it is okay to fail and being rejected. You also encouraged me, and all of your students to connect and collaborate with different people.

I would also like to thank Heather Whalley, who joined the supervision team when I entered my third year. Thank you for your patience and guidance in the field of psychology. I'm grateful that you introduced me to the journal club and connected me with people from the psychology department.

Thank you to my thesis examiners Bjorn Ross and Yelena Mejova, your feedback to this thesis is incredibly valuable to me. I want to thank Chris Lucas and Robin Hill for being on my annual review committee. I benefited a great deal from your comments and our thought-provoking discussions.

I am very thankful to be part of the big research group in ILCC. I want to thank the members of the SMASH team. Abeer, Dilara, Ibrahim, Youssef, I have learned much from our paper discussion sessions and reading group discussions. I especially need to thank Steve for initiating the group reading. I appreciate you giving feedback on most of my papers and sharing your job hunting experience. I also need to thank

the EdinburghNLP group, Adam Lopez and everyone in the AGORA team. The group discussions in AGORA have helped me expand my knowledge of NLP.

I am very grateful for the fellowship and research visit opportunities during my PhD program. In 2019, I joined the Data Science for Social Good (DSSG) program as a research fellow. I worked with Adolfo, Sebastian, Rayid and the Homeless link organization. I've learnt much from my teammates Harry, Zoe and Austin. I also enjoyed the company of all my friends in DSSG 2019. In 2020, I was invited by Emilio to join a research visit to the Max Planck Institute of Demographic Research (MPIDR). I was grateful to work with Daniela and Sophie on a project tackling mental health issues during COVID. I moved to Germany temporarily in the time of a pandemic. I'm thankful that Maria, Sarah, Gordon, Donata, Sue and Haniyeh provided me with the greatest moral support. I cannot imagine surviving a lockdown in Germany without your companionship. I would also like to thank my supervisor during my Master's degree. Tao Gong, thanks for introducing me to research and publishing the first academic paper of my life.

Thank you to all my friends in Edinburgh who gave me moral support and pulled me through the best and worst times during my Ph.D. Thank you to my office mates Mona, Irene, Spandana, Avashna, Abeer, Sameer and Clara. Especially Sameer and Clara, thank you for all your support and encouragement during my entire Ph.D. You both helped me grow as a person. I also thank my big Edinburgh family, including Yevgen, Ida, Yumnah, Elizabeth, Naomi, Stefanie, Sabine, Ramon, Andreas, Joanna, Julie-Anne, Esma, Carol, Natascia, Kate, Seraphina, Sander, Liquan and many others. I apologize if I have forgotten to mention your name. Carol, you are my true sister. And Nick, thank you for supporting me through thick and thin. I've grown a lot due to our relationship and I've had a great time being with you. Finally, I would like to thank my mom and dad, who have always had my back over the years.

Declaration

I confirm that the work submitted was composed by myself and that the work has not been submitted for any other degree or professional qualification. I confirm that the work submitted is my own, except where work that has formed part of jointly-authored publications has been included. My contribution and those of the other authors to this work have been explicitly indicated below.

The work presented in Chapter 3 was previously published in Frontiers in Psychology 11 as "The Effect of User Psychology on the Content of Social Media Posts: Originality and Transitions Matter" by Lushi Chen (student), Maria Wolters (primary supervisor) and Walid Magdy (co-supervisor).

Chapter 4 was previously published in 12th ACM Conference on Web Science as "Examining the role of mood patterns in predicting self-reported depressive symptoms" by Lushi Chen, Maria Wolters, Walid Magdy and Healther Whalley (co-supervisor).

Chapter 5 was previously published in International Conference on Social Informatics 2020 as "It's Not Just About Sad Songs: The Effect of Depression on Posting Lyrics and Quotes" by Lushi Chen, Maria Wolters, Walid Magdy and Healther Whalley.

Chapter 7 was previously published in Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology as "Similar Minds Post Alike: Assessment of Suicide Risk Using a Hybrid Model" by Lushi Chen, Abeer Aldayel, Nikolay Bogoychev and Tao Gong.

To Mom and Dad, who have always had my back

Table of Contents

Chapter 1

Introduction

1.1 Overview

Depression can greatly affect one's life if left untreated (APA et al., 2013). For example, in the case of severe depression, one could experience mutism and stupor, be subject to cognitive impairment, or even display suicidal behavior (Sonawalla and Fava, 2001; APA et al., 2013). Understanding the psychological processes (e.g., emotion, thinking, perception) that underline depression, especially the trajectory of these processes, is important for early detection and developing treatment methods.

Social media is now used in almost every part of our lives. A report from Pew Research Center pointed out that nearly 90% of the teenagers who participated in a survey in 2018 used social media at least several times a day, of which 45% used it almost constantly (Anderson et al., 2018). Social media data documented people's thoughts and emotions. Some of these thoughts are associated with unhelpful thinking patterns (Shickel et al., 2020; Simms et al., 2017), self-harm or suicidal ideations (Varathan and Talib, 2014; O'Dea et al., 2015). Moreover, the data on one's social media account often span over long periods, perhaps decades. The longitudinal information from social media data is valuable for researchers who study the trajectory of depression (Reece et al., 2017; Chen et al., 2018).

Over the past decade, researchers have explored using social media data to keep track of users' mental health status or symptomatology (Nadeem, 2016; De Choudhury et al., 2013; Glen et al., 2015; Tsugawa et al., 2015). Most existing works in this line of research are framed as an optimization towards an objective function. In particular, many research groups constructed classifiers to predict self-reported health status. Some researchers have established links between some social media behaviors

and mental health status, such as sentiment and depression (Deshpande and Rao, 2017; Mustafa et al., 2020; Wang et al., 2013a), social network and depression (Wang et al., 2013a; Saravia et al., 2016; Islam et al., 2018). Depression is a diagnosis for a cluster of symptoms, and there are many underlying psychological processes that trigger and sustain these symptoms. However, the links between social media behaviors and the psychological processes underlying the illnesses are seldom explored. For example, it is well known that affective style (Davidson, 1998; Akiskal and Akiskal, 2005) and cognitive distortions (Lefebvre, 1981; Poletti et al., 2014) perpetuate various types of affective disorders, yet few studies examine whether these psychological processes can be observed on social media data. These psychological processes are critical for researchers to understand the trajectory of the illnesses.

1.2 Important Concepts

Before we delve into the research question and goals, we explain the important concepts in this thesis.

Cognition *Cognition* refers to the mental process of learning and comprehension. Cognitive psychologists build up cognitive models to explain how perception, attention, language, memory, thinking and consciousness are involved in information processing.

Affect, Emotion and Mood *Affect* in psychology refers to feelings, emotions, or mood that we experience as part of our everyday lives. Affect can be categorized as positive or negative. We experience affect in the form of mood and emotions. There are fundamental differences between mood and emotions. *Emotions* are reactions to stimuli. Emotions are brief, less fine-grained and more intense compared with mood. For example, angry, sad and happy are emotions. *Mood* refers to the positive or negative feelings that run in the background. When we are in a good mood, we tend to have more positive emotions.

In psychopathology, *affective symptoms* refers to psychiatric symptoms related to mood or emotional responses, such as feelings of sadness, excessive or sustained feelings of enthusiasm, confidence and energy.

Affective pattern The pattern of affect (*affective pattern*) reflects our strategies to increase, maintain or decrease our emotions and feelings (Gross, 1999). Affective pattern is closely linked to psychopathology. For example, Gross (1999) suggested that suppressing one's emotions is associated with poorer psychopathology. However, the effect of emotion suppression on psychopathology is mediated by whether an individual uses cognitive strategies to interact with the situations to change their emotions (Eftekhari et al., 2009).

Affective disorders *Affective disorders* (mood disorders) are psychiatric problems that primarily affect an individual's mood, characterized by pervasive dysregulation of mood (Akiskal and Van Valkenburg, 1994). For example, prolonged depressed mood is a characteristic of major depressive disorder (APA et al., 2013). The main types of affective disorders include depression and bipolar disorder.

Feature This thesis covers theories in psychology and machine learning domains. The two fields have overlapped terms referring to different concepts. For example, *psychological feature* refers to the characteristics of the psychological life. *Feature* in machine learning refers to the independent variables used in an algorithm for predicting the dependent variable. Feature in this thesis refers to the machine learning concept.

1.3 Goals and Research Questions

Based on the existing research gaps in identifying links between psychological processes and social media behavior, we propose the following **research question**: **Can we represent social media signals in a way that reflects the psychological processes that underlie depression?** This thesis shows that we can use robust statistical approaches to extract affective patterns, cognitive distortion and thoughts related to self-harm behaviors from social media text. We define three main goals to answer the research question (see Figure 1.1). Below we list the three main goals and the corresponding research questions:

- Exploring methods to extract affective patterns from social media text and examining their implications to depressive symptoms (Chapter 3 - 5).
  - Do changes of affect correlate with users' personality traits and mental wellbeing? (Chapter 3)

  - Are mood representations associated with the severity of self-reported depressive symptoms? (Chapter 4)
  - Is posting quotes associated with levels of depression symptoms? (Chapter 5)
  - What are the themes and emotions conveyed in the lyrics and quotes? (Chapter 5)

- Identifying cognitive distortions from social media text and examining their associations with depressive symptoms (Chapter 6).

  - Does cognitive distortion have an association with users' depressive symptoms, well-being and personality dimensions?
  - What are the linguistic characteristics among users who posted content with cognitive distortion more frequently?

- Classifying suicidal risk with topics related to risky behaviors and their motivations (Chapter 7).

  - Can we infer one's suicidal risk level with their social media posts?

The three goals mainly focus on studying social media signals that reflect affective and cognitive symptoms. Affective symptoms are a major category of symptoms for affective disorders. Most of the literature represents the affective signals by averaging the affect values extracted from the text over a large period. This approach provides us with the simple information that users with more depressive symptoms use more words with negative affect on their status updates in general (Tsugawa et al., 2015; De Choudhury et al., 2013). However, people's affect changes from moment to moment. We can obtain more insights by observing the changing pattern of affect extracted from social media data. Psychopathology literature suggests that the intensity, variation, duration, frequency, and mood category are all relevant to affective disorders (Akiskal, 1996).

Besides affective symptoms, the cognitive theory of psychopathology suggests that there are common thinking patterns that expose one to greater risks of developing psychiatric problems. These thinking patterns (cognitive distortion) perpetuate one's depressive symptoms but are under-explored in the social media context. Affective patterns and cognitive distortion are both psychological processes underlying depression and many other affective disorders.

Through answering the research questions of the main goals, we show that the following information extracted from social media text might provide us insights to understand the trajectory of symptom development:

- the intensity, variation, duration, frequency, and category of mood
- cognitive distortions
- topics that reflect symptoms or risky behaviors

Data related to mental health is highly sensitive. Therefore, many researchers do not publish their datasets. Mental health researchers rely on a decentralized pool of resources. Many researchers collected and annotated their own datasets (Harrigian et al., 2021). This thesis focuses on constructing representations using social media data. Therefore, we used existing datasets to conduct our studies.

For studies in Chapter 3 - 6, we used a dataset from myPersonality. myPersonality was a Facebook application that collected self-measurement scales from social media users. After participants filled out the scale, they were given brief feedback of their scores. Then they were also given the choice of sharing their Facebook data to the application for research purposes. myPersonality collected hundreds of measurement scales, such as the BIG-5 personality tests, Center for Epidemiologic Studies Depression Scale (CES-D), Satisfaction of Life Scale and many others. There have been more studies using social media data to study psychopathology in the recent two years. However, most of the datasets with human annotations or self-reported measurement as gold standards are not published due to the sensitive nature of mental health data. Reddit suicide risk assessment (CLpsy, 2019) is one of the publicly available datasets for researchers in the recent two years. In 2021, the OurDataHelp project (UMD, 2021) started collecting large-scale longitudinal self-measurement scales from social media users to build a benchmark dataset for mental health studies with social media data. OurDataHelp only allows researchers to access the data on a server that is disconnected from the internet. Participants have been fully explained the intention of the study and how their data would be used. However, this dataset was only made available after the completion of this thesis. Future studies can apply the techniques proposed in this work to a more recently collected dataset.

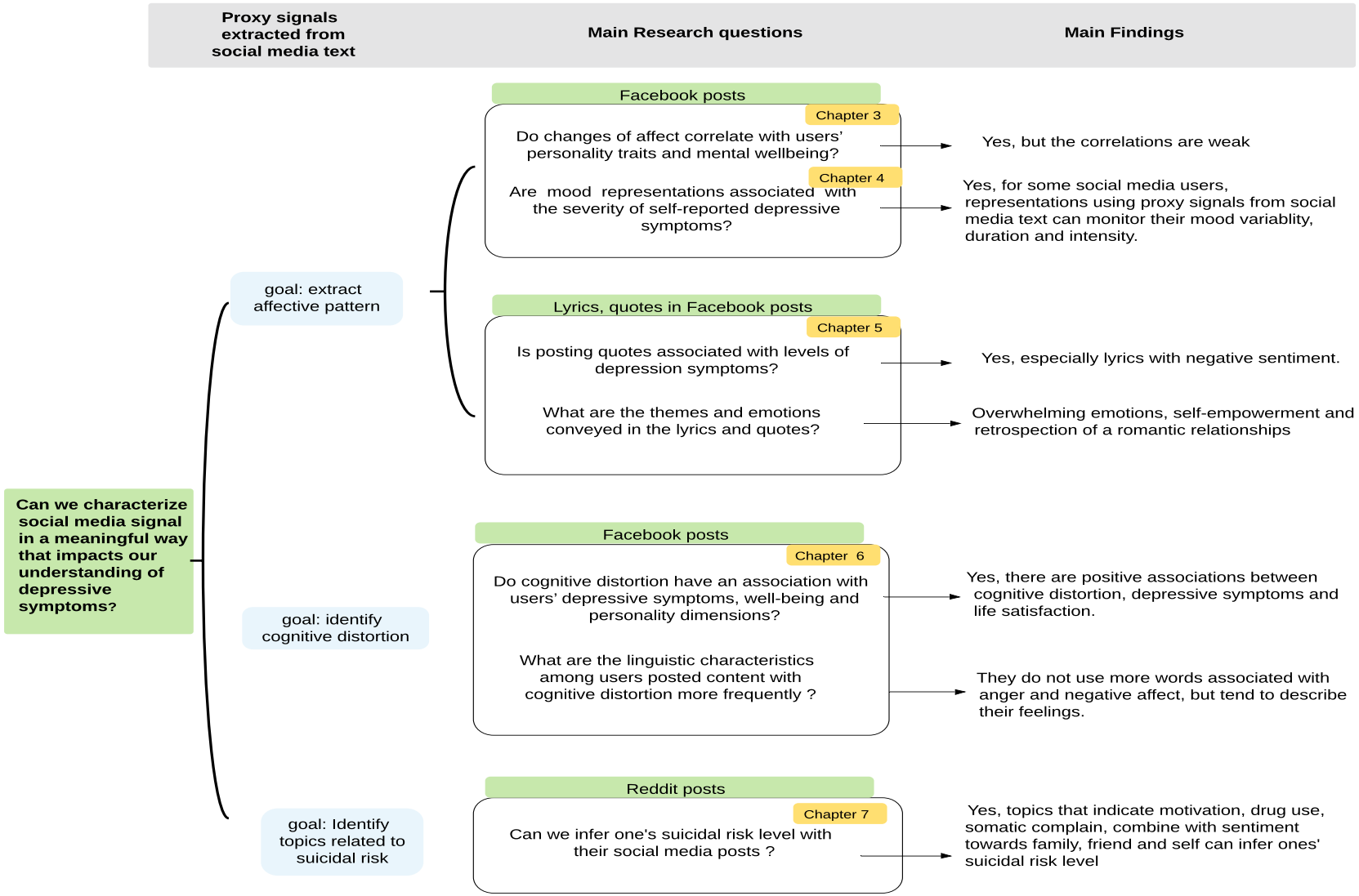


Figure 1.1: Thesis Structure

1.4 Contributions

Our work blends natural language processing, machine learning, information retrieval, and data science with insights from psychology and social computing. Although the techniques described in this thesis mainly focus on depressive symptoms, they can be translated into symptom tracking topics. We believe our findings can provide insights for researchers to observe the trajectory of depression. Most importantly, we have connected the machine learning results and findings from data analysis to the theories of depression, affective disorders and risky behaviors. Below is a list of contributions from our work:

- We have brought novel computational techniques to identify affective patterns (Chapter 3 and 4) and harmful ideations (Chapter 7) in social media text.
- We have found that non-user-generated content, such as lyrics and quotes, reflects users' depressive symptom levels.
- We have adopted a hand-annotation analysis of social media posts to identify cognitive distortions (Chapter 6).
- We have developed computational approaches to represent social media signals based on the cognitive-behavioral model of depression and affect theory (Yurica and DiTomasso, 2005; Lefebvre, 1981; Kaplan et al., 2017) (Chapter 3 and 4).
- We have contributed insights towards developing a data-driven approach to analyzing, modeling, and tracking symptoms of affective disorders.

1.5 Thesis Organisation

This thesis is organized as follows. Table 1.1 shows the publications on which each chapter is based.

Chapter 2 This chapter reviews the background material for this thesis, explains our subject of interest and the motivation for choosing the topic of representing psychological processes with social media signals. We summarize five types of social media signals that have been found to be relevant to depressive symptoms. Most of these signals fall into one of the three domains in modern psychology: affect (feelings), behavior (interactions) and cognition (thought).

We present the cognitive model of depression, affective theory and background literature that is crucial to give perspective on what social media signals have been widely studied and suggested as important to infer users' depressive symptoms.

Chapter 3 Affective literature suggests that different aspects of affect, such as magnitude, alternation, and categories, reflect on a person's tendencies for regulating emotions. These tendencies are not only behavioral features but also perpetrators for a wide range of conditions. Although researchers have found that the category and magnitude of affect extracted from social media data are associated with social media users' depressive symptoms, is it unknown whether the affective alternations extracted from social media text also reflect one's wellbeing.

In this chapter we use categorical values to represent transitions from one affective state to another. We introduce a silence day token to represent days when a user did not post any content. Our results have shown that the representations we constructed can give us a more nuanced picture of social media users' psychological traits than simply averaging the affective values over a long time. We have found that participants who are more extroverted tend to post positive content on consecutive days and that participants who are more agreeable tend to avoid posting negative content.

Chapter 4 In this chapter, we explore various approaches to represent mood (a form of affective experience) in the social media text. We use sliding window techniques, combined with different methods to measure mood in constructing the representation. To test whether these representations are associated with social media users' self-reported depressive symptom levels, we use Gaussian Process regression to measure the fluctuations in the mood representation. We observe less evidence of mood fluctuation expressed in social media text from those with low symptom scores than others with high symptom scores. We also use Hidden Markov Models to estimate latent variables with mood as observations. Assuming the latent variables are associated with depressive symptom levels, we use these latent variables to classify self-reported depressive symptom levels and achieve a high precision rate. Finally, we use these mood representations to classify users' symptom levels. Using the mood representations extracted from social media text, we are able to infer users' symptom levels. These representa-

tions also provide researchers insight into the trajectory of depression.

Chapter 5. Our experiments in Chapter 4 focus on content that is created by social media users themselves. Content not created by the users is often ignored because it might not reflect the users' own emotions. In Chapter3, our pilot study examined the links between posting non-original content (e.g., lyrics, quotes) and depressive symptoms on a small subset of sample. In this chapter, we examine whether the affect from quotes and lyrics posted on Facebook are associated with underlying symptoms of depression. We found that participants with elevated depressive symptoms tend to post more lyrics, especially lyrics with neutral or mixed sentiment. The lyrics center around overwhelming emotions, self-empowerment, and retrospection of romantic relationships. Our findings have suggested that removing quotes, especially lyrics, might eliminate content that reflects users' mental health conditions.

Chapter 6. The cognitive-behavioral theory states that individuals with depression exhibit distorted modes of thinking. These thinkings perpetuate their depressive symptoms. Cognitive distortion can be identified in the digital text, but this area is still underexplored. In this chapter, we examine cognitive distortion in the social media text. We annotate negative emotions and cognitive distortion among more than 4000 Facebook posts posted by 71 Facebook users. We have found signs of cognitive distortion presented on Facebook data. Our result have shown that cognitive distortion presented on social media text has higher associations with depressive symptoms than negative sentiment averaged over one year. We further identify the language characteristics in cognitive distortion extracted from social media text.

Chapter 7. Suicide is a significant problem globally. Most suicides are related to psychiatric diseases; individuals with depression or substance use disorders are the highest risk. We explore computational approaches of detecting suicidal ideations manifested on a set of social media data annotated by clinical specialists. We approach the problem with three separate models: a behavior model, a language model, and a hybrid model. For the behavioral model approach, we model each user's behavior and thoughts with four groups of features: posting behavior, sentiment, motivation, and content of the user's posting. We use these features as input in a support vector machine (SVM). For the language model approach, we train a language model for each risk level using all the users' posts

Table 1.1: Publication and chapters

paper title	chapter	published in
The Effect of User Psychology on the Content of Social Media Posts: Originality and Transitions Matter	3	Frontiers in Psychology 11
Examining the role of mood patterns in predicting self-reported depressive symptoms	4	12th ACM Conference on Web Science
It's Not Just About Sad Songs: The Effect of Depression on Posting Lyrics and Quotes	5	International Conference on Social Informatics 2020
Examining Cognitive Distortion in Social Media Text	6	
Similar Minds Post Alike: Assessment of Suicide Risk Using a Hybrid Model	7	Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology

as the training corpora. Then, we compute each user's posts' perplexity to determine how likely his/her posts are to belong to each risk level. Finally, we build a hybrid model that combines both the language and behavioral models. The hybrid model demonstrates the best performance in detecting the suicide risk level.

Chapter 8. This chapter is a summary of the main findings of this research. We connect our experimental design and results to the theories of psychopathology. We discuss the potential impacts of monitoring psychological processes using social media data and possible approaches for addressing biases and ethical concerns in this research area. Finally, we discuss the limitations of this research and possible future directions.

Chapter 2

Background and Literature Review

This chapter provides the literature background for this thesis. Affect, cognition and behavior are three major divisions in modern psychology. This chapter first provides an overview of the affective, cognitive and physical symptoms manifested in depressive disorders. Then we present an overview of publications that examine social media behaviors and depression. We summarize five types of social media signals associated with self-reported depressive symptoms. Among these signals, affective signals are most commonly reported as important features in the classification methods, followed by topics. *Topic* refers to textual content surrounding a certain theme. In addition to affect and topics, our overview has also suggested that cognitive distortion, a psychological process that perpetuates depressive symptoms, can be partially identified in the social media text. Finally, we identify the research gaps for mapping social media behaviors to the psychological processes underlying depressive symptoms.

2.1 Depression

Depression is a mood disorder that affects more than 250 million people around the world (WHO, 2020). In extreme cases, depression can lead to suicide. People with major depressive disorder (MDD) are 21 times more likely to commit suicide than non-depressed individuals (Arensman et al., 2015). The prevalence of depression is high compared with other psychiatric disorders. For example, one year and lifetime prevalence of depression are 12.9% and 10.8%, respectively. Prevalence for schizophrenia is 0.4% (Saha et al., 2005), bipolar is around 4% (Ketter et al., 2010). Depression is also significantly higher in women (14.4%) and countries with a medium human development index (29.2%) (Lim et al., 2018).

Traditionally, researchers study symptoms or the course of depression by collecting self-reported or interview data from the participants. The data collection process is sometimes spanning over several generations of researchers. For example, a longitudinal study on 2,320 participants would require 10,982 assessments over a few decades (Sutin et al., 2013). Most of the existing longitudinal studies about depression involve observing the change of self-reported measurement of depressive symptoms over the years. The data collection process for the longitudinal study is painstaking, and there are also gaps ranging from a few months to a year in between each data collection. These gaps need to be reduced to days or weeks if researchers want to measure the psychological processes in a fine-grained manner. For example, as people's mood changes daily, conducting a longitudinal study for mood would require collecting self-reported mood frequently over the years. Due to the constraint of data collection in longitudinal studies, little is known about the trajectory of the psychological processes underlying mental illnesses.

In the recent decade, social media platforms have provided researchers an alternative source to study the trajectory of people's psychological processes. Researchers have started to use computer algorithms to extract proxy signals that reflect depressive symptoms from social media data. By observing how these signals change over time, researchers can have a longitudinal, life-course view of their psychological processes.

This thesis focuses on studying psychological processes underlying depression as an example, and we demonstrate techniques that may be adopted to observe psychological processes of other mental disorders. We propose to represent social media signals in a way that can impact researchers' understanding of an individual's affective and cognitive processes. This chapter aims at identifying the links between social media signals and depression based on existing literature. In the coming chapters, we will use statistical approaches to represent the targeted social media signals.

2.2 Symptoms of depression

The Diagnostic and Statistical Manual of Mental Disorders Fifth Edition (DSM-5) defines many subtypes of depressive disorders: disruptive mood dysregulation disorder, major depressive disorder, persistent depressive disorder, and premenstrual dysphoric disorder, substance-induced depressive disorder, etc.

Each sub-type shares some characteristics with other sub-types. Some stereotypical depressive symptoms across different sub-types include depressed mood, irritability

or anger, decreased interest in usual activities, and feeling of worthlessness. These symptoms can be categorized into cognitive, affective, and physical symptoms. In this section, we introduce these three categories of symptoms.

2.2.1 Cognitive Symptom

Cognitive symptoms of depression involve difficulty concentrating, forgetfulness, memory loss, reduced reaction time, indecisiveness and cognitive distortion. Cognitive distortion refers to the thinking pattern that leads to misinterpretations of reality. More than half a century ago, the father of cognitive therapy, Aaron T. Beck. asked his patients to focus on their automatic thoughts. He found that the content of these thoughts was either misinterpretations or exaggerations of a situation. They were cognitive distortion that caused people to view the reality inaccurately, often in a negative way (Lefebvre, 1981; Norman et al., 1983). For example, "No one cares about me in this world". The cognitive distortion varies according to the major psychiatric problems experienced. For example, patients with depression have a general theme of self-criticism and regret. The more severe the disorder, the bigger portion of the stream of consciousness that the automatic thoughts occupied (Beck, 2019).

Later on, Beck established the theory of cognitive models of psychopathology. The cognitive model of psychopathology suggested that there were biased schemata that influence cognitive distortion. These biased schemata influence one's attitude and behavioral responses and put an individual at a greater risk of developing depression (Kovacs and Beck, 1986). Oliveira (2014) summarized 15 types of biased schemata, see Table 2.1.

Most of the cognitive symptoms require cognitive testing. Researchers cannot simply observe these symptoms in the social media context. However, cognitive distortion can be identified in the social media text. Simms et al. (2017); Bathina et al. (2020) suggested cognitive distortions can be identified in Tumblr and Twitter posts. The language markers that indicate cognitive distortion included first-person pronouns and negations. In Chapter 3, we introduce methodologies to annotate cognitive distortion on Facebook posts and we examine their associations with self-reported depressive symptom level.

Table 2.1: Checklist of Cognitive Distortions (Oliveira, 2014)

category	explanation
jumping to conclusions	Draw negative conclusion without evidence to support that conclusion.
catastrophizing(what if)	One believes the worst situation will occur
comparison	One believes that he/she is worse than others because one tends to compare himself/herself with others
dichotomous/black-and-white thinking	The tendency to view all experiences as fitting into one of two categories (e.g., positive or negative; good or bad).
disqualifying the positive	Denying a situation/trait/event is positive
emotional reasoning	Letting your emotions direct your conclusions about yourself, others, or situations.
fortunetelling	Assume that some event or events will end badly for us, that we will fail at something or we will be in danger, more as an assumption rather than an educated guess.
labeling	Labeling oneself or others using derogatory names.
magnification	Exaggerate the importance of your errors, fears, and imperfections.
mind-reading	Concluding that other people are reacting negatively, or thinking negatively toward him/her, without evidence to support the conclusion.
minimization	Minimizing or discounting the importance of some event, trait, or circumstance.
overgeneralization	When someone overgeneralizes, they see that one negative event in their life as a never-ending pattern.

personalization	Assume you are responsible for an external event over which you have no control. When you personalize, you feel guilty because you confuse influence with control over others.
selective abstraction	The process of exclusively focusing on one negative aspect or detail of a situation, magnifying the importance of that detail, thereby casting the whole situation in a negative context.
"should" statements	A pattern of having internal expectations or demands on oneself, without whether it's reasonable to have these expectations or not.

2.2.2 Affective Symptoms

Affective symptoms refer to psychiatric symptoms related to mood or emotional responses. Affective symptom is reflected in one's affective style. The "affective style" refers to the variance of quality and intensity of mood and emotional reactions (Davidson, 1998). Davidson (1998) proposed several aspects of affective styles: the intensity or magnitude of the response, the duration of the response, the frequency or the number of times the response occurs within a given period and the category of the response. Emotion-based symptoms in many DSM-5 disorders can be characterized by the aspects in the affective styles. For example, a major depressive episode is characterized by depressed mood and anhedonia; A manic episode is characterized by elevated, expansive, or irritable mood; Quickness to react angrily is a behavioral feature for paranoid personality disorder; Emotional coldness, flattened affect is the characteristics of schizoid personality disorder; Affective instability, which refers to marked reactivity of mood, inappropriate, intense anger, or difficulty controlling anger are markers for borderline personality disorder (APA et al., 2013).

Affective style underlies one's vulnerability to psychiatric disorders (Rottenberg and Gross, 2003; Akiskal, 1996). Assessing affective style is useful for observing the trajectory of a mental disorder. However, assessing self-reported affect in everyday life requires a costly extended period of data collection, even with technology (Caldeira et al., 2017). Social media data provides researchers with an alternative approach to look at the changes of affect over time with minimal effort in data collection. Studies

have found that people's affective symptoms are associated with what they post on social media. For example, participants with more self-reported depressive symptoms used more negative affective words (e.g., *sad, cry, hate*) in their social media text than those with fewer symptoms (De Choudhury et al., 2013; Park et al., 2012; Tsugawa et al., 2015; Chen et al., 2020a).

2.2.3 Physical Symptoms

Physical symptoms often accompany depression. Minor illnesses and pain are the most commonly presented symptoms. Sleep disturbance, fatigue and exhaustion, appetite change, agitation, and restlessness are also commonly presented (APA et al., 2013). A high percentage of patients with depression sought treatment for minor physical symptoms in a primary care setting. The diagnosis of depression was often masked by minor physical illnesses (Trivedi, 2004).

Physical symptoms can be identified in social media text using natural language processing techniques. However, whether a social media user mentions these symptoms on social media text is arbitrary and heavily depends on the user's self-disclosure level. Sometimes, a user's posting pattern reflects physical symptoms. For example, posting late at night reflects sleep disturbance (De Choudhury et al., 2013; Resnik et al., 2015; Nambisan et al., 2015). Wang et al. (2013a) also found depressed individuals tend to have fewer interactions with the audiences. However, findings on social media data only show a general tendency. There are always exceptions to this tendency. For example, people may interact less with other social media users because they use social media platforms mainly to obtain information.

2.3 A Review on Manifestation of Depressive Symptoms on Social Media Platforms

In the previous section, we introduce depressive symptoms described in the psychology literature. In this section, we conduct a review to understand what type of social media signals are associated with depressive symptoms based on the existing findings. Some of these signals may reflect the psychological processes underlying the symptoms. In our later chapters, we will focus on these signals.

One way of testing whether social media signals are associated with depressive symptoms is to use these signals to predict or infer a mental health status. Many re-

views have been conducted to demonstrate the techniques of using social media data to predict or infer mental health status. For example, Calvo et al. (2017) focused on analyzing the types of natural language processing techniques for detecting depressive symptoms. Guntuku et al. (2017) and Skaik and Inkpen (2020) summarized the prediction methods, prediction performance, criteria to assess mental illnesses, types of features being used, outcome types, models and evaluation metric.

Existing reviews mainly focus on the techniques being used in the classification tasks and their performances. However, the links between social media signals and depressive symptoms documented in psychopathology literature have not yet been summarized. Therefore, this overview aims to map social media signals to depressive symptoms described in psychopathology literature. By analyzing 107 publications, this overview addresses the following questions:

1. What type of techniques are applied to infer/predict depressive symptoms?

2. What key features are reported as predictive to depressive symptoms from the literature?

3. Are the key features reported as predictive to depressive symptoms different across platforms and culture?

4. What types of techniques are used to construct mood and emotion features?

This overview is structured as follows: section 2.3.1 describes the methodology for data collection, the selection criteria and the compilation process. Section 2.3.2 quantify the types of techniques applied to social media data to infer or detect depressive symptoms. Section 2.3.3 summarizes the key features reported as useful to identify signs of depression or depressive symptoms. Section 2.3.3.1 compares the key features across platforms. Section 2.3.4 presents the techniques of extracting mood or emotion features from social media text to detect depressive symptoms.

It is important to note that the “gold standard” from most of the existing literature does not represent a clinical diagnosis of depression (Guntuku et al., 2017; Chancellor and De Choudhury, 2020) because the diagnostic process is intricate, social media data do not contain all the necessary information to make a diagnosis. Therefore, in this work, we have used the phrase “infer depressive symptoms” instead of “detect depression” to refer to the task of the existing literature.

2.3.1 The Review Methodology

We have selected publications related to inferring or detecting depressive symptoms. These studies involved using user-generated data on social media platforms (e.g., Facebook, Twitter, Weibo, Reddit). The selected works involve:

- inferring depressive symptoms level of individual social media user
- inferring the depressive symptoms level expressed in one single post
- examining the language characteristics from users who self-reported having depression

2.3.1.1 Record Identification

The search was done on 30th December 2020. A total of 808 related papers were identified after searching PubMed, ACM Digital Library, IEEE Explore, PsycInfo, Web of Knowledge and Google Scholar. Articles were selected following the Boolean search strings in title and abstract:

- PubMed: (social media[tw] OR Facebook[tw] OR Twitter[tw] OR Instagram[tw]) AND (predict[tw] OR detect[tw]) AND (depress*[tw])
- PsycInfo: (“social media” OR “Facebook” OR “Twitter” OR “Instagram”) AND (“predict” OR “detect”) AND (“depression”)
- ACM Digital Library: (“social media” OR “Facebook” OR “Twitter” OR “Instagram”) AND (“predict” OR “detect”) AND (“depression”)
- IEEE Explore: (“social media” OR “Facebook” OR “Twitter” OR “Instagram”) AND (“predict” OR “detect”) AND (“depression”)
- Google scholar: “depression” AND “social media” AND (“predict” OR “detect”)

2.3.1.2 Record Selection

The search identified 808 articles in total, 158 duplicates were removed. Based on the title and abstract manual screening, 457 studies were identified as irrelevant. 193 studies were assessed full-text for eligibility. 85 articles were excluded based on the following eligibility criteria and exclusion criteria, resulting in 107 papers after the full-text screening. Studies were excluded if:

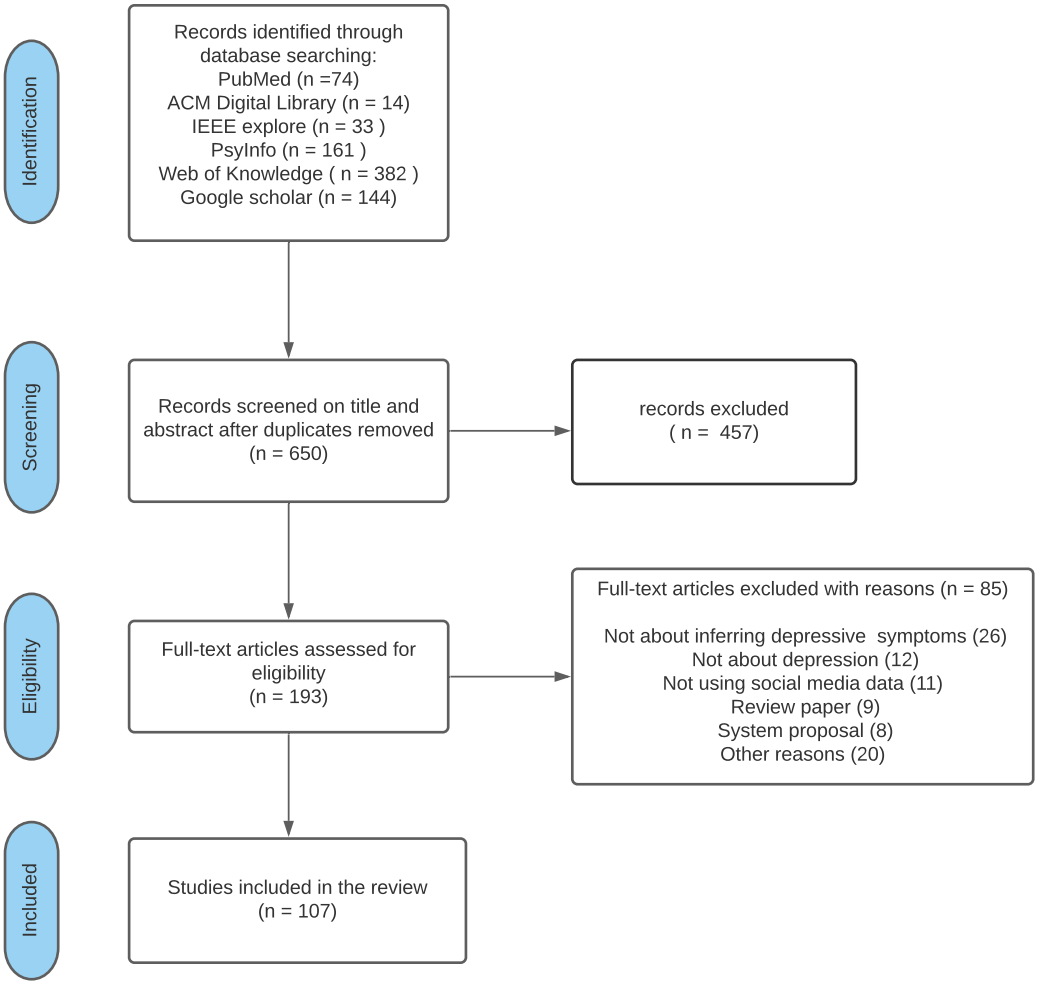
- They use other than social media data, such as mobile sensor data;
- They are review papers or system proposals;
- They study the frequency of using social media platform and associations with users' depression;
- They study social media data that reflects users' mood or emotions but not they do not link the findings to depression;
- The study is not conducted at a population level.

We used the following inclusion criteria to select the paper for this review:

- The paper must involve using social media data to infer or predict depressive symptoms. Social media platform refers to any web-based technology that facilitates the creation and sharing of information, including but not limited to Facebook, Twitter, Reddit, Instagram, Weibo, etc.
- The paper examined factors reflected from social media data that contribute to depression, including postpartum depression.
- The type of paper must be "research article" published in journals or conferences.
- The paper can involve predicting or inferring multiple mental disorders, including depression
- The full text must be written in English.
- The paper can be qualitative or quantitative analysis.

The query "social media" enables the returned results to include other social media platforms, such as Reddit and LiveJournal (see Section 2.3.3.1 for the details). Based on the above eligible criteria, ultimately, 107 publications were selected for the current review. The full list of all the selected papers was provided in the Appendix (see Figure 2.1 for the selection process).

Figure 2.1: Procedural flowchart



2.3.1.3 Data Extraction

To assist with the systematic extraction of information from the papers, we used a data extraction spreadsheet. The spreadsheet included columns for characterizing papers by title, author, publication year and other columns that describe the following variables (see Table 2.2).

Table 2.2: Information extracted from selected literature

column variables	values
does the paper involve feature analysis	yes, no
types of social media features that are found to be useful to infer depressive symptoms	affect, personal pronouns, ego network and social capital, topics, posting time and posting volume, other features
modeling methods or other methods to identify the features	classical machine learning (ML), deep learning, both classical ML and deep learning, textual analysis, statistical analysis
language of the data	English, Chinese, Spanish, Thai, Arab, Portuguese
source of the data	Twitter, Reddit, Facebook, Weibo, combing multiple sources, other sources

Each selected paper was analyzed using this template. In Section 2.3.4, we also analyze the feature engineering techniques for extracting social media signals that were used in modeling depressive symptoms.

2.3.2 Techniques Applied to Infer Depressive Symptoms

Figure 2.2 shows that the number of studies on using social media signals to infer or predict depressive symptoms has dramatically increased since 2017. Most of the existing studies in this line of research mainly fall into two categories:

1. Identify themes and topics in the social media posts (Cheng et al., 2016; Feldhege et al., 2020; Bataineh et al., 2019; Resnik et al., 2015; Cavazos-Rehg et al.,

2016) and their comments (Andalibi et al., 2017);

2. Using social media proxy signals to classify a mental health status (De Choudhury et al., 2013; Tsugawa et al., 2015).

There are 11 studies (10%) in the first category. These studies mainly used the qualitative analysis or manual annotations to identify themes and content characteristics that indicate certain types of symptoms, such as cognitive biases, hopelessness and suicidal ideations (Cheng et al., 2016; Feldhege et al., 2020; Bataineh et al., 2019).

In this overview, the majority of the studies (N = 94, 88%) use social media proxy signals to classify whether a social media user has a high level of self-reported depressive symptoms. Similar to Guntuku et al. (2017)'s finding, we find that our selected sample of papers often adopted classical machine learning (e.g., regression, SVMs, nearest neighbor, decision trees, PCA, naive Bayes classifier), deep learning, or both methods to classify self-reported depressive symptom levels (De Choudhury et al., 2013; Tsugawa et al., 2015; Shen et al., 2018; Shah et al., 2020).

Figure 2.4 shows classical machine learning (e.g., decision trees, support vector machine) is a dominant approach in our selected sample. Works adopted classical machine learning models often involve feature analysis in the modeling process (Yang et al., 2020; Mustafa et al., 2020; Benamara et al., 2018). However, as deep learning techniques became more popular since 2018 (see Figure 2.4), fewer papers included feature analysis (see figure 2.3).

The functional process of the deep learning model can be very complex, often resulting in "blackbox". Among the 18 papers using deep learning techniques, only one work included feature analysis (Wu et al., 2019). We have identified nine papers using both classical machine learning and deep learning techniques and comparing the performances of multiple models.

Despite efforts to improve algorithm performance, inherent biases in the dataset (e.g., population bias, content bias, see Chapter, 1, Section 1.1) and modeling processes suggests that the classification technology was far from ready to be used by clinicians or researchers.

2.3.3 Key Social Media Signals for Inferring Depressive Symptoms

Among the selected publications, 58 of them have reported what type of social media signals are important to infer depressive symptoms. We summarize five types of

Figure 2.2: Number of Publications

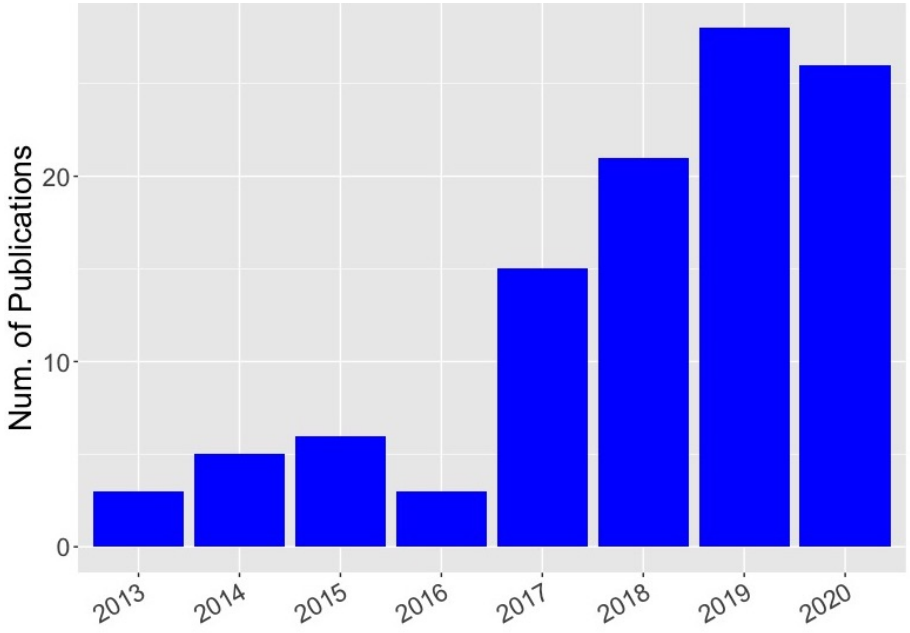
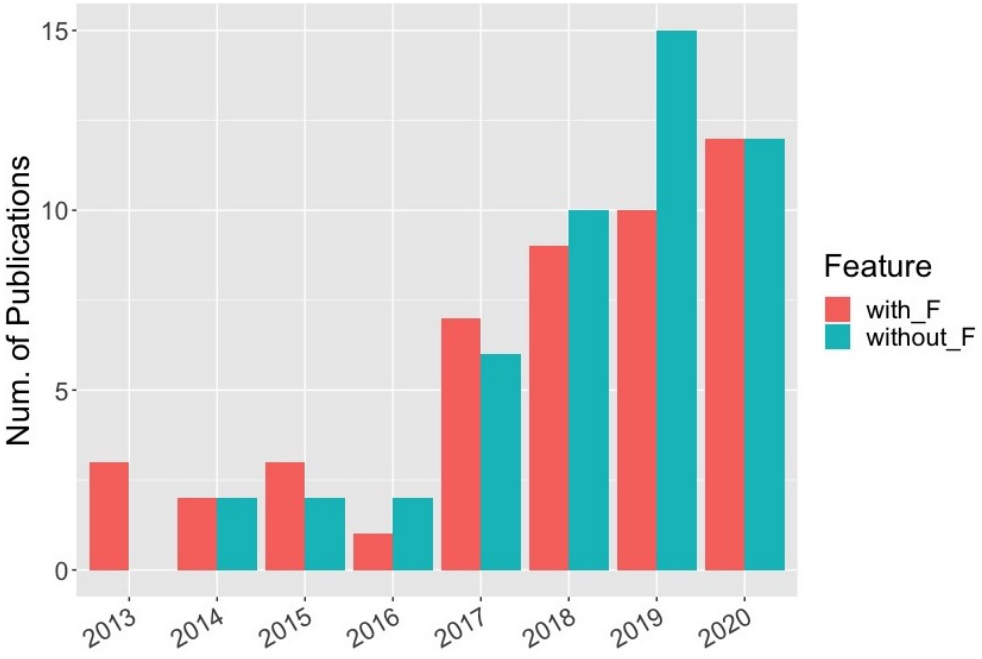
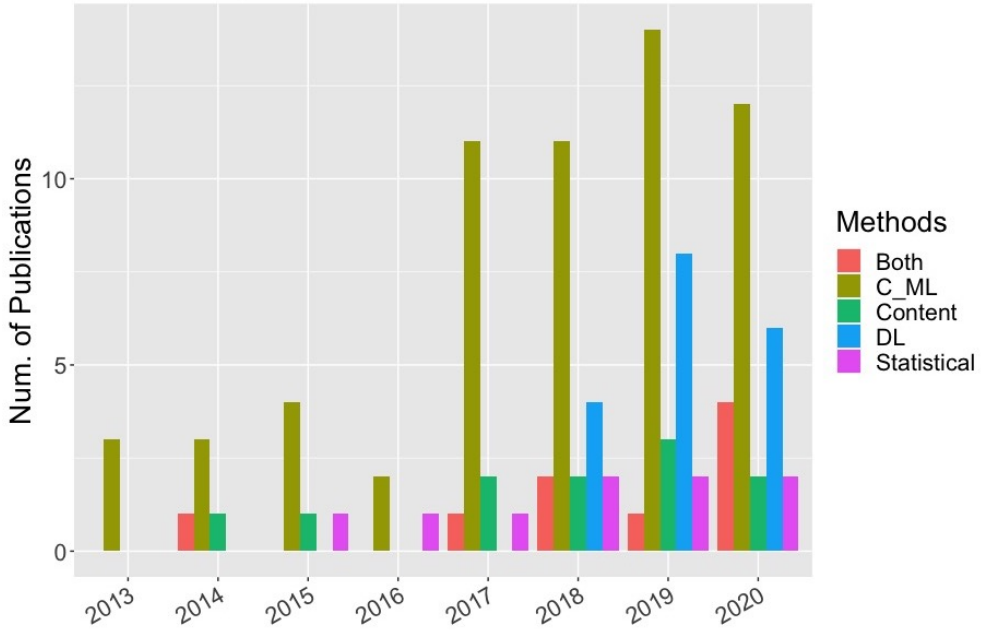


Figure 2.3: Feature Analysis by Year



Note: with_F: papers reported features analysis, without_F: papers do not report feature analysis

Figure 2.4: Methods Used in Publications



Note: C_ML: classical machine learning, DL: deep learning, Both: papers contain classical machine learning models and deep learning models, Content: qualitative content analysis, Statistical: statistical analysis of term frequencies

features that are commonly found to be associated with depression (see Table 2.4 for publications in each category). The five types of features include:

- affect or emotions (positive and negative affect, hostility, sadness, e.g., angry, sad)
- personal pronouns (1st and 3rd personal pronouns, e.g., I, his)
- ego network (number of followers, followees, comments, likes)
- topics (a cluster of certain themes, e.g., physical symptoms)
- posting time and posting volume

The affect or emotions, personal pronouns and topics are linguistics features extracted from social media text. Section 2.3.4 explains the common approaches used to extract each type of signal. For affective symptoms, the majority of the studies find that users with a high level of depressive symptoms tend to use words with negative affect (De Choudhury et al., 2013; Tsugawa et al., 2015; Fatima et al., 2018).

Frequent use of first-person pronouns is also an indicator for depressive symptoms because it may indicate a tendency to focus on self. (De Choudhury et al., 2013; Shen et al., 2018; Wu et al., 2019). An ego network refers to a network made up of a central node (e.g., an individual) and all other nodes (e.g., social ties) connected to the central node. Studies focused on ego networks find that social media users with more depressive symptoms tend to have less influence in their ego network. For example, they receive fewer comments and likes and they have fewer followers (Yang et al., 2020; De Choudhury et al., 2014; Shen et al., 2018).

Moreover, many studies have suggested that posting time reflect users' sleeping pattern (Wongkoblap et al., 2018; Benamara et al., 2018; Shen et al., 2018; Li et al., 2020). Some users may struggle with insomnia if they are awake at night, although others may be awake at night due to work reasons. Multiple studies find that users with a high level of depressive symptoms tend to post more often. However, people who have extremely high symptom levels should post less due to low energy levels.

Besides papers that have adopted a classification approach to infer depressive symptoms, there are a total of 9 papers involved in textual analysis of the social media posts. These papers focus on observing how social media users share their distress and daily life experiences. For example, Lachmar et al. (2017),Tian et al. (2018) and Michikyan (2020) identified themes from posts generated by users with depression. Ophir et al. (2017) hand-annotated cognitive distortion in the social media text. Bathina et al. (2020); Nambisan et al. (2015) used a lexicon-based approach to identify cognitive distortion and rumination.

The type of proxy signals used to infer users' depressive symptom levels are slightly different among papers that adopted classification methods and papers that focus on textual analysis (see Table 2.3). Studies focusing on textual analysis have identified topics related to one's thinking process and life situations. For example, aggression and a feeling of worthlessness (Marinelarena-Dondena et al., 2017; Thorstad and Wolff, 2019), a feeling of loneliness (Schwartz et al., 2014), suicidal thoughts (Shen et al., 2013; Cavazos-Rehg et al., 2016; Nguyen et al., 2014; Schwartz et al., 2014) and homesickness (Resnik et al., 2015). Table 2.5 lists the topics that have been identified as important features in each work that involves manual textual analysis.

In contrast, machine learning algorithms are not ideal for identifying cognition-related topics because these topics are largely context-dependent. Topics identified with machine learning approaches are mainly related to somatic symptoms, biological words, and medications. These topics can be easily captured by spotting the nouns in

the text.

Some of the signals related to one's thinking process or life situations are important to the onset or perpetuation of depression. For example, cognitive distortions, life challenges, sadness, substance abuse and suicide ideation (see Table 2.3). In Table 2.3, we list the social media signals identified with manual textual analysis and the corresponding theories from psychopathology that suggest these links.

Representing and extracting these signals from social media text has the potential to help researchers to understand the development of the illnesses with a theoretical background. There have been studies exploring approaches to detect of cognitive distortion (Simms et al., 2017), life events (Di Eugenio et al., 2013), support (Andalibi et al., 2017) and suicidal ideations (Varathan and Talib, 2014) from social media text.

2.3.3.1 Important features differ across platforms

Of all the studies in this review, 22 of them use Facebook data, 35 use Twitter, 26 use Reddit, 7 use Weibo (Shen et al., 2018; Tian et al., 2018; Wang et al., 2013a; Hu et al., 2019; Chen et al., 2020a; Hu et al., 2015; Wang et al., 2020), 4 use Instagram (Mann et al., 2020; Reece and Danforth, 2017; Ricard et al., 2018; Chiu et al., 2020), 4 use other data sources and 7 papers trained models on multiple sources and compared their performances (Lin et al., 2014; Tai et al., 2015; O'Dea et al., 2018; Seabrook et al., 2018; Aldarwish and Ahmad, 2017). The majority of the studies were conducted on datasets in English. Only 16 studies examine a foreign language, of which seven papers studied Chinese language data. The lack of language varieties in these studies indicates a strong representation bias presented in the existing literature.

Figure 2.5 shows the percentage of papers that reported a specific type of social media signals as an important feature to infer depressive symptoms across various social media platforms. Studies using Facebook and Twitter data are most likely to report the affective words as an important feature (De Choudhury et al., 2013; Yang et al., 2020; Tsugawa et al., 2015; Shen et al., 2017), whereas studies using Reddit data are least likely to report affective features as important. We believe the difference is related to the platforms' functionalities.

Kietzmann et al. (2011) identify seven functional building blocks in multiple social media platforms: identity (users reveal their identity), conversations, sharing, presence (the extent to which users can know if other users are accessible), relationships, reputation, and groups. Functionalities vary across platforms, and users' motivations for posting were influence by the functionalities. The motivations behind posting can be

Table 2.3: Proxy signals reflect social media users' depressive symptoms (papers focus on textual analysis)

Social media signals	psychology literature
cognitive distortion and depressive rumination (Ophir et al., 2017; Nambisan et al., 2015; Bathina et al., 2020).	(Beck, 1991, 2019)
explicit reference to depressive symptoms (Ophir et al., 2019; Tian et al., 2018).	~
lifestyle challenges (e.g., no appetite), relationships, social struggles or rejections (Michikyan, 2020; Lachmar et al., 2017; Ophir et al., 2019).	(Gore et al., 1993; Monroe and Harkness, 2005; Monroe et al., 2009; Slavich et al., 2010)
apathy, sadness and negative experience (Lachmar et al., 2017; Tian et al., 2018; Michikyan, 2020; Ophir et al., 2019).	(Bowlby, 1998; Brown and Siegel, 1988)
seeking positive relief (e.g., seek support) (Lachmar et al., 2017; Tian et al., 2018).	~
seeking negative relief (e.g., substance abuse) (Lachmar et al., 2017).	(Levy and Deykin, 1989; Abraham and Fava, 1999)
transitions in life (Michikyan, 2020).	(Praharso et al., 2017)
suicidal ideations (Ophir et al., 2017; Lachmar et al., 2017).	(Levy and Deykin, 1989; Orsolini et al., 2020)
share medical information (Tian et al., 2018).	~

categorized as image-related (e.g., "I want to be famous") or intrinsic-related (e.g., "it is fun", "I need support") (Kankanhalli et al., 2005; Wasko and Faraj, 2005). Users from mental health support communities focused on sharing, conversations, and fostering groups (Zhang et al., 2017; De Choudhury and De, 2014). They are more likely to discuss problems that they didn't feel comfortable with (De Choudhury and De, 2014; Johnson and Ambrose, 2006). However, they leave the community when they didn't have any experiences to share (e.g., problems are resolved) or when they were unable to form strong connections with the community (Wasko and Faraj, 2005). Users' retention rate strongly depends on the collective identity the community fosters (Zhang et al., 2017).

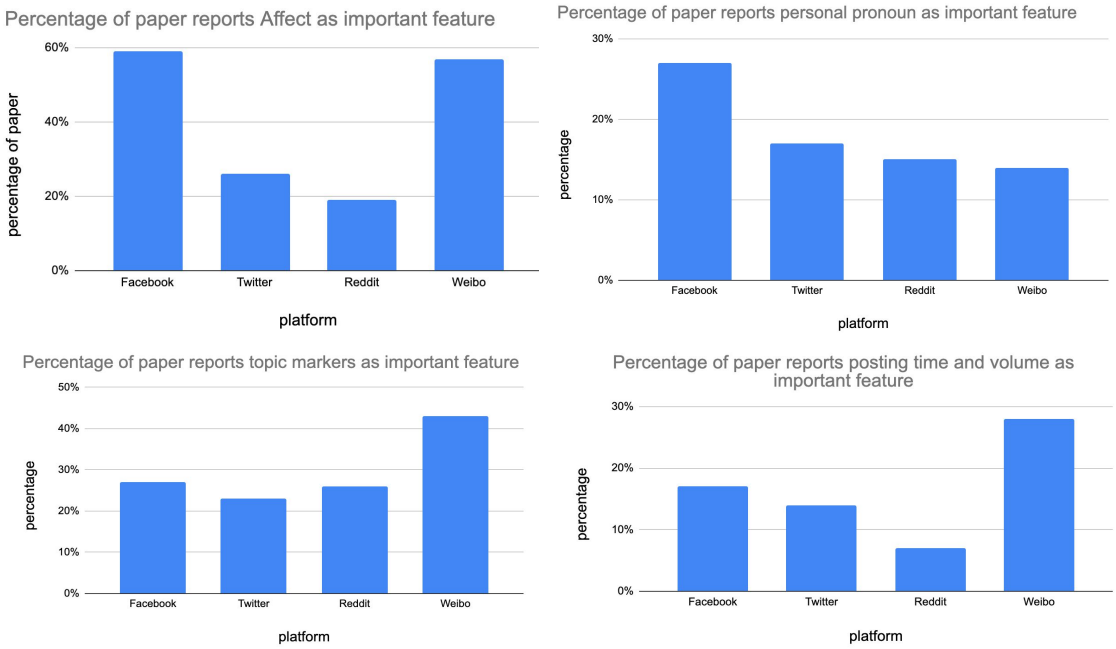
Since users' motivations for joining mental health support communities are mainly sharing and seeking support, most of the users tend to be less active once their problems were resolved. Therefore, users' posting history in a single mental health subreddit is usually sparse. For example, (De Choudhury and De, 2014) study the posting behavior in multiple mental health related subreddits. They find that half of the users posted less than one post within a range of two months. On average, users post about 1.5 posts in r/SuicideWatch since joining the subreddit (Chen et al., 2019). Researchers cannot aggregate mood based on users' posts in a single subreddit. Using affect extracted from one or two posts to infer users' mental health status, especially affective disorder, is not ideal.

On the contrary, researchers may be able to extract the pattern of affect from users' microblogging because microbloggers post more frequently. Microbloggers often focus on sharing information and building connections with others. Many microbloggers post content regularly to form connections and social reputations. For example, Toubia and Stephen (2013) find that microbloggers who started with a low initial number of followers and gained more followers later tend to create more content. Therefore, microbloggers are much more likely to create longitudinal social media records than mental health support communities. Researchers can aggregate the affect extracted from microblogging to estimate the mood of the users (Chen et al., 2020d; De Choudhury et al., 2013).

Other types of features also differ across platforms due to the functional characteristics of the platform. For example, social network features are often examined in platforms with social ties, and image features are often examined in microblogging (Huang et al., 2019; Shen et al., 2018; Guntuku et al., 2019; Huang et al., 2019). Some features, such as personal pronouns and topics, are commonly reported as important

features to infer mental disorder symptoms across different platforms (see Figure 2.5).

Figure 2.5: Percentage of papers report certain type of social media signals as important feature



2.3.4 Techniques for Feature Engineering

Affect and topics are most frequently reported as an important category of features to infer depressive symptoms with a classification approach. This section analyzes the current techniques applied to extract affect and topics in the social media text.

Affect Researchers measure mood in social media text using sentiment (feeling) as a proxy (Reece et al., 2017; Chen et al., 2018). Sentiment analysis is a natural language technique to extract opinions, feelings and contextual information from text. Methods of sentiment analysis include supervised learning, in which the algorithm learns the sentiment features from labeled datasets. Another approach is a lexicon-based method, where researchers label the words which express opinions and feelings. The predefined sets of words are used to assign sentiment score to text (Sharma et al., 2020).

As we mentioned in Section 1.2, mood refers to feelings that run in the background. However, sentiment as opinion or feeling is a reaction to a situation. To construct a variable representing a general feeling, we can aggregate the sentiment within a time

window. A few works used aggregated sentiment in a sliding time window as a feature in the mental status predictive techniques (De Choudhury et al., 2013). However, most of the works have used averaged sentiment over a large time window (e.g., 1 year) as a feature (Benton et al., 2017; Chen et al., 2018; Tsugawa et al., 2015; Wang et al., 2013a; Leis et al., 2019; Tong et al., 2019; Saravia et al., 2016).

Extracting topics from text Topics are usually identified with clustering approaches (e.g., Latent Dirichlet Allocation (LDA)) or lexicon-based approach, such as Linguistics Word Count and Inquiry (LIWC) (Tausczik and Pennebaker, 2010) and Empath (Fast et al., 2016). LDA topic modeling is a commonly used tool to model the latent structure of a collection of text documents (corpus). Blei et al. (2003) propose that each document can be represented as a mixture of a small number of topics, and distribution of words characterizes each topic. The document collection (corpus) can be described as a distribution over the latent topics (Maier et al., 2018).

LIWC is a type of computerized text analysis that counts words in meaningful psychological categories (Pennebaker et al., 2001), such as emotions and cognition. LIWC contains a human-curated dictionary that includes more than 60 topics related to psychological processes. The algorithm assigns topic scores to the document by counting the presence of vocabularies in each document. LIWC is able to detect meaning in a wide range of context, including the following categories described in Pennebaker et al. (2001)'s work:

- attentional focus: pronouns and verb tense
- emotionality: positive and negative emotions
- social relationships: words provide information about who has more status, whether a group is working well together and quality of a relationship.
- thinking styles: conjunctions, nouns, verbs and cognitive mechanisms
- individual differences: self-focus, cognition complexity, social references and other cues that help to identify individual differences.

Empath is similar to LIWC except that it uses neural embeddings to discover new related terms, in addition to a small set of seed words in a category. The new related terms were validated with a crowd-powered filter. Emapth's categories were highly correlated with categories in LIWC (r = 0.906)(Fast et al., 2016).

A large amount of literature has shown that topics extracted with LIWC and Empath were both associated with users' self-reported depressive symptom levels (De Choudhury et al., 2013; Tsugawa et al., 2015; Reece et al., 2017).

2.3.5 Conclusion

This overview explains that the binary classification approach is the most commonly applied technique to infer depressive symptoms. Fewer than 10% of the studies adopted qualitative or textual analysis. Affect, ego-network, posting frequency and volume, topics, and personal pronouns are the most common social media signals reported as important to predicting users' depressive symptom levels in classification. Topics and themes associated with cognitive processing, such as self-worthlessness, suicidal ideations and cognitive distortions were often identified in studies that focused on textual analysis.

These important key signals vary according to the types of social media platforms. Affective signals extracted from microblogging are more important for inferring depressive symptoms than affective signals from discussion forums. Finally, we identify that aggregating sentiment over time was the most commonly used approach to extract affective features from social media text. There is a potential to improve the representation of affective patterns so that their structures are more similar to mood defined in psychology literature. For topic features, clustering methods (e.g., LDA) and lexicon-based approach (e.g., LIWC) are often used to extract themes or topics in the social media text. There is a potential to expand the lexicon dictionary, thus enabling it to cover more psychological constructs or symptoms related to other types of disorders or risky behavior.

2.4 Research Gaps

The literature review in this section summarizes five types of features that were often used in machine learning pipelines to infer social media users' depressive symptoms. We identify the challenges and research gaps in the current literature: the affective feature is most frequently reported as an important feature to infer depressive symptoms. This finding echoed with the psychopathology literature, which suggested mood (a form of affective experience) is an important category of symptom for depressive disorders. However, the representation of affective patterns in existing studies that use

social media data does not reflect the structure of mood. To address this challenge, Chapters 3 and 4 focus on representing mood expressed in social media text.

Another important finding is that topics are also frequently reported as an important category of feature, especially topics that reflect somatic complaints, suicide, and medications. However, the current techniques on topic extraction mainly rely on lexicon-based algorithms such as LIWC. Lexicon-based algorithms may be good at capturing topics that can be reflected by single keywords, such as personal pronouns and somatic symptoms but can hardly capture complicated psychological processes, such as the cognitive process that reflects depressive symptoms. Therefore, in Chapter 6, we explore methods to identify cognitive distortion from social media text.

A number of studies find that cognitive distortion can be identified in social media text (Simms et al., 2017; Ford et al., 2019; Zogan et al., 2020). Ophir et al. (2017) suggest the cognitive distortions expressed in social media text were highly correlated with depressive symptoms in a sample currently undergoing therapy treatment. However, it is uncertain whether this finding generalizes well in other samples.

The lexicon-based topic approach does not cover topics that are specific to a certain type of mental disorder or risky behavior. In Chapter 7, we experiment with manually compiling a dictionary of topic features to predict suicidal risk.

Finally, all the existing studies focused on users' content and reposted content are often removed because they may not reflect the users' thoughts. There is also copy-and-paste content in the post. This content is usually lyrics or quotations. It is unclear whether this content also reflects users' well-being and personalities. In chapter 5, we explored whether lyrics and quotes also contribute to social media users' depressive symptoms.

Table 2.4: Social Media Signals Inferring Depression

Feature group	Components	Publications
Affect	positive and negative affect, mood, emotion	De Choudhury et al. (2013), De Choudhury et al. (2014), Yang et al. (2020), Tsugawa et al. (2015), Shen et al. (2017) ,Jamil (2017), Shen et al. (2018), Seabrook et al. (2018), Wongkoblap et al. (2018), Husain (2019), Eichstaedt et al. (2018), Merchant et al. (2019), Guntuku et al. (2019), Chen et al. (2020d), Ophir et al. (2017), Michikyan (2020), Ehrenreich and Underwood (2016), Reece et al. (2017), Fatima et al. (2018), Karmen et al. (2015), Benamara et al. (2018), Thorstad and Wolff (2019), Ramiandrisoa and Mothe (2020), Fatima et al. (2019), Tadesse et al. (2019), (Li et al., 2020), (Reece and Danforth, 2017), Mustafa et al. (2020), Leis et al. (2019), Tian et al. (2018), Hu et al. (2019), Chen et al. (2020a)
social network and social capital	replies, followees, ego-network	De Choudhury et al. (2013), De Choudhury et al. (2014), Yang et al. (2020), Wu et al. (2019), Tsugawa et al. (2015), Jamil (2017), Shen et al. (2018), Negriff (2019), Huang et al. (2019), Vedula and Parthasarathy (2017)
Personal pronouns	1st, 3rd person pronouns	De Choudhury et al. (2013), De Choudhury et al. (2014), Yang et al. (2020), Tsugawa et al. (2015), Jamil (2017), Shen et al. (2018), Negriff (2019), Huang et al. (2019), Vedula and Parthasarathy (2017), Wongkoblap et al. (2018), Eichstaedt et al. (2018), Ophir et al. (2017), Benamara et al. (2018), Tadesse et al. (2019), Leis et al. (2019), O'Dea et al. (2018), Ophir et al. (2019), Mann et al. (2020), Marinelarena-Dondena et al. (2017)
Topic specific linguistics markers	markers or topics of depressed mood, loneliness, suicidal thoughts and many others	De Choudhury et al. (2013), De Choudhury et al. (2014), Wang et al. (2013b), Schwartz et al. (2014), Tai et al. (2015), Nambisan et al. (2015), Ehrenreich and Underwood (2016), Ophir et al. (2017), Shen et al. (2017), Reece et al. (2017), Marinelarena-Dondena et al. (2017) Shen et al. (2018), Eichstaedt et al. (2018), Benamara et al. (2018), Tian et al. (2018), Chen et al. (2018), Tadesse et al. (2019), Merchant et al. (2019), Thorstad and Wolff (2019), Fatima et al. (2019), Hussain et al. (2020), Li et al. (2020), Mann et al. (2020), Shatte et al. (2020), Hosseini-Saravani et al. (2020), Tlachac and Rundensteiner (2020)
Sleep pattern Posting activities	posting time, posting volume	De Choudhury et al. (2014), De Choudhury et al. (2013), Shen et al. (2018), Shen et al. (2018), Shen et al. (2017), Benamara et al. (2018), Tian et al. (2018), Li et al. (2020), Wongkoblap et al. (2018), Wu et al. (2019), Husain (2019), Huang et al. (2019), Cacheda et al. (2019), Wang et al. (2013b), Tsugawa et al. (2015), Ehrenreich and Underwood (2016)

Table 2.5: Social Media Signals Inferring Depression

Other features	
Feature group	Publications
Question-centric content	De Choudhury et al. (2014)
visual features	(Huang et al., 2019), (Shen et al., 2018), (Guntuku et al., 2019), (Huang et al., 2019)
topic features subgroup	
health, biological words, somatic symptoms	Benamara et al. (2018), Ehrenreich and Underwood (2016), Shen et al. (2018), Schwartz et al. (2014), Tai et al. (2015), Nambisan et al. (2015), Marinelarena-Dondena et al. (2017), Eichstaedt et al. (2018), Merchant et al. (2019), Thorstad and Wolff (2019), Mann et al. (2020)
aggression, hostility	Schwartz et al. (2014), Wang et al. (2013b)
loneliness	Schwartz et al. (2014), Eichstaedt et al. (2018), Thorstad and Wolff (2019), Wang et al. (2013b)
self-worthless	Marinelarena-Dondena et al. (2017), Thorstad and Wolff (2019), Hosseini-Saravani et al. (2020)
suicide	Ophir et al. (2017), Reece et al. (2017), Schwartz et al. (2014), Nambisan et al. (2015), Wang et al. (2013b), O'Dea et al. (2018)
cognitive distortion	Lachmar et al. (2017), Bathina et al. (2020), Ophir et al. (2017), Peng et al. (2019)

Chapter 3

Identify Affective Symptoms from Social Media Text

Affective literature suggests that different aspects of affect, such as the magnitude, alternation, and categories, reflect on a person's tendencies for regulating emotions. These tendencies are symptoms and perpetrators of depression. Existing studies examine social media signals and depression have found the variation of affect reflects depressive symptom level (De Choudhury et al., 2013; Reece et al., 2017). However, the alternation of different affective categories is not yet studied in the context of social media data. In this study, we analyze the content originality and affect polarity of 4086 posts from 70 adult Facebook users contributed over two months. Social media data is usually sparse because users do not post content everyday. Existing studies often use zero or mean to represent the days when users did not post any content. Here we introduce a silence token to represent days when the user does not post any content. Our results show that more extrovert participants tend to post positive content continuously, and that more agreeable participants tend to avoid posting negative content. We also observe that participants with stronger depression symptoms posted more non-original content. We recommend that transitions of affect pattern derived from social media text and content originality should be considered in further studies on mental health, personality, and social media.

3.1 Motivation

Many people express rich moods and emotions in their social media posts. Psychologists use the word "affect" to describe these experiences of feelings and emotions.

Affect plays an important role in cognition (Gross et al., 1998) and wellbeing (Silvera et al., 2008). Therefore, affective expressions on social media text have emerged as a key variable for making inferences about users' personality traits (Bachrach et al., 2012; Golbeck et al., 2011; Farnadi et al., 2013) or mental health (De Choudhury et al., 2013; Coppersmith et al., 2014; De Choudhury and De, 2014; Bazarova et al., 2015).

Existing studies formulate the associations between affect and wellbeing based on the frequencies of affective words used in social media text (Schwartz et al., 2013; Yarkoni, 2010; Golbeck et al., 2011; Park et al., 2015; Chen et al., 2020a). However, patterns of affect are an important class of symptoms of affective disorders (Rottenberg, 2005; Frijda, 1993; Bylsma et al., 2011; Sheppes et al., 2015; Thompson et al., 2012; Houben et al., 2015; Carlo et al., 2012). Personality may also predispose individuals to specific moods (Rusting and Larsen, 1995; Rusting, 1998). With this in mind, we examined how patterns of affect expressed in social media text is related to with users' mental health and personality.

While non-original content has been extensively studied in opinion mining (Agarwal et al., 2011; Balahur et al., 2009), it has been comparatively neglected in the study of psychological interpretations of social media data. However, social media users often use lyrics or quotes to communicate their emotions. Such content comes from other media, such as literature, videos, films, or music, which can evoke strong emotional experiences (Juslin and Laukka, 2004; Scherer et al., 2001; Scherer, 2004). Since the affect of the non-original content may be different from the social media users' affect when they are post this content, we differentiated between original and non-original content in our analysis.

This pilot study was designed to examine the following research questions:

1. **Changes in Affect:** To what extent do changes in the affect of social media posts correlate with users' personality traits and mental wellbeing?

2. **Originality:** To what extent does the use of non-original material in their posts correlate with users' personality traits and mental wellbeing?

Following best practice in sentiment analysis and opinion mining, we distinguish between positive, negative, neutral, and mixed (both positive and negative) affect (Moilanen and Pulman, 2007; Rosenthal et al., 2015; Agarwal et al., 2011).

We used a well known dataset, myPersonality (Bachrach et al., 2012; Youyou et al., 2015), that enriches Facebook posts with many validated psychological measures. In MyPersonality, positive mental wellbeing is measured using the Satisfaction

with Life Scale (Diener et al., 1985, 1999), while the presence of depressive symptoms is assessed using the Center for Epidemiologic Studies Depression scale (CES-D) (Radloff, 1977). Personality traits are established following the OCEAN model (McCrae and John., 1992), which consists of the five traits Openness to Experience, Conscientiousness, Extroversion, Agreeableness, and Neuroticism.

We included all 70 adult users who provided sufficient, regular Facebook data for two months before completion of the CES-D questionnaire, and corrected for multiple comparisons in our statistical analysis. We find that the transitions from one affective state to another expressed in social media posts give us a highly nuanced view of personality traits. While the amount of non-original posts in ones' social media status updates is closely linked to depression symptoms, this link is mediated by neuroticism.

3.2 Background

Affect refers to both mood and emotion. Moods are slow-moving states that can be influenced by people, objects or situations, whereas emotions are quick reactions to stimuli (Rottenberg and Gross, 2003; Watson, 2000), and highly situation- or object-specific (Bylsma et al., 2008). Mood influences the probability of having emotions of the same valence—negative mood facilitates negative emotions, and positive mood makes positive emotions more likely (Rottenberg, 2005; Fredrickson, 1998). Affect is is an important predictor of mental wellbeing, including a person's overall satisfaction with life (Headey et al., 1993; Singh and Jha, 2008; Chen et al., 2017), and the level of symptoms of depression (Tsugawa et al., 2015; Coppersmith et al., 2015; Resnik et al., 2015).

Personality also predisposes people to certain affective states (Rothbart et al., 2000). While neuroticism is associated with negative affect (Pishva et al., 2011), positive affect is strongly linked to extroversion (Watson and Clark, 1997; Fujita et al., 1991). Extroverts experience more positive affect because they engage in more social situations (Diener and Emmons, 1984; Ryan and Deci, 2001). Individuals who score high on agreeableness have a greater ability to regulate negative affect (Meier et al., 2006; Haas et al., 2007). This relationship between affect and personality is also reflected in social media studies (Lin et al., 2017; Golbeck et al., 2011; Schwartz et al., 2013; Pennebaker and King, 1999). For example, people who use negative affective words in their social media posts tend to have lower conscientiousness, lower agreeableness (Golbeck et al., 2011), and higher neuroticism (Pennebaker and King, 1999).

In psychology, quantitative representations of affect are typically multidimensional (Russell, 1980). In this study, we focus on valence, which is represented in many classic affect models. Traditional measures, such as the Positive and Negative Affect Schedule (PANAS) (Watson et al., 1988), report the strength of positive and negative valence. Mixed valence can occur when people experience 'dialectic' emotion, which is a mix of positive and negative emotions (Russell, 2003; Schimmack et al., 2002).

The personality trait measurements in myPersonality are based on Costa and McCrae's well-validated OCEAN model (McCrae and John., 1992). The model consists of five dimensions: extroversion, agreeableness, conscientiousness, neuroticism, and openness to experience. Neuroticism refers to the degree of emotional stability. Openness reflects the degree of creativity and curiosity. Conscientious individuals tend to be careful and diligent. Extroversion refers to a tendency to be energetic and friendly. Agreeableness reflects the tendency to be compassionate and to cooperate with others (Digman, 1990). The five-factor structure has proved to be robust in both self and peer ratings (McCrae and John., 1992), children and adult (Mervielde et al., 1995), across different cultures (McCrae and Allik, 2002), and stable over time (McCrae and John., 1992).

3.3 Data and Methodology

The myPersonality data set (Bachrach et al., 2012; Youyou et al., 2015) contains more than 180,000 Facebook users, enriched with a variety of additional validated scales (Bachrach et al., 2012). The collection of myPersonality data complied with the terms of Facebook service, informed consent for research use was obtained from all users, and researchers had to seek permission to use the dataset. Permission for the use of this database was obtained before it closed for new studies in 2018. The study was granted Ethical Approval by the Ethics Committee of the School of Informatics, University of Edinburgh.

3.3.1 Choice of Scales

From the extensive data collected within myPersonality, we chose two scales for quantifying mental wellbeing, the *Center for Epidemiologic Studies Depression Scale (CES-D)* and the *Satisfaction with Life Scale (SWL)*. The CES-D scale measures a key aspect of mental health, the presence of depression symptoms (Radloff, 1977). The scale

has high internal consistency, test-retest reliability (Radloff, 1977; Orme et al., 1986; Roberts, 1980), and validity (Orme et al., 1986). Following previous social media studies (De Choudhury et al., 2013; Park et al., 2012), we adopt a score of 22 or higher as a cut-off value for likely depressive disorder (maximum score: 60). The 5-item SWL scale has been tested across different cultures and age groups (Pavot and Diener, 2009) and has been found to have high internal consistency and temporal reliability (Diener et al., 1985). Personality traits were measured using a 100 item scale using items from the open source International Personality Item Pool (Goldberg et al., 2006) that were validated against the the NEO-PI-R (Schwartz et al., 2013) instrument.

3.3.2 Selection of Participants

The data set was originally designed for a study of the effect of mental wellbeing and values on social media disclosure. We therefore selected only those participants who had completed the CES-D scale, the SWL scale, and the Schwartz Value survey (Schwartz, 1992) in addition to the full personality questionnaire. 301 participants in myPersonality provided full data for all four scales.

To ensure we had enough posts to assess the frequency of affect transitions, we only included users in our sample that regularly updated their public Facebook feed (*regular users*). We defined regular users as individuals who posted on average twice a week or more. We estimated posting frequency using the average post count per day during the sampling frame. If an individual had a post count per day of 0.3, this individual made around 110 posts in 365 days, which was roughly equivalent to an average of 2 posts per week. Of the original 301 participants, 122 (40.5%) were regular users.

Since the CES-D asks about symptoms in the past week, we excluded a further 31 users who had not posted any content in the week before completing the CES-D scale. We then focused on a 60-day span (two months) before CES-D completion, to ensure we had sufficient data to track the development of users' moods. We removed 14 users who contributed less than 20 posts during that time. Finally we removed four users who were under 18 year old and three users with more than 20% of the posts written in a language other than English, because English was the common language of the annotation team. The final sample consisted of 4086 posts from 70 users.

3.3.3 Corpus Annotation

3.3.3.1 Social Media Affect

For the purpose of this study, we refer to the affect shown in social media posts as *social media affect*. In this study, we operationalize valence as the post author's attitude towards a primary target of opinion, following (Mohammad, 2016). We refer to the 'dialectic' affective state as *mixed valence*. If there is no clear trend towards positive or negative affect, the associated valence is *neutral*.

After extensive piloting, we created an annotation guideline (available as part of the supplementary material) that was largely based on (Mohammad, 2016)'s work on defining the valence of a social media post. Each post is assigned one of four affect polarities: + (positive), - (negative), $\pm$ (mixed), or 0 (neutral). We used manual annotation since this is commonly used in computational linguistics to create a baseline gold standard data set for further analysis (Teufel, 1999).

Out of the 4086 posts, 2698 (66%) were annotated by a team of six trained annotators and 1185 (29%) by the first author. 5% of all posts were annotated by all seven annotators to establish inter-rater reliability, which was measured using Cohen's κ (Gamer et al., 2012). Average inter-rater reliability between the first author and the annotators is 0.88, and 0.78 among the six annotators.

After annotation, most of the posts were of positive valence (N= 1588, 39%), followed by negative valence (N=1164, 28%), neutral valence (N=982, 24%) and mixed (N=312, 8%). 40 posts were excluded from analysis, since they did not contain English text.

3.3.3.2 Originality

We define posts that consist of quotes from sources such as song lyrics, books, or movies as non-original content; all other content was defined as original. Since non-original content might not directly reflect the user's moods or emotions, annotators were instructed to annotate such posts according to the likely emotions of the author. For example, if a post consists of an uplifting motivational quote, annotators considered the underlying valence to be positive.

In order to establish the originality of a post, we retrieved the first page of results obtained by searching for the post text using the Google API. For each web page on the first page of results, we computed the cosine similarity between the the post content and the page content. Posts with a cosine similarity greater than 0.96 were labeled

as non-original, and posts with a cosine similarity between 0.92 and 0.96, where the website links or website names included the words 'lyrics' or 'quote' were labeled as potentially non-original. Posts with a cosine similarity lower than 0.92 were labeled as original. The cutoff points were determined based on a sample of 300 posts manually annotated for originality by the first author. On these posts, the classifier yields 100% recall, 81% precision, and an F1-score of 0.89. In our data set, 287 (7%) of all posts were identified as non-original.

3.3.4 Modeling Affect Transitions

We examine two types of transitions:

- **Post-Level versus Day-Level:** *Post-level* transitions focus on changes in affect between subsequent social media posts, whereas *day-level* transitions focus on changes in overall dominant affect between subsequent days.
- **Silence versus Non-Silence:** Not all users post every day. In our *default* models, these silent days are ignored, whereas in our *with-silence* models, days without posts are explicitly modeled as *Silence*.

The post-level social media affect is likely to be influenced by *underlying emotions*, which change more quickly, whereas the day-level social media affect is likely to be influenced by *underlying mood* during the day. Day-level affect was calculated as follows. If the majority of the posts p_{ij} on day d_j have the same affect a, then the affect of day d_j is set to a. If there is an equal number of positive (+) and negative (-) posts, or if the number of mixed affect ($\pm$) posts is equal to the number of posts with other types of affect, affect is set to $\pm$ (mixed). For transitions between original and non-original posts, we only consider the post-level representation. Table 3.1 shows an example of the affect and originality representations.

3.3.5 Statistical Analysis

Demographic differences between users above and below the CES-D cut-off score for probable depression were assessed using Wilcoxon-Mann-Whitney tests (R-package 'Stats').

We used Pearson correlation coefficients to assess the significance of correlations between social media data on one hand and personality traits and mental wellbeing on

Table 3.1: Affect and originality representation for a sample week

	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Affect							
Post-Level	+ - -	+ - +	+ +	S	±	0	+ -
Day-Level	-	+	+	S	±	0	±
Originality							
Post-Level	O N O	O O N	N N	S	O	O	N N

Note: $\leftrightarrow$, negative valence: $-$, positive valence: $+$, mixed valence:$\pm$, *S*: silence day, original content: O, non-original content: N

the other hand. Due to the small sample size and the number of correlations computed, all correlation coefficients were estimated using a permutation approach (Higgins, 2003) as implemented in the R Package `jmuOutlier` (Garren, 2017). Correlations that reach $p < 0.01$ or better are reported as significant; correlations that reach $p < 0.05$ are reported as trends in the data. For all correlations reported in the paper, we give the estimated correlation coefficient, the bootstrap 95% confidence interval, and the corresponding coefficient of determination r^2.

3.4 Results

3.4.1 Demographics and Baseline Statistics

Table 3.2 shows the basic statistics of our sample. Our data predominantly comes from single female Caucasian young adults. The average CES-D score is above the cut-off for possible depressive disorder.

Thirty-nine (56%) participants had a CES-D score of 22 or higher (mean: 33, SD: 6.5), which means that it is possible that that they have depressive disorder, and 31 (44%) had a score of 21 or lower (mean: 12, SD: 6). Figure 3.1 Plot 1 shows the density distributions of personality trait and SWL scores for three groups, the full sample, people above the cut-off, people below the cut-off..

Participants with possible depressive disorder are less extroverted (Z = 375, p<0.005), have higher levels of neuroticism (Z = 990, p<0.001), lower levels of conscientiousness (Z = 375, p<0.001), and lower satisfaction with life (Z = 323, p<0.001). Detailed results are reported in Figure 3.1 Plot 2.

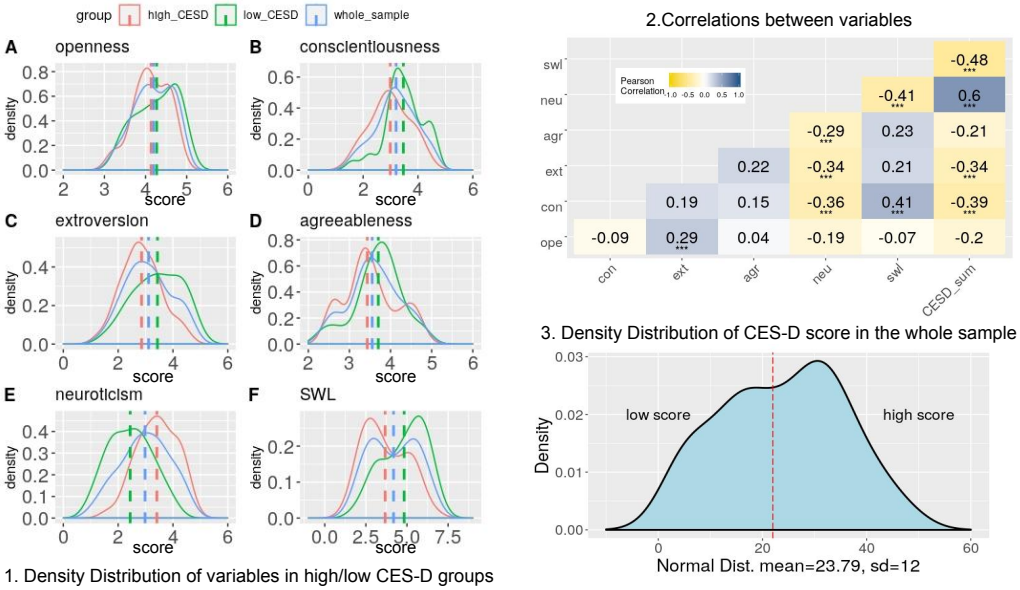


Figure 3.1: Basic statistics for personality trait scores, SWL and CES-D scores. Plot 1 shows density plots of the distribution of of personality traits and SWL for all participants, participants with CES-D $>=$22 (high CES-D), and participants with CES-D $<$ 22 (low CES-D). The dotted line shows the median. Plot 2 is a heat map of correlations between personality traits, SWL and CES-D scores (***: $p < 0.001$). Plot 3 illustrates the distribution of the CES-score in the entire sample (N = 70). The dotted line indicates the cutoff score of 22.

Table 3.2: Demographics of the sample.

Variable	N (%)	Variable	Mean (SD)
Gender		*Age*	
- Female	49 (70%)	- Female	23.52 (6.56)
- Male	21 (30%)	- Male	22.84 (7.13)
Ethnicity		*Personality*	
- Caucasian	54 (75%)	- Openness to Experience	4.19 (0.46)
- Black	3 (4%)	- Conscientiousness	3.20 (0.75)
- Asian	5 (7%)	- Extraversion	3.11 (3.83)
- other	8 (14%)	- Agreeableness	3.55 (0.68)
		- Neuroticism	2.98 (0.89)
Living Status		*Mental Wellbeing*	
- Living with partner	8 (10%)	- SWL	4.18 (1.44)
- Single	54 (77%)	- CES-D	23.79 (11.86)
- Married	5 (7%)		
- Unknown	3 (4%)		

Note: Caucasian includes White people of American, British, and other origins; Black includes African Americans and Black people from Europe. SWL: score for Satisfaction with Life Scale. CES-D: Center for Epidemiologic Studies Depression Scale,

All scales are normally distributed (Shapiro-Wilks test), except for openness to experience (W=0.96, $p < 0.05$), and satisfaction with life (W=0.95, $p < 0.05$), which are bimodal. Figure 3.1 Plot 2 shows the correlations between different personality dimensions. As expected, the five personality dimensions are not orthogonal.

3.4.2 Social Media Affect: Frequencies versus Transitions

For **overall frequencies of affect category**, the only clear correlation is between extroversion and positive content. Overall, more extroverted participants are more likely to have days where they make predominantly positive posts (r=0.29, $p < 0.01$, 95%CI = (-0.15, 0.32), $r^2 = 0.08$). In addition, participants who score higher on agreeableness tend to post fewer negative posts and have fewer days with predominantly negative posts (both r=-0.26, $p < 0.05$, 95%CI = (-0.48, -0.04), $r^2 = 0.07$).

When we look at **transitions between affect categories**, however, a more nuanced picture emerges. Table 3.3 summarizes the correlations between personality, well being and transition types. Significant correlations are summarized in Table 3.4. Due to the number of correlations presented, we choose a cut-off of $p < 0.01$, which is stricter than the normal $p < 0.05$.

Several transition types are correlated positively and negatively with Extroversion and Agreeableness. Neuroticism, conscientiousness, and SWL show interesting trends ($p < 0.05$) that do not reach significance (c.f. Table 3.3

More extroverted participants are more likely to post predominantly positive content several days in a row (*day-level*, +↔+, r=0.30, $p < 0.001$, 95% CI = (0.06, 0.54), r^2=0.09). They have more transitions to or from a silence day with a positive post (*post-level with-silence*, S↔+, r=0.29, $p < 0.01$, 95% CI = (-0.01, 0.46), r^2=0.08). This pattern fits well the overall predominance of posts with positive affect. Extroverts are also less likely to alternate between days with neutral and days with non-neutral content (*day-level*, for both 0↔+ and 0↔-, r=-0.28, $p < 0.01$, 95% CI = (-0.52, -0.09), , r^2=0.08).

People who score higher on agreeableness are less likely to follow a post with negative affect with another negative affect post (-↔-, *post-level with-silence*: r=-0.37, $p < 0.001$, 95% CI = (-0.50, -0.06) , r^2=0.14); This tendency is much less pronounced on the day-level (-↔-, r=-0.22, $p < 0.1$, 95% CI = (-0.44, -0.02) , r^2=0.04). On top of that, they are more likely to alternate between days with mixed valence and silence (*day-level*, ±↔S, r=0.28, $p < 0.01$, 95% CI = (-0.01, 0.46), r^2=0.08, post-level with-

Table 3.3: Correlations between personality, SWL, and CES-D scores and affect transitions. Number of participants N=70

	S↔S	-↔-	+↔+	±↔±	0↔0	+↔-	±↔+	±↔-	±↔0	0↔+	0↔-	S↔+	S↔-	±↔S	S↔0
Post-level representation (Post Plus Silence)															
N_{Occ}	1238	346	542	29	230	599	143	134	100	424	414	641	384	137	211
ope	0.09	-0.17	-0.17	-0.16	-0.05	-0.14	-0.07	-0.08	0.11	0.01	0.03	0.17	0.00	0.13	0.03
con	-0.06	0.01	0.09	-0.09	-0.15	0.11	0.00	-0.01	-0.14	-0.07	-0.08	0.16	0.00	0.15	-0.15
ext	0.04	-0.12	0.16	-0.10	-0.19	-0.06	-0.03	-0.12	-0.09	-0.09	-0.17	**0.29****	-0.04	0.00	-0.18
agr	0.14	**-0.37*****	0.03	0.02	-0.15	**-0.22***	0.08	0.04	0.04	-0.04	**-0.23***	**0.23***	-0.04	**0.29****	-0.13
neu	-0.07	0.19	0.18	0.18	-0.03	**0.23***	0.11	0.04	0.02	0.05	-0.05	**-0.22***	-0.03	**-0.23***	0.13
swl	0.04	-0.10	-0.13	-0.10	0.06	-0.03	0.02	-0.05	-0.04	0.02	-0.08	0.02	0.16	-0.02	0.18
CESD	-0.04	0.19	0.08	0.09	0.00	0.04	0.15	0.07	0.03	-0.06	0.11	-0.20·	0.00	-0.11	-0.03
Post-level representation (Post only), N = 70															
N_{Occ}		396	694	34	313	728	188	166	142	547	502				
ope		-0.16	-0.05	-0.06	-0.02	-0.05	0.06	-0.01	0.14	0.09	0.13				
con		-0.07	0.18	-0.07	**-0.23***	0.08	0.14	0.10	-0.11	-0.13	-0.12				
ext		-0.04	**0.33*****	0.04	**-0.24***	0.05	0.08	-0.10	-0.15	-0.16	-0.20·				
agr		**-0.28****	0.18	0.00	-0.16	-0.10	**0.26***	**0.28****	0.13	0.03	**-0.26***				
neu		0.14	0.00	0.11	-0.02	0.16	-0.14	-0.09	-0.08	0.01	-0.12				
swl		0.00	-0.12	-0.11	0.11	0.02	0.09	0.09	-0.02	0.08	-0.04				
CESD		0.14	-0.04	0.03	0.04	-0.03	-0.06	-0.11	0.04	-0.11	0.13				
Day-level representation, N = 70															
N_{Occ}	228	281	271	267	304	287	303	296	298	261	311	242	259	261	261
ope	0.12	-0.17	-0.11	-0.05	-0.02	-0.08	0.00	-0.14	-0.07	-0.01	-0.02	0.12	-0.02	0.19	0.13
con	-0.06	-0.03	**0.25***	0.05	-0.01	0.03	-0.03	-0.04	-0.16	-0.19	-0.12	0.08	0.10	0.06	-0.07
ext	0.06	-0.11	**0.30*****	-0.03	-0.14	0.04	0.14	-0.13	0.01	**-0.28****	**-0.28****	**0.24***	-0.08	0.02	-0.17
agr	0.11	-0.22·	0.15	-0.05	0.08	-0.12	0.16	-0.06	0.11	-0.08	-0.17	0.15	-0.07	**0.28****	-0.09
neu	-0.08	0.16	0.00	0.19	-0.17	**0.21***	0.09	0.11	-0.01	0.12	0.08	-0.14	-0.12	**-0.26***	0.03
swl	0.02	-0.08	-0.01	-0.08	**0.25***	-0.03	-0.06	-0.10	0.03	-0.06	-0.04	-0.02	0.12	0.06	0.08
CESD	-0.03	0.11	-0.10	0.08	-0.18	0.02	0.10	0.08	0.08	-0.01	0.21·	-0.18	0.03	-0.16	0.05

Note: Pearson correlation P-value (permutation testing): $\cdot < 0.1$, * $< .05$, ** $< .01$, *** $< .001$, bidirectional transition types: $\leftrightarrow$, negative valence: $-$, positive valence: $+$, mixed valence: $\pm$, neutral: 0, silence day: S, N_{Occ}: number of occurrences of each transition type, ope: openness, con: conscientiousness, ext: extraversion, agr: agreeableness, neu: neuroticism, swl: Satisfaction with Life Scale, CESD: Center for Epidemiologic Studies Depression Scale

Table 3.4: Summary of the significant correlations between transition states and the five personality traits ($p < 0.01$)

	Transitions	Post-Level (with-silence)	Post-Level (without-silence)	Day-Level
Extraversion	S $\leftrightarrow$ +	$\uparrow$	—	—
	0 $\leftrightarrow$ +	—	—	$\downarrow$
	0 $\leftrightarrow$ -	—	—	$\downarrow$
	+ $\leftrightarrow$ +	—	$\uparrow$	$\uparrow$
Agreeablness	- $\leftrightarrow$ -	$\downarrow$	$\downarrow$	—
	$\pm \leftrightarrow$ S	$\uparrow$	—	$\uparrow$
	$\pm \leftrightarrow$ -	—	$\uparrow$	—

Note: $\downarrow$ indicates a significant negative correlation at $p < 0.01$ or better, $\uparrow$ indicates a significant positive correlation at $p < 0.01$ or better. — indicates that the correlation is not significant at this level. Bidirectional transition types: $\leftrightarrow$, negative valence: $-$, positive valence: $+$, mixed valence:$\pm$, neutral: 0, silence day: *S*.

silence, $\pm\leftrightarrow$S, r=0.29, $p < 0.01$, 95% CI = (0.08, 0.52), r^2=0.08) .

Participants with higher neuroticism tend to alternate between positive and negative content, but this is only evident when we take silence into account (+$\leftrightarrow$-, *post-level with-silence*: r=0.23, $p < 0.05$, 95% CI = (0.00, 0.47), r^2=0.04, *post-level without-silence*: r=0.16, 95% CI = (-0.08, 0.41), r^2=0.025, *day-level*: r=0.21, $p < 0.05$, 95% CI = (-0.46, -0.10), r^2=0.04).

There are interesting differences in transition patterns that incorporate information about silence days and those that do not. When disregarding silence days, we observe that people with higher conscientiousness or extroversion are slightly less likely to follow a neutral post with another neutral post (*post-level without-silence*, conscientiousness, 0$\leftrightarrow$0, r= -0.23, $p < 0.05$, 95% CI = (-0.41, -0.04), r^2=0.07; extroversion, 0$\leftrightarrow$0, r = -0.24, $p < 0.05$, 95% CI = (-0.41, -0.04), r^2=0.07).

When we take into account silence days for computing transitions, we find several more interesting trends. People who are more satisfied with life are more likely to follow a neutral post with another neutral post (0$\leftrightarrow$0, *day-level*: r=0.25, $p < 0.05$, 95% CI = (-0.01, 0.44), r^2=0.06). In addition, people with higher neuroticism are more likely to alternate between positive and negative posts (0$\leftrightarrow$-, *day-level*: r=0.21, $p < 0.05$, 95% CI = (-0.01, 0.40), r^2=0.04), but less likely to make a positive post after a period of one or more silence days (S$\leftrightarrow$+, *post-level with-silence*: r=-0.22, $p < 0.05$,

95% CI = (-0.48, 0.00), r^2=0.04). We found that silence to silence transitions are not correlated with personality or mental health.

3.4.3 Post Originality

High CES-D scores are significantly correlated with posting non-original content (r=0.29, $p < 0.01$, 95% CI = (0.10, 0.46), r^2=0.08). There is a similar tendency for participants with higher neuroticism scores (r=0.25, $p < 0.05$, 95% CI = (0.06, 0.43), r^2=0.07). Examining transitions between post originality shows that these effects stem from slightly different posting patterns. Users with higher CES-D scores tend to follow non-original content with non-original content (N↔N, *post-level with-silence*, r=0.26, $p < 0.05$, 95% CI = (0.07, 0.43), r^2=0.07) or to alternate between original and non-original content (N↔O *post-level with-silence*, r=0.27, $p < 0.05$, 95% CI = (0.08, 0.44), r^2=0.07). Users with higher neuroticism scores tend to post sequences of non-original content (N↔N, post-level with-silence, r=0.25, $p < 0.05$, 95% CI = (0.06, 0.43), r^2=0.05), and are less likely to post original content before or after a period of silence (O↔S, *post-level with-silence*, r=0.28, $p < 0.05$, 95% CI = (0.09, 0.45), r^2=0.08).

Since neuroticism is closely linked to depression symptoms, we also computed a partial correlation between content originality and CES-D while controlling for neuroticism. The resulting correlation was no longer significant (r=0.14, $p = 0.22$, r^2=0.02). Therefore, the association between content originality and depression symptoms might be moderated by neuroticism.

3.5 Discussion

3.5.1 Main Findings

Many studies have found associations between the frequency of affective words used in social media text and personality. However, existing studies often see affect as static and only focused on the strength of bipolar valence (positive/negative). Instead, our work focuses on affect patterns. We encode posting behavior, transitions between affect states, and content originality. From a practical point of view, our technique can supplement experience sampling techniques (Myin-Germeys et al., 2018) to help clinicians and patients develop a more comprehensive view of a person's affect patterns, arrive at a better substantiated diagnosis, and make improved treatment decisions. However, this depends on whether the patient is willing to share information

from their social media feed with their therapist.

Overall, the correlations seen between affect transitions and personality traits are in line with the consensus in the early literature (Gross et al., 1998). Extroverts tend to produce sequences of positive posts. This behavior fits well with the positive emotional core in extroverts stipulated in (Watson and Clark, 1997). Participants with higher agreeableness are less likely to post sequences of negative posts. This could be due to their ability to regulate negative affect (Meier et al., 2006; Haas et al., 2007).

Although the psychology literature suggests a strong association between negative mood states and neuroticism (Rusting and Larsen, 1995), we did not find this in our data. Our results are in line with previous studies of verbal cues to personality traits in social media (Schwartz et al., 2013; Yarkoni, 2010; Golbeck et al., 2011; Park et al., 2015). (Golbeck et al., 2011) found social media users who were more likely to talk about anxiety were on the higher end of the neuroticism scale. We speculate that self-presentation bias may influence how social media users regulate their expression of negative emotions in their public posts. The only relevant association we found was that social media users on the high end of neuroticism are more likely to switch between posting positive and negative affective content. This finding aligns well with the fact that high neuroticism is associated with high emotional instability (Costa and McCrae, 1992).

The link between posting non-original content and elevated depression symptoms appears to be moderated by neuroticism. This suggests that high levels of neuroticism predispose users both to depressive symptoms, and to an indirect disclosure of emotions through quotes and lyrics.

In our sample, the prevalence of depressive symptoms is higher than what would be expected in general population. In the original CES-D paper, Radloff (1977) proposed three levels of depression severity: low (0-15), mild-to-moderate (16-22), and high (23-60). They found that only 21% of the general population scored above the low symptom level. In contrast, in our sample, nearly half of the participants exhibit a high level of symptoms ($>$22). Within the context social media studies of depression, however, our data set is not exceptional. For many studies in the area, high symptom individuals account for nearly half of the data set (Reece et al., 2017; Tsugawa et al., 2015; Nadeem, 2016; De Choudhury et al., 2013; Husseini Orabi et al., 2018).

Our results support the claim that affect expressed in social media data text is associated with social media users' affect patterns in real life. However, the data set used in this study is from the early 2010's, and only covers the well established social me-

dia platform Facebook. The associations found in this study are likely to be slightly different from those found in another social network (e.g., Instagram) or in a new data set collected ten years later.

3.5.2 Limitations.

Due to the restrictions imposed by the need for sufficient Facebook updates to allow analysis, our final sample is relatively small. Given the size of the significant effects we found in the data, power calculations indicate that a well-powered study should include data from around 200 users (Schönbrodt and Perugini, 2013). It also skews heavily towards younger female Caucasians with relatively low satisfaction with life and strong depression symptoms. It is possible that other groups of users (e.g., non-Caucasians, males) are less likely to disclose personal information about mood and emotions on their public Facebook (Dosono et al., 2017; McDonald et al., 2019).

3.6 Conclusion

In this chapter, we conducted a pilot study to demonstrated the benefits of detailed representations of social media affect for unpacking the relationship between personality, mental wellbeing, and the content posted on social media. Importantly, our representations include non-binary affect categories (positive, negative, mixed, neutral), and take into account content originality. As a consequence, we were able to obtain a more detailed picture of the link between patterns of affect and depressive symptoms.

However, this sample size of this pilot study is too small to provide powerful effect size. In the next chapter, we enrich our data set with more in-depth analyses of original versus non-original content, extend coverage by including a larger sample of the myPersonality data set, and construct statistical models that allow us to observe long-term trends in posting patterns.

Chapter 4

Using Affective Patterns from Social Media to Infer Depressive Symptoms

In the previous chapter, I demonstrated representing the transition states of social media affect is beneficial to revealing the pattern of well-being and personality dimensions. We further ask the question, are these representations associated with depressive symptoms? In this chapter, we use vector and sliding window technique to construct mood representations. These representations capture the variation and alternations of mood. By using these representations to classify symptom levels, we have found that they are associated with social media users' self-reported depressive symptom levels. However, binary classification results are not very meaningful to researchers who study depression. We further use Gaussian Process Regression to monitor users' latent mood based on the mood representation. We observe less evidence of mood fluctuation expressed in social media text from those with low symptom measures compared to others with high symptom scores. Next, we leverage a daily mood representation in Hidden Markov Models to estimate binary latent states that influence the observed mood. We find these estimated latent states are associated with self-reported depressive symptom level. However, we also find presence of potential subgroups driving these findings. Our findings support the claim that for some people, derived mood from social media text can be a proxy of real-life mood, in particular depressive symptoms. Combining the mood representations with other proxy signals can potentially advance responsibly used semi-automatic screening procedures.

4.1 Introduction

Depression is the leading cause of disability worldwide. Initial efforts to detect depression signals from social media posts have shown promising results (De Choudhury et al., 2013; Coppersmith et al., 2014; Park et al., 2012; Tsugawa et al., 2015; Nguyen et al., 2014; Nadeem, 2016; Almeida et al., 2017). Given the high internal validity (Reece et al., 2017; De Choudhury et al., 2013), results from such analyzes are potentially beneficial to clinical judgement. The existing models for automatic detection of depressive symptoms learn proxy diagnostic signals from social media data, such as help-seeking behavior for mental health or medication names (De Choudhury et al., 2013; Coppersmith et al., 2014). However, in reality, individuals with depression typically experience depressed mood, loss of pleasure nearly in all the activities, feeling of worthlessness or guilt, and diminished ability to think (APA et al., 2013). Therefore, a lot of the proxy signals used in these models lack the theoretical underpinnings for depressive symptoms. It is also reported that the social media posts from many patients in the clinical setting do not contain these signals (Ernala et al., 2019). Based on this research gap, we propose to monitor a type of signal that is well-established as a class of symptom in affective disorders — mood. Mood is an experience of feeling that can last for hours, days or even weeks (APA et al., 2013). In this work, we attempt to enrich current technology for detecting symptoms of potential depression by constructing a "mood profile" for social media users.

The variance of quality and intensity of mood and emotional reactions are referred to as " affective style" (Davidson, 1998), which underlies one's risks of developing psychological disorders (Rottenberg and Gross, 2003; Akiskal, 1996). Assessing affective style in everyday life is difficult in an experimental context because it requires a costly extended period of data collection. In contrast, social media data contains longitudinal information that reflect one's emotional reactions to stimuli. Therefore, it can provide researchers with an alternative lens to examine the affective style of an individual, based on the premise that approval is obtained from social media users, and data privacy is well-protected.

Existing models for detecting symptoms of potential depression often include mood as a feature variable in the modeling process. However, there are a few methodological gaps in these models. First, most of them do not distinguish between mood and emotions. Emotion is a brief reaction to a specific stimulus, whereas mood has longer temporal duration (Morris, 2012). Researchers using social media data to study mood

or emotions often see a single post as reflecting mood (Bollen et al., 2011; Thelwall et al., 2011; Celli et al., 2016). However, a single social media post is likely to reflect a participant's emotions at the time rather than ongoing mood (Batson et al., 1992; Rottenberg, 2005). In this current work, we adopted the definition of mood from The Diagnostic and Statistical Manual of Mental Disorders (APA et al., 2013): "mood is the pervasive and sustained 'emotional climate', and emotions are 'fluctuating changes in emotional 'weather' ". We sought to determine whether temporal mood representation derived from social media text is associated with subsequent self-reported depressive symptoms, and if so, what are the best approaches to represent mood as a time dependent variable for future work?

Furthermore, a majority of models in this line of research often ignore the fact that affect is inherently time dependent. Only a few models have adopted temporal affective patterns (Reece et al., 2017; De Choudhury et al., 2013). Most of these models also formulate the associations between affect and depressive symptoms based on the averaged affect (Schwartz et al., 2013; Chen et al., 2020a), but the transitioning from one affective state to another was largely ignored (Rottenberg, 2005; Frijda, 1993; Bylsma et al., 2011; Sheppes et al., 2015). In this work, we explored and tested multiple approaches to represent the temporal affective patterns and the transitions of affective states.

Nevertheless, social media users often post sporadically. The sparsity of social media data posits a big challenge in the modeling process. Most of the existing studies imputed missing values with the mean or simply removed users with a lower word count (De Choudhury et al., 2013; Wang et al., 2013a). Removing outliers is beneficial to the modeling process. However, it may result in removing those with severe symptoms from the sample, because disinterest in social contact and social withdrawal (e.g., posting sparsely) is the core symptom of major depressive disorder (MDD) (APA et al., 2013). Therefore, it is necessary to use some modeling techniques to include the outliers.

Towards addressing the methodological gaps described above, we designed multiple mood representations with the following characteristics: (i) Temporal features (ii) Transitions from one mood state to another (iii) Posting behavior. Here we see all the mood representations as a *Mood Profile* for social media users. We formulate the following questions to explore the roles of mood in predicting depressive symptoms:

1. Are mood representations derived from social media text associated with the severity of self-reported depressive symptoms?

2. Which representation in the mood profile is most predictive of the severity of self-reported depressive symptoms?

Our main contributions in this study are:

1. Constructing a mood profile for social media users based on their status updates. The mood profile encompasses representations that encoded the variance of mood intensity, alternations of mood states and the behavior of not posting.

2. Examining the associations between the social media mood profile and users' depressive symptoms level.

3. Examining which representation in the mood profile is more predictive to depressive symptoms level.

In our work, we analyzed a set of 93,378 posts from 781 Facebook users who had consented to the use of their posts and answers to related questionnaires for research reasons. For each user, a mood profile is constructed based on their social media text. We found that people with low symptom level tend to have less fluctuations in their mood pattern. We also modeled the mood representation with a Hidden Markov Model and we found the hidden states estimated based on the mood representation is highly related to depressive symptoms. Nevertheless, combining several representations in the mood profile is more predictive to depressive symptom levels (f-score: 0.62) than using one representation only. Our results suggest the mood profile derived from social media text can potentially serve as a reference for an individual's depressive symptom level. The data-driven, evidential nature of our approach provides us with better insight into the relationship between mood derived from social media data and depression.

4.2 Background

4.2.1 Depression and Mood

Moods are slow-moving states of feeling, influenced by others, objects or situations (Rottenberg and Gross, 2003; Watson, 2000). The pattern of mood reflects one's vulnerability to developing affective disorders (Rottenberg, 2005; Rottenberg and Gross, 2003). Depressed mood is a symptom of mood disorders, such as major depressive disorder (characterized by a persistent feeling of sadness) and dysthymia (persistent mild depression) (APA et al., 2013).

It is also well established that mood fluctuation and irritability are associated with many somatic and sensory dysfunctions in the psychology literature. Frequent alternating between moods (typically a few days) and irregular cycles of mood underlie the behavioral features of a wide variety of conditions (Akiskal, 1996). In this study, we expect to find associations between mood derived from social media text and depressive symptoms similar to the psychology literature. Some level of associations has been found in the existing studies. For example, participants with depressive symptoms use more negative affective words (e.g., sad, cry, hate) in their social media text than those who do not (De Choudhury et al., 2013; Park et al., 2012).

4.2.2 Detecting Depressive Symptoms with Sentiment

Studies which examine emotions derived from social media data often adopt sentiment analysis. This is a computational process that categorizes affect or opinions expressed in a piece of text. The extracted affect is called sentiment (Pang et al., 2008). Most of the existing works use averaged sentiment over a long period of time (e.g., one year) as a feature to predict depressive symptoms (Coppersmith et al., 2014; Tsugawa et al., 2015; Benton et al., 2017; Park et al., 2012; Tsugawa et al., 2015; Wang et al., 2013a).

In addition to that, the change of sentiment over time is also an important aspect to infer affective disorders. However, only a few studies have included sentiment as a time dependent feature in the model (De Choudhury et al., 2013). For example, (De Choudhury et al., 2013) used the momentum of the feature vector in the screening detection. (Eichstaedt et al., 2018) include temporal posting patterns, but not the temporal affect pattern. (Chen et al., 2018) used temporal measures of fine grained emotions to predict users' depressive states. Recently, (Reece et al., 2017) adopted a Hidden Markov Model (HMM) to analyze the change of language in social media posts and users' depressive symptom. They found that the shift of words in status updates indicate depression and (expand) PTSD symptoms. The above mentioned studies adopted a sliding window technique to define dynamic sentiment (De Choudhury et al., 2013; Chen et al., 2018; Reece et al., 2017). However, none of them systematically explored the size of time window and the slide increment, and most studies only use a continuous sentiment value. In this work, we aggregated the sentiment in a sliding window based on its dominant valence (e.g., positive, negative) or average value. We also included the changes of affective states as a feature variable.

4.2.3 Posting Behavior and depressive symptoms

Social media users are known to communicate selectively due to self-presentation biases (Kim and Lee, 2011; Vogel et al., 2014). They are less likely to reveal events that project negatively on themselves (Mehdizadeh, 2010) due to stigma and fear of potential repercussions. Therefore, self-presentation biases leads to fundamental differences between real-life mood and social media mood.

In addition to that, social media behavior can be counter intuitive. For example, people with who are more depressed would be expected to post less than people with fewer symptoms, however, several studies found that individuals with a history of depression (determined from past medical history) tended to post more often compared with people without depression (Smith et al., 2017). There are several potential reasons for this. A person might not be severely depressed, they might be more comfortable with talking about their feelings, they might see their social media as a place where they can escape stigma, or they might have a social media support network for their mental health. In this study, we see the behavior of not posting as a variable in itself and observe if posting frequency has any predictive capacity with regards depressive symptoms.

4.3 Data

For this study, the myPersonality data set (Bachrach et al., 2012; Youyou et al., 2015) was used. It contains Facebook posts of 180,000 participants collected from 2010 to 2012, enriched with a variety of additional validated scales (Bachrach et al., 2012). The collection of myPersonality data complied with the terms of Facebook service, and informed consent for research use was obtained from all participants. Permission for the use of this database was obtained in 2018, and Ethical Approval for this piece of secondary data analysis was obtained from the Ethics Committee of the School of Informatics, University of Edinburgh. Other publications using this dataset include (Freudenstein et al., 2019; Sun et al., 2019).

4.3.1 Screening for Depressive Symptoms

From the participants in the myPersonality data, we focused on 1047 participants who completed the Center for Epidemiologic Studies Depression Scale (CES-D). The CES-D is a 20 item scale that measures the presence of depressive symptoms in the general

population (Radloff, 1977). It is one of the screening tests most widely used by health service provider. The symptoms measured in CES-D include mood, anhedonia, the feeling of being worried, restless, changes in sleeping pattern and physical symptoms (such as lost of appetite) and irrational thoughts. The scale has been found to have high internal consistency, test-retest reliability (Radloff, 1977; Orme et al., 1986; Roberts, 1980), and validity (Orme et al., 1986).

(Radloff, 1977) proposed three groups of depression severity: low (0-15), mild to moderate (16-22), and high (23-60). For using mood profile to predict self-reported depressive symptoms, we followed the practice from previous social media studies (De Choudhury et al., 2013; Park et al., 2012; Reece et al., 2017; Tsugawa et al., 2015) and adopted 22 as a cutoff point to divide participants into high symptoms and low symptom groups. This allows us to compare our model's performances with previous studies. For examining the mood fluctuation, we were additionally interested in a more nuanced picture in different symptom levels. Therefore, we further distinguish moderate and high symptom by following the original study from (Radloff, 1977). Participants were divided into three groups using two cutoff points: 16 and 22.

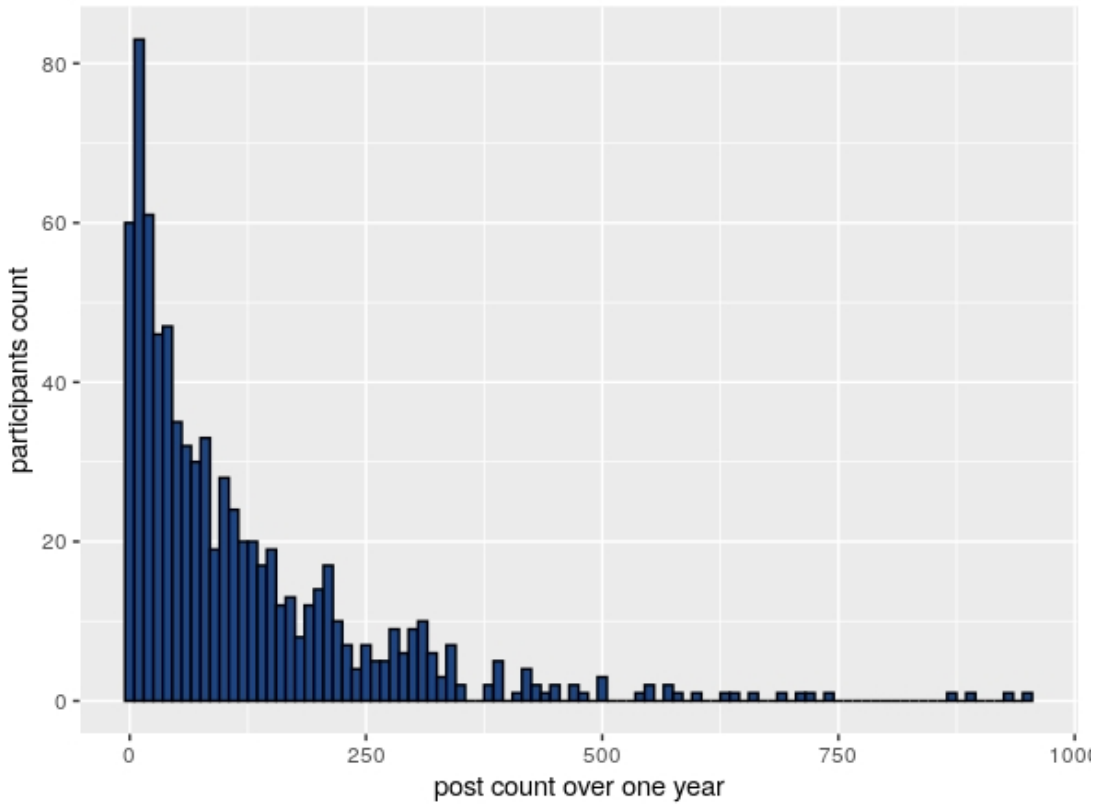
4.3.2 Summary Statistics

Among the 1047 participants who completed the CES-D scale, we removed 110 participants who were less than 18 years old. The CES-D survey was open from 2010 to December 2012, but MyPersonality only collected participants' status updates from January 2009 to December 2011. Since 2012 status updates were not available, we further removed participants who completed the scale in 2012 and who posted at least one post in the past year. Eventually we yielded a final set of 781 participants who had posted 93,378 posts over the past year before they took the test.

The average number of posts per user over one year was 120, this distribution was skewed by a small number of frequent posters, as evidenced by a median value of 73 posts per user. Figure 4.1 shows participants' count of posts up to one year before they completed the CES-D scale. The mean age of the participant is 26 (sd = 11.7), 333 (43%) participants are male and 448 (57%) are female. Table 4.1 shows further details of the participants, including the ethnicity, gender and marital status.

Overall, our sample has a relatively high mean CES-D score (m = 26.3, sd = 8.9), and the proportion of high symptom class to low symptom class is 1.6:1 (cutoff 22), see Figure 4.2. (Radloff, 1977) found only 21% of the general population scored at and

Figure 4.1: Distribution of post count from participants



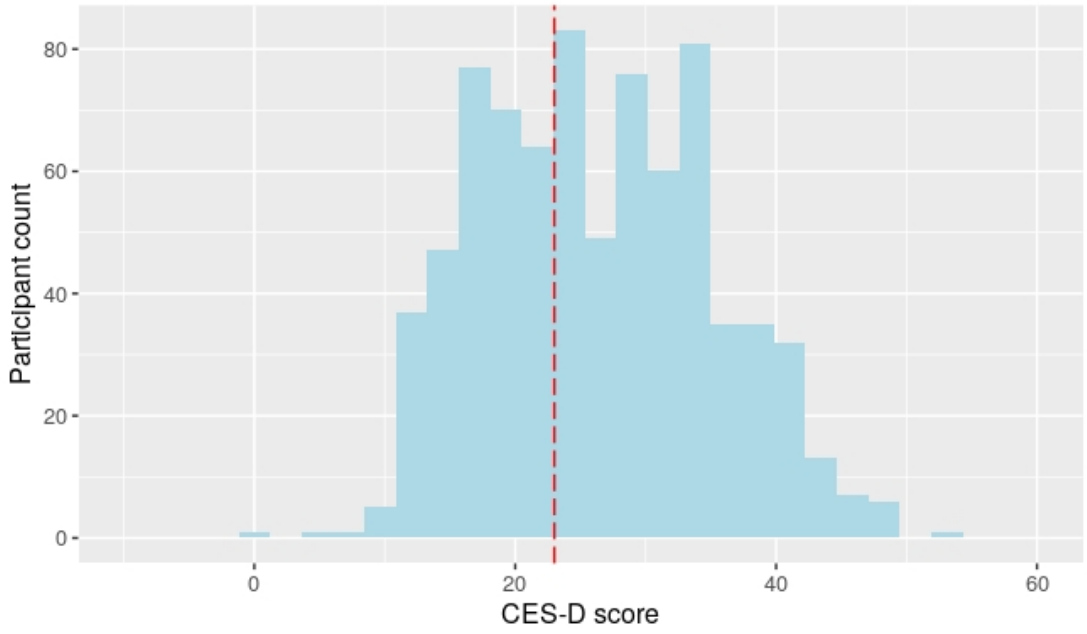
Note: Figure demonstrates the distribution of post count over one year before participants completed the CES-D survey scale. Size of the bin is 10.

Table 4.1: Demographic Information of the 781 Participants

Ethnicity	No.	%	Marital Status	No.	%
Black	38	4.3	Single	574	73.8
Asian Chinese	26	3.3	Divorced	28	3.5
Middle Eastern	13	1.7	Married	27	3.4
Native American	13	1.6	Married with Children	38	4
Other Asians	84	10.8	Partner	78	10
Not Specified	96	12.2	Not specified	36	4.5
White-American	309	39.2			
White-British	71	8.9			
White-Other	131	17.1			

above an arbitrary cutoff score of 16. However, we note the current dataset is not an exceptional case. For example, (Reece et al., 2017) used a dataset that contained 105 depressed participants and 99 non-depressed participants, other studies have a proportion of high symptom to low symptom class as 2:3 (De Choudhury et al., 2013; Tsugawa et al., 2015; Nadeem, 2016), 3:5 (Husseini Orabi et al., 2018). All of these studies recruited a sample biased towards potentially high symptom individuals compared with empirical studies which selected participants in a random trail. We speculate that there is a bias in those individuals self-selecting for this type of research.

Figure 4.2: Distribution of CES-D score



Note: Figure demonstrates the density distribution of the CES-D score, red line indicate the cutoff point 22

4.4 Constructing Mood Profile

A mood profile is constructed for each participant. Each mood profile encompasses sets of features which represent mood, the change of mood and the transition of mood states. Since mood is time dependent, we use a sliding window technique to construct the temporal features. A window starts from day 0 (the day when users completed the CES-D scale) and moves backwards for up to one year. Choosing the size of a time window presents a challenging question, how granular should a time window be? (De Choudhury et al., 2013) look at a user's tweets in a single day. (Reece et al., 2017)

use both day and week as the time window because most of the participants did not generate enough daily content. In this paper, we define the size of the time window as measured by day $d \in D := \{3, 7, 14, 30\}$, see Table 4.2 for the notations. The size of the slide increment determines how much information the two adjacent windows share. The slide increment is also measured by day $s \in S := \{3, 7, 14\}$.

Another challenge is to decide how far back do social media posts indicate symptom level. Earlier studies use data up to one year before participants completed the self-reported symptom measurement (De Choudhury et al., 2013), (Reece et al., 2017) found that symptoms can be predicted up to nine months before the official disclosure of the illness. In the current work, each representation in the mood profile was constructed with posts written up to one year before the participant completed the CES-D survey.

4.4.1 Sentiment Scores

We used the sentiment scores retrieved from SentiStrength (Thelwall et al., 2010). SentiStrength extracts sentiment from the text based on a function that describes how well the words and phrases of the text match a predefined set of sentiment-related words.

4.4.2 Temporal Mood Representations

Since many social media users do not post every day, we encoded the behavior of not posting as "Silence" and we defined four mood states: positive, negative, neutral and silence. We adopted two approaches to define mood within a time window: most frequent mood state over a time window and average sentiment over a time window, see Table 4.2. If two mood states had the same high frequency in the same time window, we defined the mood as mixed. Since neutral mood state is relatively less frequent in compare with the rest of the mood states, we tend to give neutral more weights. If other mood states have the same frequencies as neutral, we defined the mood as neutral. For the average sentiment, silence days as missing values are imputed by the mean. We also constructed features that represent the change of mood (De Choudhury et al., 2013), see mood momentum in Table 4.2.

Table 4.2: Notations for Mood Profile

Variable	Notation	Description
Window Size	d	A period of time within $d \in D$ days, $d \in D := \{3,7,14,30\}$
Slide Increment	s	A sliding window move forward by every $s \in S$ days, $s \in S := \{3,7,14\}$
Sentiment	v	Sentiment score of a single post
Day Sentiment	V	Arithmetic mean of sentiment in one day $V = \frac{v+...v_i}{i}$
$Mood_\mu$	M_μ	Arithmetic mean of day sentiment over a time window, $M_\mu = \frac{V+...V_d}{d}$
$Mood_\omega$	M_ω	Most frequent sentiment over a time window, categorical
Mood Momentum	ΔM	Difference between M_μ in two time windows
Mood States Transition	Tr	The probability of a user transfer from one mood-state to another, a mood state is defined by M_ω
Mood States Transition	ΔTr	Difference between Tr in two time windows

4.4.3 Temporal Mood Transition Representations

We also encoded the probability of a user transferring from one mood state to another as a representation in the mood profile. We have in total 16 transition states (e.g., positive to negative, negative to silence) from the four classes (positive, negative, neutral and silence). Note that if we set the slide increment as one day, we would have 365 $\times$ 16 mood transitions features. To prevent the large dimensionality, which might led to sparse representation, we defined d as 30 and s as 30, so that we have 12 $\times$ 16 feature columns for Mood Transition Representations.

4.5 Association Between Mood Profile and Depressive Symptoms

We first observed whether the pattern of the mood profile is related to symptom level. Then we tested the mood profile's predictive power on symptom level.

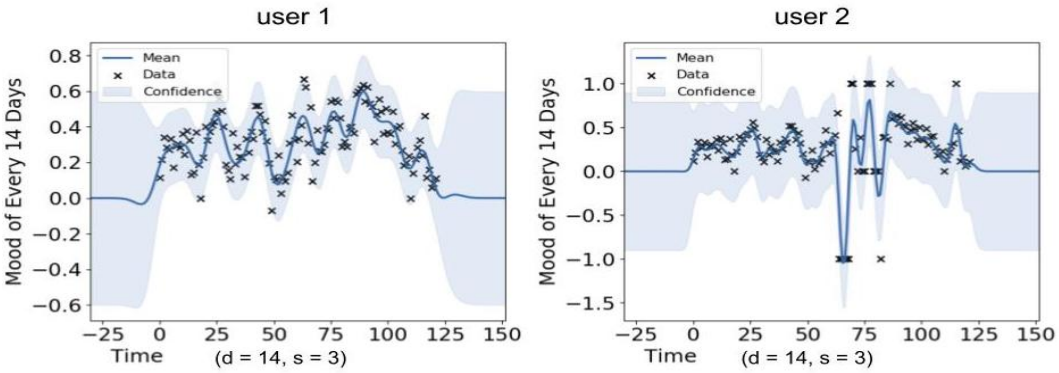
4.5.1 Mood Fluctuations

We modeled mood fluctuations using Gaussian Process (GP) regression. GP regression is a Bayesian approach that assumes a Gaussian process prior over functions (Quiñonero-Candela and Rasmussen, 2005). In this analysis, we see the temporal mood representations as noisy representations of participants' mood. We use GP regression to estimate participants' latent mood based on their mood representation. For participants with few data points, the GP regression is modeling the mean of the sample due to the imputation approach we adopted. Thus, for this experiment, we excluded participants who posted less than 10 posts over year before they completed the depressive symptom scale. Eventually, this yielded 690 participants for the current analysis. We used mood representations with $d \in D := \{1,3,7,14\}$ and $s \in S := \{1,3,7\}$ as input of the GP regression model. The GP regression is best fitted on mood vector with $d = 14$ and $s = 3$, see Figure 4.3. Each dot on the graph represents mood (averaged sentiment) in a time window $d = 14$, x axis shows the count of time windows. Since the entirety of the dataset includes posts of one year (365 days), there are 122 time windows for each participant.

We constructed one model for each participant. Here we are not interested in making prediction with the GP regression model, instead, our focus is on the covariance function parameter, lengthscale. In the GP regression, the kernel (covariance function) choice determines most of the model's generalization properties. The lengthscale, which is a parameter of the covariance function, determines the "wiggles" of the function. We adopted the Matérn covariance function, which works better than the standard Gaussian kernels (e.g., Squared Exponential Kernel, Rational Quadratic Kernel) in capturing the physical processes due to its finite differentiability. It is less "smooth" than the standard Gaussian kernel. A small lengthscale means the function value changes quickly, while a large lengthscale means that its value changes slowly (Chalupka et al., 2013). By fitting a GP regression model on each user, we obtain a lengthscale of each user's latent mood, and we compare the lengthscale among participants with different symptom levels (low, moderate, high).

Here we used a time window of 14 days and a step size of 3 days as input for the GP regression model. We used a nonparametric test (Mann-Whitney U test) to compare the lengthscale differences between groups. The lengthscale of the high symptom group ($Median = 2.77$) is identical to the moderate symptom ($Median = 2.77$) group ($U = 35424$, $p = 0.01$). However, the low symptom group ($Median = 2.98$) has a

Figure 4.3: Example of GP Regression



Note: here shows examples from two participants, each data point represents mood of every 14 days estimated by the GP regression model. N = 690.

significantly larger lengthscale than the high symptom group (U = 17231, p = 0.01). The moderate symptom group was also significantly different from the low symptom group (U = 7244, p = 0.02). Our result suggest that people with high or moderate depressive symptom level have more mood fluctuations than people with low symptom level.

4.5.2 Classifying Symptom Levels using Daily Mood Representation

Another approach to examine whether the mood profile is associated with depressive symptom is to see if a particular mood state is influenced by depressive symptoms level. We assume the mood states are serially dependent and we used Hidden Markov Model (HMM) (Beal et al., 2002) to model two unobservable states based on a daily mood state representation. This representation comprises four mood states (positive, negative, neutral and silence). Since the behavior of not posting (silence) is included in the modeling process, we did not remove any less active users in this analysis (N = 781).

4.5.2.0.1 Hidden Markov Model We used a multinomial (discrete) emission Hidden Markov Model (HMM) to model users' observed mood for one year (Johansson and Olofsson, 2007). The major parameters used for the model are:

1. Observed mood O_t (time series), daily mood transition representation (d = 1, s = 1).

2. Transition matrix (A), gives the probability of a transition from one state to another.

3. Transition state j.

4. Observation emission matrix (B), which gives the probability of observing O_t when in state j.

An HMM model (denoted by λ) can be written as:

$$\lambda = (\pi, A, B) \tag{4.1}$$

The idea behind this approach is to use the observed mood to estimate the parameter set (π, A, B), A shows us the probability of transferring from one hidden state to another, and B tells us the probability of emitting a certain mood when a user is in a specific symptom state.

We used hmmlearn python library (Gao et al., 2017) to fit emission, transition matrices (using expectation-maximization) and hidden state sequence (using the Viterbi path algorithm), see Section 9 for the initialized probabilities. We trained the model on the entire set of data and observed if the emission probabilities align with our existing knowledge of affect and depressive symptoms. Here we were not to find the optimal model to forecast a new observation sequence, hence we did not test the training model on a test set. Instead, we were interested to know whether the hidden states decoded from the HMM model were associated with depressive symptom levels.

The HMM model decodes a binary hidden state for each day. We speculate that one of the hidden states represents the user experiencing more depressive symptoms (high symptom state), and another represents fewer symptoms (low symptom state). Although the CES-D scale measures an overall symptom level in one week, it is entirely possible for an individual to have more symptom on some days (e.g., sleep disturbance, loss of appetite) but less on others. To test our speculation on the hidden states, we classify participants' self-reported symptom level according to the count of high symptom state. Here we use cutoff score 22 to divide participants into two groups for comparing the results with the existing models. However, there is a challenging questions, up till when shall we count the high symptom states? Since the CES-D scale measures an overall symptom level in the past one week (e.g less than 1, 1-2, 3-4, 5 or more), and the Diagnostic and Statistical Manual of Mental Disorders, 5th Edition (DSM-5) defines depressive symptom as "The individual must be experiencing five or

more symptoms during the same 2-week period". Therefore, we defined our classification criteria as whether participants have at least x days experiencing high symptom in the last y days before they completed the CES-D scale, $x \in X := \{1,2,3,4,5,6,7\}$, $y \in Y := \{7,14\}$.

4.5.2.1 Evaluation of Hidden States

4.5.2.1.1 Emission Probabilities We observed whether the hidden states' emission probabilities align with our existing knowledge in depressive symptom and affect. Table 4.3 shows two hidden states and their emission probabilities to each observation. Given an observed day, we can see both hidden states were most likely to emit silence day because social media users posted sparsely. However, the high symptom hidden state has lower probability to emit silence days compared with low symptom hidden state. The high symptoms state also has a higher probability to emit negative mood or neutral mood, but the low symptoms state has a higher probability to emit positive mood. Therefore, results from the HMM model aligns with our existing knowledge in depressive symptom and affect.

Table 4.3: Emission Probabilities

$N = 781$	Positive	Negative	Neutral	Silence
Low Symptom	8.51	5.20	4.65	81.6
High Symptom	3.15	12.8	7.00	76.9

Note: less symptoms: hidden state that represents less symptoms on a particular day, more symptoms: hidden state that represents more symptoms on a particular day, N: training sample size

4.5.2.1.2 Transition Probabilities of Observations We are also interested to know whether people are more likely to transfer from certain mood states to another. We constructed a transition probability matrix for the observations (daily mood representation). Table 4.4 again shows us that social media users in general are more likely to become silent after they posted any social media content, although high symptom group is less so. High symptom individuals have higher probabilities of changing in between any mood states other than silence. This result aligns with the findings from the GP regression that low symptom individuals shows less fluctuations in their mood representation.

In general, people were more likely to have a positive mood if they had a positive mood in the previous time window. The probabilities of $+ \rightarrow +$, $- \rightarrow -$ were similar among the two groups, but high symptom participants are slightly more likely to transfer from negative to negative. When low symptom participants have a neutral mood, they have similar chances of having a neutral or negative mood in the next time window, whereas, high symptom participants are also more likely to have a negative mood in the next time window. Our result shows that while people, in general, are more vocal when they have a negative mood, but high symptom participants are more likely to vocal about the negative content for a more extended period.

Table 4.4: Transition Probabilities of Observations

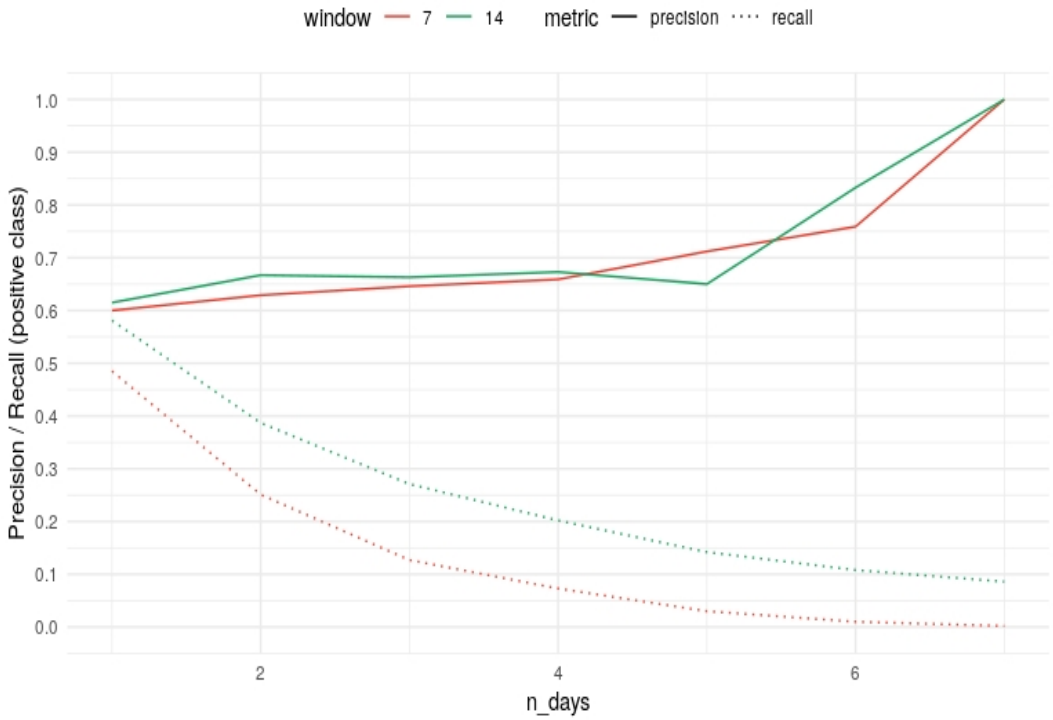
	High symptom				Low symptom			
	+	-	0	S	+	-	0	S
+	21.1	15.7	13.4	49.6	19.5	13.3	12.3	54.8
-	22.3	16.2	14.1	47.3	20.5	13.3	12.9	53.3
0	19.3	14.5	12.8	53.3	17.6	11.6	11.8	58.9
S	5.82	37.5	4.21	86.2	5.92	37.1	4.33	85.9

Note: $+$: positive, $-$: negative, 0 neutral, *S*: silent

4.5.2.1.3 Using Hidden States to Classify Symptom Level Figure 4.4 shows the precision and recall of the high symptom class by counting the hidden states from the HMM model. This figure is plotted based on Table 1 in Appendix .1.2. The baseline model is formulated using a stratified dummy classifier that predicts based on the most frequent training labels. Precision increases as the criterion of x increase. Table 4.5 shows some of the best classification results, see Table 1 in Appendix .1.2 for full results. Assigning participants with six high symptom states within 14 days to the high symptom class results in very low recall (10.8%) but high precision (71.2%). Assigning participants with one high symptom state within 14 days results in a more balanced recall (60.3%) and precision (58.1%) to high symptom class. Result from this classifier does not surpass the baseline in f1 score but when using a higher x as criteria, the precision rate is much higher than the baseline. Our result supports the claim that daily mood representations inferred from social media text is highly associated with depressive symptoms. When a social media user shows specific mood patterns, it is highly likely that the person developed high level of depressive symptoms. However,

only using this approach to identify high symptom individuals would result in a lot of false negative cases.

Figure 4.4: Precision and Recall of High Symptom Class (HMM) with Various Assignment Criterion



Note: window: size of the time window, x days before participants completed the CESD scale. ndays: count of high symptom state within the time window. This figure is plotted based on the result in Appendix .1.2 table 1

4.6 Representation Predictability of Depressive symptoms

The previous analysis suggests that the mood profile is highly associated with depressive symptoms. Now we examine which representation in the mood profile is most predictive of depressive symptoms. We combine the representations with sets of proxy signals in a classification task.

Table 4.5: Predicting depressive symptom with hidden states

Criteria	P	R	f1
baseline	61.2	100.0	76.0
$x = 1, y = 7$	61.5	48.5	45.2
$x = 1, y = 14$	60.3	58.1	59.2
$x = 6, y = 14$	71.2	10.8	18.9

Note: high: high symptom class, low: low symptom class, R: recall of high symptom class, P: precision of high symptom class, f1: average macro-f1 score of both classes, criteria: criteria for classifying high symptom class.

4.6.1 Feature Extractionn

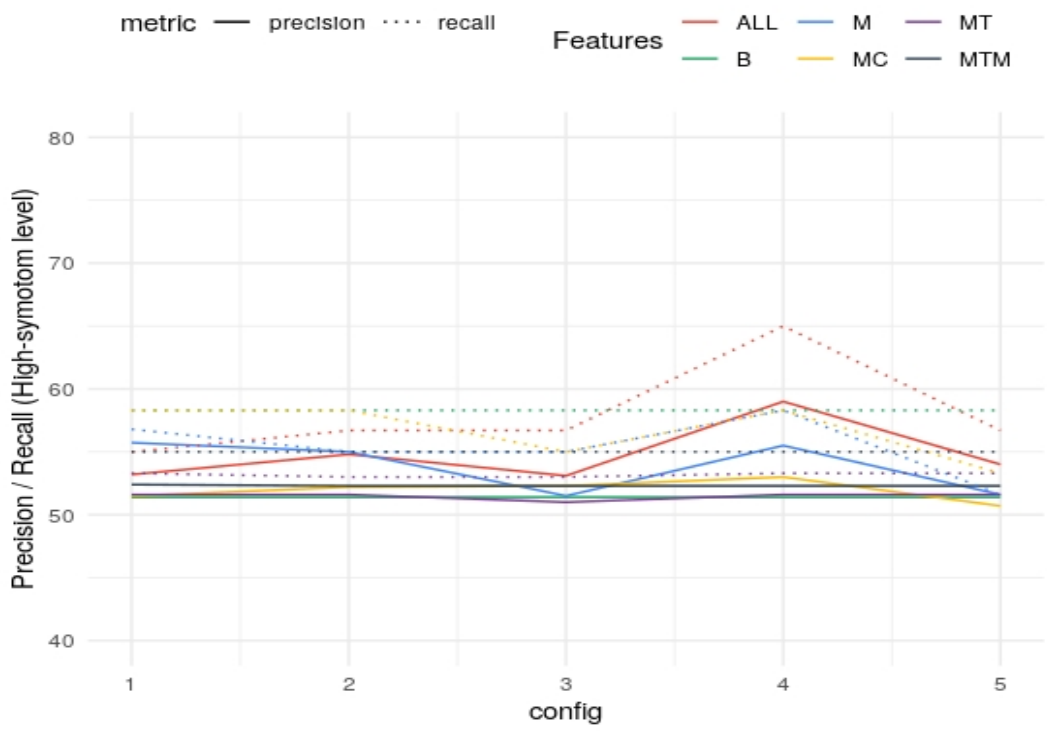
We extracted multiple features for the posts of each user to train multiple models for high-symptoms prediction. Our extracted featured included: 1) n-gram word representation, where $n \in N := \{1,2,3\}$); 2) topic modeling from Latent Dirichlet allocation (LDA) and 3) all the entries from Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2001). N-gram were ordered by term frequency across the corpus, we grid searched the number of most frequent n-gram and number of topics for LDA (see Section 9). We found the most frequent 1500 n-grams and 30 LDA topics gave us the optimal results. These feature variables were commonly used in detecting signs of potential depression (De Choudhury et al., 2013; Park et al., 2012; Coppersmith et al., 2014; Reece et al., 2017). We compare the precision and recall between models with different representations from the mood profile.

Our dataset has an exceptionally high proportion of high symptom individuals as discussed earlier. Given that we have only 303 low symptom participants among 781 participants, we randomly selected 303 participants in the high symptom sample to have a dataset with a balanced class proportion that is closer to the existing literature (1:1), $N = 606$, So that we can have results that are more comparable with the existing literature. We split the data into train (80%, N = 486), and test set (20%, N = 120) in stratified fashion. Stratified five-fold cross-validation was used to optimize the parameters in the model training. A grid search of parameters was carried out for several candidate classification algorithms (e.g., decision trees, support vector machine, logistics regression) (Suykens and Vandewalle, 1999), see Section 9 for the grid search parameters.

4.6.2 Model Evaluation

A baseline model is formulated using a stratified dummy classifier that generates predictions according to the training set's class distribution. Out of several candidate algorithms, logistic regression demonstrated best performance. Hundreds of classification models were trained and evaluated for this task. The models with different representations from the mood profile can be evaluated by precision and recall. We grid searched d and s that maximizes the metrics. Figure 4.5 shows the precision and recall of the high symptom class from models with various configurations and feature sets. Models with configuration 4 (time window 30 days and increment slide 3 days) yield the best scores. Table 4.6 shows the precision, recall and f1 score of the high symptom class from configuration 4. The model with mood, mood momentum and mood transition representations yields the highest scores, and the model with averaged mood over a time window gives second highest scores, 0.59 precision, 0.65 recall, and an F-score of 0.62.

Figure 4.5: Precision and Recall of logistic regression



Note: config 1: d = 7, s = 3, config 2: d = 14, s = 3, config 3: d = 14, s = 7, config 4: d = 30, s = 3, config 5: d = 30, s = 7, B: basic features (n-gram, topic modeling, LIWC), M: B + $Mood_{\mu}$, MC: B + mood momentum, MT: B + mood transition, MTM: B + mood transition momentum, All: all features excluded MTM

Table 4.6: Prediction result of depressive symptom (d=14, s=3

Features	P	R	F1
RB	47.6	50.0	48.8
B	51.4	58.3	54.7
B + M_μ	55.5	58.3	56.9
B + Δm	53.0	58.3	55.6
B + Tr	51.6	53.3	52.4
B + ΔTr	52.3	55	53.6
B + M_μ+ Δm + B + Tr	**59.0**	**65**	**61.9**

Note: R, P, F1 are recall, precision, and f1 score of high symptom classes respectively. B: basic features (tfidf bag-of-words, topic modeling, sentiment, LIWC). RB: random baseline, model parameters: penalty: l2, Inverse of regularization strength: 0.1

4.7 Discussion

4.7.1 The Role of Mood in Predicting Depressive symptoms

Mood is a time dependent variable, using time series approaches to model mood inferred from social media text provides us with better insight about mood and depressive symptoms. Participants in this study demonstrated significantly fewer mood fluctuations if they reported a low symptom score. This finding aligns with the well-established connection between emotionality and depression in the psychology literature. We also found the hidden states from the HMM model are highly relevant to self-reported depressive symptoms, see Table 4.5. Our model suggests that an individual having one high symptom state in 14 days is highly likely to have high symptom level. It is worth to note that the criteria we used in here is different from the criteria in the CES-D scale, where individuals need to have experienced symptoms 1-2 days in the past 7 days to score on a criterion. However, we cannot assume that people will talk about their symptoms every time they experience them. This result suggests that individuals who show specific mood pattern in social media text are highly likely having high depressive symptoms, however, most of the individuals with high symptom do not display this mood pattern.

Existing studies that use a sliding window technique to create dynamic sentiment features have not yet explored which representations and configurations tend to yield a better result in classifying symptoms. We explored various configurations of the slid-

ing window and found that mood in a 30 days time window and move the time window every 3 days is most predictive to depressive symptom level. This result suggests that a less granular mood representation is more beneficial in identifying symptoms. Moreover, combining several representations in the mood profile together can dramatically enhance the model performance. Our best model (f-score: 0.62) encompasses the mood profile and a set of basic features commonly used in existing works. Other studies using multiple sets of proxy signals to predict depressive symptoms achieved a precision score ranging from 0.48 (Coppersmith et al., 2014) to 0.87 (Reece et al., 2017; Guntuku et al., 2017). (Schwartz et al., 2014), using the same data set, achieved correlation of 0.386 with continuous scores. The mood profile can potentially enhance the current screening technology by combining it with more advanced engineered features.

The transition probabilities of mood showed that participants, in general, were more vocal on social media when they were in a negative mood. We speculate that some depressed individuals react to negative mood by posting, and some by silence. Those who are more vocal on social media when in a negative mood might be using social media to reach out to others or use posts as a way to reflect. The associations between negative mood and being vocal, and the association between high symptom scores and a specific mood pattern, suggest that posters could be stratified into several groups, those that withdraw, those that reach out, and those that do not disclose potential signs of depression on social media.

Our results show that a temporal mood profile derived from social media text is highly associated with users' subsequent self reported depressive symptom level. In order to examine the potential of mood momentum and mood transition further, advanced time series analysis techniques need to be applied. Most importantly, mood profiles can potentially provide more information to clinicians than a classification system with binary output.

4.7.2 Technological and Ethical Implications

Similar to the existing studies, the present finding of the derived mood pattern has implications on symptom level but does not provide an accurate interpretation for participants' mental health condition. An accurate interpretation of one's mental health condition requires a holistic view, and any diagnosis requires a strong understanding of an individual's case history. The daily life information contained in social media

data is just a tip of the iceberg of one's life experience.

Our approaches provide a useful source of information for assessing participants' derived mood pattern over time. However, as with all social media related research, ethical and privacy issues need to be considered, given the potential for misusing social media data (Fiske and Hauser, 2014; Lumb, 2016; Cadwalladr and Graham-Harrison, 2018). Using social media analysis techniques in practice requires that the user whose data is being analyzed is comfortable with their social media timeline being used in this way, and that they consent to it. The scope of their consent also needs to be clear, i.e., whether it is for research or whether it is also for potential clinical use.

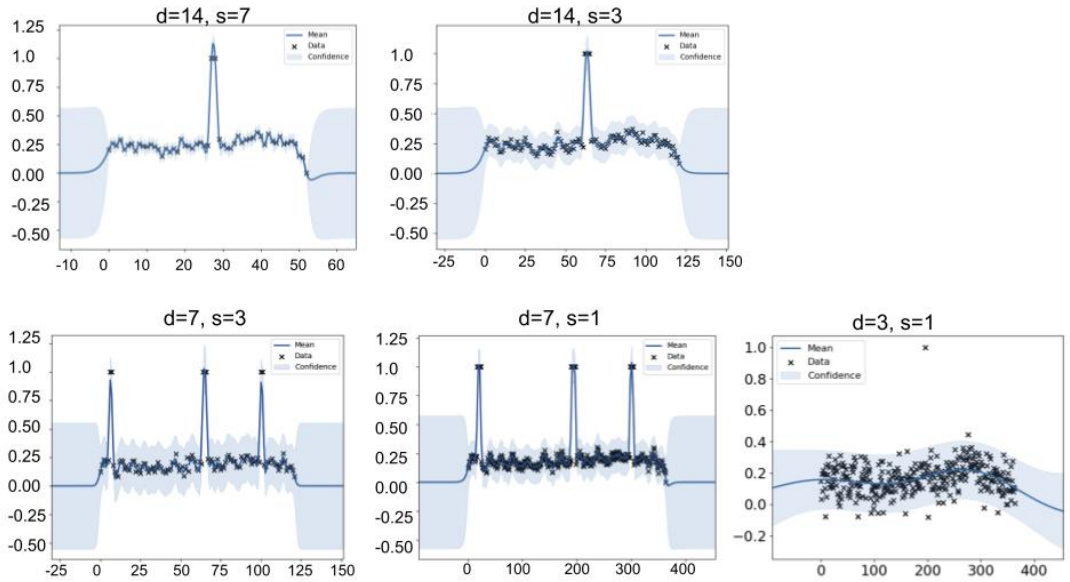
4.7.3 Limitations

Our sample contains participants who allowed researchers access to their Facebook posts and to complete a symptom screening scale. Therefore, this sample may be strongly biased towards those who were comfortable to disclose and reach out on social media. It is still unclear about what the biases are in a sample with these tendencies compared with a random patient sample. Of particular interest is the relatively high depressive symptom score from most of the participants in this sample, and this bias is prevalent in studies in this line of research (Guntuku et al., 2017). We speculate that people who have depression are more curious about taking part in mental health related studies.

In this work, the symptom screening test was conducted once only. There were also no tests controlling for the presence of other disorders, such as bipolar, which greatly affect behavior and mood variability. Those at the high end of the scale could have other types of affective disorders but showing depressive symptoms at the time when they carried out the self-reported measurement. Therefore, the measurement of self-reported symptoms is not an accurate reflection of whether the person has depression.

In addition, the sentiment scores employed in this study were retrieved with SentiStrength, which is a word counting approach to identify positive and negative affect. Although numerous studies have validated the word counting approach, the ideal method to retrieve less noisy sentiment is to construct the sentiment classification model with the examined dataset. Future studies can train their model for sentiment annotation to retrieve more accurate sentiment.

Figure 4.6: Impact of Mood Representation Structure on GP regression (Example 1



Note: d: timewindow size, day as unit; s: step size, day as unit.

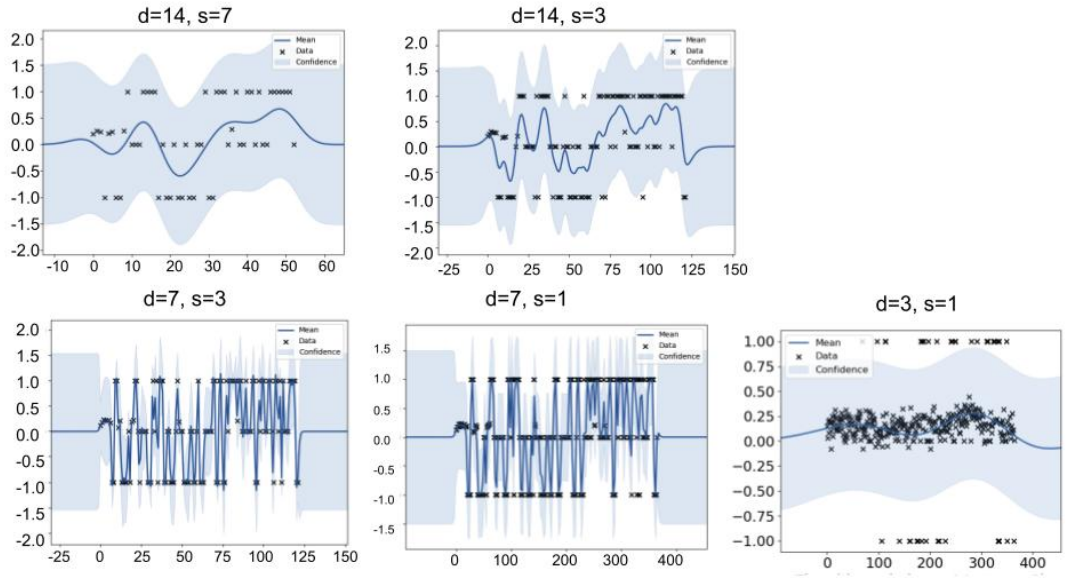
4.8 Further Analysis

This section includes a more detailed explanation of the models and further analysis of the HMM and GP regression models.

4.8.1 Impact of Mood Representation Structure on GP Regression

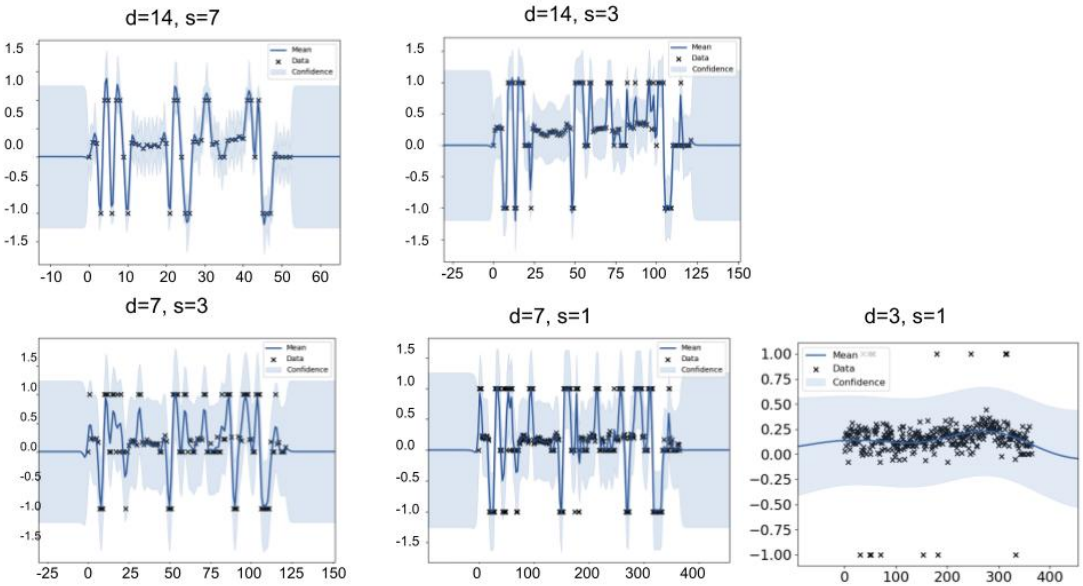
In the published work, we used a time window of 14 days and a step size of 3 days as input for the GP regression model. In this follow up session, we test the impact of step size and time window on the GP regression model by experimenting the input of the model with mood representations of various step sizes and time windows (time window $d \in D := \{7, 14\}$ and step size $s \in S := \{1, 3, 7\}$). A smaller step size ($<$ time window size) enabled the time window to capture information in the previous and current time window. Therefore, we test the following combination of step size and time window (d=7, s=1), (d=7, s=3), (d=7, s=14), (d=14, s=3), (d=14, s=7). Figure 4.6, 4.7 and 4.8 show the GP regression models from three participants as examples to demonstrate our results, each figure contains the d and s combination mentioned above. We can see the regression fits better in representation with a relatively bigger time window d = 14, and a smaller step s = 3. Therefore, the parameters we selected for our published work (d=14, s=3) is the optimal among the tested pairs.

Figure 4.7: Impact of Mood Representation Structure on GP regression (Example 2



Note: d: timewindow size, day as unit; s: step size, day as unit.

Figure 4.8: Impact of Mood Representation Structure on GP regression (Example 3



Note: d: timewindow size, day as unit; s: step size, day as unit.

4.8.2 Model Biases

In this section, we investigate why using the HMM hidden states to classify depressive symptom levels can achieve a high precision but at a great cost from the recall (see Table 4.5). In this task, high precision is important, especially if we use social media data as extra information for a person's mental health status. It helps to raise the red flag for some symptoms. However, this approach also left most of the high symptom users out. We further examined what leads to low recall. We observed that both users with very high symptom level (CESD score > 45, $N = 14$) [1] and those with low symptom level (CESD score < 16, $N = 70$), have a mean number of posts of 143 and 166 respectively, which is less than the mean number of posts in the sample ($M = 194$).

In the HMM model, we included silence days in the daily mood representation. Table 4.3 shows that silence days accounted for most transition states. Therefore, the pattern of silence days strongly influences the model. We examined the results from the model that used the HMM hidden states to assign symptom levels. We found that the algorithm tends to misclassify high symptom users. For example, by using the criteria of $x = 6$, $y = 14$ [2], the mean CES-D score for the misclassified cases ($M = 31.2$) is much higher than the mean of the sample ($M = 26.3$). Nevertheless, among 14 users who had very severe symptom levels (CES-D score > 45), our algorithm only correctly assigned high symptom levels to 1 of them. We speculate that the model is strongly influenced by the fact that users with very high or very low symptoms have similar posting pattern – they both post sparsely.

4.9 Conclusion and Future direction

Mood is an important signal to indicate the development of a depression episode. This chapter unpacks the opportunities and challenges of designing technologies that use social media data to track mood patterns. We provided an outline of utilizing the sliding window technique to construct temporal representations of mood based on sentiment expressed in a social media text. We explored approaches to estimate users' mood when they did not post on a social media platform and we attempted to use a token to represent the behavior of not posting. Our result suggests that our approaches of representing mood inferred from social media data are highly relevant to social media

[1] CES-D score > 45 which means that these individuals self-reported they have had at least 15 symptoms most of the time in the past two weeks.

[2] Users who have six high symptom state in the past 14 days are assigned "high symptom level"

users' depressive symptom levels.

Our follow-up analysis revealed that the social media users' irregular posting pattern poses a great challenge for researchers to infer mental illness symptoms. Posting frequency or posting pattern is often considered a manifestation of depressive symptoms in the existing literature (De Choudhury et al., 2013; Tsugawa et al., 2015). However, users with very high symptom levels and those with very low symptom levels both post less. Among the 14 users who have very severe symptom levels (CES-D score > 45), our algorithm using the hidden states from the HMM model to assign symptom level only correctly assigned high symptom level to one user. These hidden states were estimated using a daily mood representation as input, and silence day was included as one of the observed states in the representation. Therefore, future work should reduce the weight of posting frequency when using social media data to infer depressive symptoms.

In Chapter 3-4, we focused on the affective symptoms manifested in social media text. Existing studies using social media text to infer users mental health status, including Chapter 3-4, only focused on content created by the account users. Reposted and intext copy-and-paste content, such as quotes and lyrics, are not examined separately. We raise the question: are these content noise for inferring mental health status? Shall we include or remove them in the analysis? In the next chapter, we will examine the content that is not created by the account users and its association with depressive symptoms.

Chapter 5

Affect in Non-self-created Content and Depressive Symptoms

Chapter 4 focus on examining the self-created content on Facebook. When studying how mental illness may be reflected in people's social media use, content not written by the users is often ignored because it might not reflect their own emotions. Chapter 3 shows evidence that participants with stronger depression symptoms posted more non-self-created content in pilot study involves a small subsample ($N = 70$). In this chapter, we examine non-self-created content in a larger sample. We extract quotes and song lyrics from the feeds of 781 Facebook users from the MyPersonality database who had also completed the CES-D depression scale. We find that participants with elevated depressive symptoms tend to post more lyrics, especially lyrics with neutral or mixed sentiment. By analyzing the topics of those lyrics, we find they center around overwhelming emotions, self-empowerment and retrospection of a romantic relationship. Our findings suggest removing quotes, especially lyrics, might be eliminating content that reflects users' mental health conditions.

5.1 Introduction

Social media records provide psychologists with a novel way of examining mental illness symptoms (De Choudhury et al., 2013; Glen et al., 2015; Park et al., 2012; Chen et al., 2020d). Existing studies often focus on the emotional content written by the social media users themselves, which we refer to as "self-created Content" (SC)(Seabrook et al., 2018; Schneider and Carpenter, 2019). Non-self created content, such as reposts, music and videos, are seen as reflecting indirect emotions of the user

and thus receives less attention in the analysis (Wang and Zhuang, 2017) .

There are two types of non-self created content: *repost* (e.g., shares and retweets), and *copy-and-paste quotes* (e.g., song lyrics, religious verses, and famous quotes). A repost is easy to identify since it is a functionality on social media platforms to share a post from another user. However, quotes are more complicated to identify, since usually there are no quote marks or reference to the source. Furthermore, sometimes quotes and lyrics are posted from memory, which can introduce distortions. There are limited number of studies that examined the reposts (Tsugawa et al., 2015; Guo et al., 2009), while, to our knowledge, quotes have not been studied yet.

There is extensive work on how affective disorders, especially depression (De Choudhury et al., 2013; Seabrook et al., 2018; Tsugawa et al., 2015), are reflected in social media data (Chancellor and De Choudhury, 2020). However, existing studies in this line of research focus on self-created content only. We are not aware of work that analyzed the usage of lyrics and quotes in general in social media posts of users with different levels of depressive symptoms. This is surprising, because music is associated with mood regulation process (Gladding et al., 2008; Hunter et al., 2011; Sachs et al., 2015).

Our research questions are:

1. Is posting quotes associated with levels of depression symptoms?

2. What are the themes and emotions conveyed in the quotes posted by people with high symptoms of depression, and how might they relate to symptoms?

This study is the first to our knowledge that examines whether quotes in social media is associated with users' emotional state. We analyzed a set of 93,378 posts from 781 Facebook users who consented to take part in the myPersonality study (Bachrach et al., 2012; Youyou et al., 2015) who completed the CES-D test in addition to the personality scales.

We constructed a classifier to automatically detect potential quotes in Facebook posts, and distinguish song lyrics from other types of quotes. Then we used generalized linear modeling to examine potential links between the emotions in the quotes, especially lyrics, and the levels of depressive symptoms reported in the CES-D scale. Finally, we use topic modeling to examine the themes of lyrics that are posted by people with high versus low depressive symptoms.

We found that quotes account for more than 10% of the content in 12.6% of par-

ticipants [1]. Users with higher depressive symptom levels tend to post more lyrics with neutral sentiment. Our findings suggest that lyrics are used as an agent for users to communicate their emotions indirectly. Therefore, not all non self-created content should be regarded as noise.

5.2 Background

5.2.1 Social Media Behavior and Depressive Symptom Level

Experiencing negative emotions can increase the amount of social interaction and sharing of emotions (Luminet IV et al., 2000), both of which are part of the mood regulation process that leads to mood improvement (Hill, 1991). Posting patterns on social media and the mood of posts may reflect symptoms of depression (Chen et al., 2020a,d; Seabrook et al., 2018; Tsugawa et al., 2015). To study the emotions expressed in text, researchers often use sentiment analysis that categorizes affect or opinions expressed in the text (Pang et al., 2008). Several studies have shown that users with the presence of depressive symptoms use more negative affective words (e.g., sad, cry, hate) in their text than those who do not (De Choudhury et al., 2013; Tsugawa et al., 2015).

5.2.2 Effects of lyrics and quotes on depression

A recent study by (Chen et al., 2020c) on a subset of the current data set found a potential association between posting quote on Facebook and users' depressive symptom levels, but they did not examine the sentiment expressed in the quotes. Surprisingly, the link between the content of quotes and symptoms of depression has not been examined in the existing literature, even though music is strongly linked to emotions.

People often choose music that is in congruence with their mood. Listening to songs centered around hurt, pain, and grief is part of the mood regulation process for coping with aversive life events (Gladding et al., 2008). Retrieving nostalgic memories from music may enhance the mood, especially when these memories are related to meaningful moments in life (Routledge et al., 2012; Taruffi and Koelsch, 2014). Listeners may find some consolations in lyrics because when they realize they are not alone in dealing with the painful situations (Gladding et al., 2008; Sachs et al., 2015). Quotes, especially song lyrics, may induce congruent emotions (Gladding et al., 2008;

[1] for distribution details, see Figure 1 (Appendix .2.2))

Hunter et al., 2011; Sachs et al., 2015).

5.3 Data Collection and Preparation

5.3.1 MyPersonality Dataset

We used the myPersonality data set (Bachrach et al., 2012; Youyou et al., 2015). myPersonality collected Facebook posts from 180,000 participants from 2010 to 2012, with the consent from Facebook users. The data collection process complied with Facebook's terms of service, and we obtained the required permission to use the data. 781 participants over the age of 18 also completed the Center for Epidemiologic Studies Depression Scale (CES-D) in 2011 or 2010. We extracted all 93,378 posts that these participants posted up to one year before completing the CES-D. On average, participants posted 120 posts during the selected time window. Most participants are young ($M = 26$, $SD = 11.7$) female ($N = 448$, 57%) and white-American ($N = 309$, 39%). Detailed demographics are provided in Table 2.

5.3.1.0.1 Depressive Symptom Screening Test The 20-item CES-D scale is one of the most widely used tools that measure the presence of depressive symptoms in the general population (Radloff, 1977). It has high internal consistency, test-retest reliability (Radloff, 1977; Orme et al., 1986), and validity (Orme et al., 1986). Scores range between 0 and 60. Following common practice, we adopted 22 as a cutoff point to divide participants in our dataset into high symptom (P_{HS}, $N = 478$) and low symptom groups (P_{LS}, $N = 303$) (De Choudhury et al., 2013; Park et al., 2012; Reece et al., 2017; Tsugawa et al., 2015).

5.3.2 Identifying Quotes in User Timelines

For each post, we retrieved the first page of search results via the Google search API, which included the link, title, and snippet. Since we observed that quotes often contain misspellings or small variations, we created a rule-based classifier outlined in 1, which assigns each post to one of three classes: 1) Lyric, 2) non-lyric quote (NL-quote), and 3) self-created content (SC).

In order to calculate the cosine similarity between post and snippet, we created document vectors for each by converting each word to word vectors using the pre-

Algorithm 1 Quotes identification algorithm. $\cos(\theta)$ is the max cosine similarity between post and each of the retrieved snippets. C_q and C_l are the counts of search results that contain the word "quote" or "lyric" respectively

$\cos(\theta) = argmax(cosine(post, snippet))$
$C = C_q + C_l$
if $(\cos(\theta) > X)$ OR $(X > \cos\theta > Y$ and $C > 0)$ OR $(Y > \cos\theta > Z$ and $C > N)$ **then**
    label ← Quote
    if $(C_l > N_l)$ **then**
        label ← Lyric
    end if
else
    label ← Self-create
end if

Table 5.1: Result of Quote and Lyrics Classifier shown in Algorithm 1

	validation			test		
	F1	Recall	Precision	F1	Recall	Precision
NL-quote	87.8	88.3	87.2	89.1	85.3	94.0
Lyrics	76.6	78.0	80.3	79.6	80.0	82.2

trained word vectors from Python Package Spacy (Honnibal and Johnson, 2015), and summing the word vectors into a single document vector.

The first author annotated a subset of posts for the quote classifier to determine the values of the thresholds in Algorithm 1 $\{X, Y, Z, N, N_l\}$ using grid search, and to test the performance on a separate test set. 750 posts were annotated, where 523 were used as validation for threshold optimization and 227 were used for testing. The final values for the threshold are: X = 0.998, Y = 0.975, Z = 0.85, N = 3, and N_l = 2. Table 5.1 shows the classifier performance on the validation and test sets for detecting quotes (NL-quotes and lyrics).

Among 93,378 Facebook status updates in our collection, we identified 3,722 posts classified as quotes posted by 305 (39%) of our 781 participants, out of which 1,488 (40%) are song lyrics posted by 102 (13%) of the participants. Figure 1 (Appendix .2.2) shows the percentage of posts which were quotes for all 305 participants. For roughly a third ($N = 99$), 10–40% of their posts were quotes.

Table 5.2: Sentiment distribution in identified quotes and lyrics for P_{HS} and P_{LS}

	High symptoms (P_{HS})				Low symptoms (P_{LS})			
	total	pos	neg	neut/mix	total	pos	neg	neut/mix
Lyrics	1056	40%	24%	36%	432	40%	27%	33%
NL-Quotes	1597	41%	25%	34%	637	41%	22%	37%
Total Quotes	2653	40%	25%	35%	1069	40%	24%	36%

5.4 Quotes and Depressive Symptom Levels

5.4.1 Frequency and Sentiment of Quotes

When considering all 781 participants, 198 (41%) of the P_{HS} (high symptoms) share quotes compared to 107 (35%) of P_{LS} (low symptoms). (P_{HS}) also share significantly more quotes and lyrics on their timeline ($M = 5.55, SD = 14.07$) than those with low symptoms (P_{LS}, $M = 3.52$, $SD = 9.12$) (t-test, t(2.4) = 778.61, p = 0.015).

The sentiment of quotes was analyzed based on the sentiment scores calculated by SentiStrength (Thelwall et al., 2011). Posts that are dominated by a polarized sentiment score were labeled correspondingly as Positive or Negative, while posts with no dominant polarized sentiment (neutral or equivalent magnitude of positive and negative words) are labeled as Neutral/Mixed. Table 5.2 shows the full distribution of sentiment in the quote posts shared by participants for each P_{HS} and P_{LS}.

5.4.2 Sentiment of Quotes

Since P_{HS} are more likely to post quotes than P_{LS}, we now examine the relationship between the sentiment of those quotes and the level of depressive symptoms more closely. We use logistic regression, with symptom group (low versus high) as the dependent variable, and frequency of lyrics and quotes and sentiment of lyrics and quotes as independent variables. Frequencies are expressed as ratios, to account for differences in the number of posts per participant.

We consider two models, one with all independent variables (Model 1), and one excluding variables that have high collinearity with others (Model 2). The correlations are summarized in Figure 1. Most variables are weakly correlated ($r < 0.25$) with each other. However, the magnitude of sentiment variables are moderately correlated with the ratio of lyrics and quotes ($r > 0.40$); therefore, we drop them from Model 1 to

Table 5.3: Logistics Regression Models. B: Beta coefficient, SE: standard error of the coefficient, $*: p < 0.05$

	Model 1		Model 2	
Variables	B	SE	B	SE
ratio of positive lyrics to quote	0.52	1.02	0.40	0.79
ratio of negative lyrics to quote	-0.59	0.85	-0.14	0.69
ratio of neutral or mixed lyrics to quote	1.82	0.85*	1.89	0.84*
ratio of lyrics to total post count	-8.75	4.15*	-6.92	3.76*
ratio of positive NL-quote to quote	-0.26	0.61	-0.23	0.40
ratio of negative NL-quote to quote	0.56	0.77	1.15	0.58*
ratio of neutral or mixed NL-quote to quote	-0.22	0.41	-0.10	0.41
averaged sentiment magnitude of positive lyrics	-0.06	0.19		
averaged sentiment magnitude of negative lyrics	-0.18	0.18		
averaged sentiment magnitude of positive NL-quote	-0.02	0.19		
averaged sentiment magnitude of negative NL-quote	-0.17	0.18		
ratio of mixed/neutral posts to total post count	0.95	0.70		
ratio of negative posts to total post count	0.90	0.73		

youeld Model 2.

ANOVA show that Model 1 (AIC = 1028) and Model 2 (AIC = 1021) are not significantly different from each other (F(2,6) = -4.86, $p > 0.05$). We see the strongest association between symptom level and content variables for the ratio of lyrics to total post count. There are no clear associations between negative mood and depression level (see Table 5.3).

5.4.3 Themes in Quotes

We have seen that people with higher levels of depressive symptoms are more likely to post quotes, in particular lyrics, but we did not see any clear links between quotes with negative mood and symptom level. Therefore, we decided to use LDA topic modeling (Newman et al., 2006; Blei et al., 2003) to extract common themes in quotes and lyrics posted by P_{HS} versus P_{LS}. We used verbs, nouns, and adjectives as input for the LDA topic model. Each input word was labeled whether it came from a post by P_{HS} or P_{LS}. The best-performing model yielded 15 topics. Details of the topic modeling are given in 9 Table .2.1.

Most of the topics in the lyrics reflect hurt and grief in romantic love. Among the

five most prevalent topics in lyrics, nearly all of them mainly comprise of words from P_{HS}.

Table 3 (Appendix .2.2) shows the 10 most frequent keywords and themes of the seven most prevalent topics. Three of the most common topics of lyrics deal with empowerment, in particular self-empowerment. Topic 0, with keywords such as love, want, feel, need, think, is highly emotionally charged and mainly comprised of words from P_{HS}, while Topic 7 comprises of lyrics that indicate introspection (e.g., feel, know). Topics from non-lyrics quotes are less varied. Most of the non-lyrics posts are dominated by two topics, which center around life, love, and feelings towards various entities. The dominant words from topic 13 are mainly from high symptom individuals, whereas those in topic 9 are from low symptom individuals.

5.5 Conclusion

In this study, we showed that people with high levels of depressive symptoms are more likely to post quotes and lyrics on Facebook.

However, the ratio of negative sentiment posts does not appear to be associated with depressive symptom levels (Tsugawa et al., 2015). Instead, people with more depressive symptoms are more likely to post lyrics, and the sentiment of those lyrics tends to be mixed or neutral.

Most of the lyrics centered around hurt, pain, and grief in a romantic relationship, which may indicate a mood regulation process (Gladding et al., 2008). Some of the lyrics reflect introspection and the desire for self-empowerment, which is part of the coping process. Therefore, we argue that lyrics and quotes should not be excluded from studies of the ways in which people with depression use social media—they may hold important clues to coping strategies.

5.6 Further Analysis

In the published work, we used SentiStrength, a lexicon-based sentiment tool, to extract sentiment from the text. In recent years, word embeddings methods have been widely adopted in sentiment classification tasks (Rezaeinia et al., 2019). In the follow-up analysis, we adopted several pre-trained deep learning models to re-annotate the sentiment in our dataset. We used human annotation to evaluate the sentiment annotations using SentiStrength and deep learning approach. We re-examined our analysis of

the published work using the sentiment annotated by the deep learning model.

SentiStrength has a predefined list of positive and negative words. The algorithm scores the text by counting these words' occurrences in the text. The sentiment score of the post is the sum of two scores in the published work. Posts with the following situations were considered as neutral:

- Posts with 0 positivity or negativity.
- Posts with an equal magnitude of positivity and negativity.

With the neural network approach, we defined posts with 0 positivity or negativity as neutral. We compared the quality of the sentiment score from SentiStrength and deep learning models on a set of annotated data. We randomly selected 150 posts from a set of lyrics or quotes (N = 150) identified by our classifier. Two annotators followed an annotation guideline adopted in Chapter 3, [2]. Each post was assigned positive, negative or neutral sentiment. The inter-rater reliability of the two annotators was 0.75 (Cronbach's alpha). We did not annotate the magnitude of the post because we found our annotators' perception of the sentiment magnitude was too subjective. We found a few challenges during our annotation process:

- Users often repost motivational quotes or religious quotes, for example, "Trust in the Lord with all your heart, and lean not on your understanding; In all your ways acknowledge Him, and He shall direct your paths." These quotes are usually positive or neutral, but users seem to be enduring some challenges and using these posts to encourage themselves to face the challenges.
- Some lyrics do not provide contextual information to the story. Therefore, our annotators tend to assign lyrics without context information to neutral. For example, "All you have to do is close your eyes and just reach out your hands and touch me. Hold me close. (More Than Words)"

We used pre-trained neural network models to predict Facebook post sentiment. We selected a few pre-trained models that use the transformer architecture: "cardiffnlp/twitter-roberta-base-sentiment" [3] Barbieri et al. (2020), "nlptown/bert-base-multilingual-uncased-sentiment" [4] and "distilbert-base-uncased-finetuned-sst-2-english" [5] (huggingface, 2021).

[2] https://github.com/luciasalar/AnnotationTasks

[3] a roBERTa-base model trained on 58M tweets, it is finetuned for sentiment analysis with the TweetEval benchmark

[4] bert-base-multilingual-uncased model finetuned for sentiment analysis on product reviews in six languages: English, Dutch, German, French, Spanish and Italian.

[5] DistilBERT architecture model fine-tuned on a dataset called SST-2 for sentiment analysis tasks.

Table 5.4 shows the evaluation of the machine annotations. We found that the Cardiff NLP model generated more accurate sentiment scores than other pre-trained models and SentiStrength. It is expected that the pre-trained embedding trained on tweets is more suitable for our task because Facebook posts are similar to tweets in terms of length and content type. The word-based approach SentiStrength has the worst performance. It is not surprising that the neutral class is most difficult to be identified since it is even difficult for a human to identify neutral sentiment.

Finally, we defined an ensemble classifier score by averaging the scores from the three models. In the ensemble classifier, neutral was assigned to the post if both Cardiff NLP and NLP town assigned it to neutral. Cardiff NLP tends to predict more negative cases as neutral, resulting in a high recall and low precision rate in the neutral class but a low recall rate in the negative class.

Our published work found that the ratio of neural or mixed lyrics to quotes, ratio of lyrics to total post count, and ratio of negative non-lyrics quotes contribute to depressive symptoms. In this follow-up work, we also used lyrics and quote ratio to predict symptom levels in a logistics regression model. Here we replaced the SentiStrength score with sentiment scores annotated by the ensemble model (see Table 5.5). We still found that the ratio of neutral lyrics and number of lyrics contribute to symptom level.

5.7 Review for Follow-up Analysis and Next Steps

Our published work shows that people with more depressive symptoms are more likely to post lyrics, and the sentiment of those lyrics tends to be mixed or neutral. In our follow-up analysis, we re-annotated the sentiment score using a pre-trained deep learning model. We found that using pre-trained embeddings on our sentiment annotation is better than SentiStrength. Our logistics regression model using the re-annotated data showed a similar result as the model in our published work. The number of lyrics and neutral lyrics was associated with users' depressive symptom levels. From our human annotation process, we learned that the neutral lyrics were usually lyrics that can be interpreted positively or negatively based on the context. It is still unknown why this type of content was associated with users' depressive symptoms. Future studies should investigate this issue.

Table 5.4: Machine sentiment annotation evalution

model	class	precision	recall	f1-score
SentiStrength	negative	0.58	0.44	0.5
	neutral	0.16	0.36	0.22
	positive	0.36	0.28	0.31
	f1 weighted average	0.44	0.38	0.40
Cardiff NLP	negative	0.91	0.42	0.57
	neutral	0.28	0.86	0.41
	positive	0.76	0.68	0.71
	f1 weighted average	0.77	0.57	0.60
NLP town	negative	0.77	0.53	0.63
	neutral	0.15	0.18	0.16
	positive	0.5	0.70	0.58
	f1 weighted average	0.59	0.54	0.55
Distilbert	negative	0.72	0.62	0.66
	neutral	0	0	0
	positive	0.49	0.80	0.61
	f1 weighted average	0.53	0.59	0.55
ensemble	negative	0.74	0.60	0.66
	neutral	0.27	0.14	0.18
	positive	0.51	0.78	0.62
	f1 weighted average	0.60	0.59	0.58

Table 5.5: Logistics regression model with sentiment annotated by ensemble classifier. B: Beta coefficient, SE: standard error of the coefficient, $*: p < 0.05$

	B	SE
ratio of positive lyrics to post count	-11.88	6.73
ratio of negative lyrics to post count	-8.28	7.16
ratio of neutral lyrics to post count	-28.92	15.37*
number of lyrics	0.16	0.06*
number of NL-quote	0.006	0.03
ratio of positive NL-quote to post count	3.31	4.00
ratio of negative NL-quote to post count	3.83	3.62
ratio of neutral NL-quote to post count	-7.54	13.99

Chapter 6

Identify Distorted thinking from Social Media Text

This chapter aims to identify cognitive distortion expressed in social media text. Cognitive distortion is irrational thought occurring at the onset or perpetuation of mental disorders (Beck, 2019, 1991). We annotated cognitive distortion (see Chapter 1, Section 2.2.1) among more than 4,145 Facebook posts to establish a gold standard. Although cognitive distortions only accounted for 1% of the posts in our sample, they are positively associated with users' self-reported depressive symptom scores, and life satisfaction score but not personality dimensions. Compared with negative affect expressed in the post, cognitive distortions showed a stronger association with users' depressive symptom levels. Among users who posted cognitive distortion more often, we found that they did not use more words associated with anger and negative affect, but tended to describe their feelings in their accounts. They also used more words related to risk but fewer terms related to reward. Our work reaffirms that cognitive distortions can be identified in Facebook posts. Our finding suggested that cognitive distortion extracted from social media text may be a more robust feature than affective words to identify people who were at risk of developing depression.

6.1 Background and Prior Work

When we experience failure, rejection or trauma, we sometimes develop automatic thoughts that lead us to think inaccurately about reality. For example, "I'm all alone in this world, no one will help me." These negative, irrational systematic thoughts are "cognitive distortions" (Beck, 2019). Cognitive distortions cause an individual

to perceive the reality inaccurately. Psychologists suggest that cognitive distortions are driven by biased schemata. These schemata influence information processing by guiding the encoding, organization and retrieval of the stimuli. Adverse events that occur earlier in life might lead to the development of biased schemata (Zhang et al., 2011; Poletti et al., 2014; Beasley et al., 2003).

Cognitive distortion is generally characterized by negative self-referential beliefs. Once activated, the self-referential schemata lead to specific impairments in attention, interpretation and memory (Clark et al., 2000; Disner et al., 2011). The activation of these dysfunctional attitudes also increases the likelihood that a person pays more attention to the mood-congruent stimuli. This process increases one's awareness for depressive stimuli in the environment, thus blocking the positive emotions from a pleasing event (Smith et al., 1994; Mahoney, 2013).

The process of biasing towards negative stimuli is also closely linked to neuroticism. Individuals with a high neuroticism level have consistent, unique patterns in emotional responses (Martin, 1985), specifically when the individual recalls experience related to self. For example, multiple studies have found that the degree of neuroticism was significantly correlated with the time an individual recalls pleasant memories (Ruiz-Caballero and Bermúdez, 1995; Lloyd and Lishman, 1975). The systematic biases in attending negative memories give rise to related cognitive distortions. For example, people with schema bias tend to give more negative interpretations and self-attribution for adverse events.

Since the schema bias, depressed mood and neuroticism are inter-correlated, the schema bias not only leads to psychiatric disorders such as anxiety and depression (Beck, 2019, 1991, 2008), but also plays a significant role in maintaining psychiatric disorders (Beck, 1991, 2008). One of the goals of cognitive-behavioral therapy for anxiety and depression is to help individuals adjust these biases (Harvey, 2004). Psychologists commonly use a checklist to assess cognitive distortions. Beck proposed six categories of cognitive distortion: arbitrary inference; selective abstraction; overgeneralization; magnification and minimization; personalization; dichotomous thinking (Hollon and Beck, 1979) (see Table 6.1). Over the years, researchers expanded the list based on the existing model (Oliveira, 2014; Newman, 2015).

In the last decade, researchers have extracted social media signals to infer depressive symptoms (De Choudhury et al., 2013; Tsugawa et al., 2015; Chancellor and De Choudhury, 2020). One of the most widely studied social media signals is affective words in the social media text (Chen et al., 2020d; Saravia et al., 2016). In

contrast, cognitive distortions as a major element for depressive symptoms is barely studied in the social media context.

A few studies found that cognitive distortions can be identified in the social media context (Simms et al., 2017; Ophir et al., 2017). Ophir et al. (2017) examined the Facebook posts from teenagers who had received psychotherapy treatment and found that cognitive distortion express in their posts was highly associated with depressive symptoms. However, Ophir et al. (2017)'s study only contained 190 Facebook status updates from teenagers who received psychotherapy treatments. It is also unclear whether the finding can be generalized to another sample. In this work, we attempted to identify cognitive distortions in Facebook posts and examined their association with users' depressive symptom levels. In contrast to Ophir et al. (2017)'s work, our sample is not exclusive to people who didn't go through a therapy treatment. In our sample, more than half of the users reported a low level of depressive symptoms.

Ophir et al. (2017) suggested that the annotation process of cognitive distortion in Facebook posts largely overlapped with the process of annotating negative affect. Therefore, we annotated cognitive distortion and negative affect in 4145 posts from 77 Facebook users according to the Cognitive Distortions Questionnaire (CD-Quest). CD-Quest is a brief 15-item questionnaire that assesses the frequency and intensity of cognitive distortions (Morrison et al., 2015; Oliveira, 2014). CD-Quest has been validated in samples of various populations (Kaplan et al., 2017; De Oliveira et al., 2015; Qian et al., 2020).

We examined the correlation between the cognitive distortions and users' self-reported depressive symptoms, life satisfaction and personality. To take a step further, we also examined language characteristics of users who posted cognitive distortions more often on their social media wall. To be specific, we designed the research questions as follows:

1. RQ 1: Do cognitive distortions reflected in social media text have associations with users' depressive symptoms, well-being and personality dimensions?

2. RQ 2: Is cognitive distortion expressed in social media text a better indicator of users' depressive symptoms than affective words?

3. RQ 3: What are the language characteristics among users who post content with more cognitive distortions?

6.2 Methods

This section describes our methods for annotation and statistical analysis. First, we annotated the negativity and the positivity of the users' affect in social media text. Cognitive distortions are most likely to be negative, therefore, we only annotated the cognitive distortions among posts with negative affect in this pilot study. While we are aware that cognitive distortions may carry positive sentiments in rare situations, we leave the analysis of those cases for future work.

Next, we adopted correlation tests to determine if affect and cognitive distortions from social media data were associated with participants' self-reported depressive symptoms, satisfaction with life, and personality. Finally, we examined the language characteristics in posts with cognitive distortion with Linguistics Inquiry and Word Count (LIWC) (Pennebaker et al., 2001).

6.2.1 Data

This corpus consists of 4696 Facebook posts from individuals who participated in the myPersonality project from January 2009 to December 2011. Our methods for data analysis were carried out in accordance with the approved guidelines from myPersonality. myPersonality was a Facebook-based application collecting psychometric tests from users. Participants opt to allow myPersonality to collect their account information and public Facebook posts. Collection of myPersonality complied with the terms of Facebook service. All data are anonymized and gathered with opt-in consent for research purposes. Our study sample contains 301 participants who have completed the Center for Epidemiologic Studies Depression Scale (CES-D), Satisfaction with Life Scale and Big-5 Personality Scale.

6.2.2 Sampling Approach

To ensure we have enough posts to conduct a longitudinal study, we only included regular posters in our sample. We defined regular posters as individuals who posted twice per week or more. We estimated this using the average post count per day during the sampling frame. If an individual had a post count per day of 0.3, this individual made around 109.5 posts in 365 days, roughly equivalent to an average of 2.1 posts per week. In our sample, 122 out of 301 participants were regular posters. Since the CES-D score provides us with a " ground truth", we analyze data from posts that have

been made around the time that the score was obtained. Then we collected a sample of 4696 posts from 91 regular users, these posts were produced two months before the CES-D score was obtained. We further eliminated 14 posters who contributed less than 20 posts during the two months. We also removed posts that were not written in English. As a result, we were left with a sample of 4145 posts from 77 users.

6.2.3 Self-reported Measurement Scale

From the extensive data collected within myPersonality, we chose the Center for Epidemiologic Studies Depression Scale (CES-D) (Radloff, 1977) for measuring Facebook users' depressive symptoms. The scale was test for its internal consistency, test-retest reliability Radloff (1977); Orme et al. (1986); Roberts (1980), and validity Orme et al. (1986). We also adopted the *Satisfaction with Life Scale (SWL)* for quantifying users' mental wellbeing. The 5-item SWL scale has been found to have high internal consistency and temporal reliability across different cultures and age groups Diener et al. (1985); Pavot and Diener (2009). For personality and life satisfaction, we adopted the Five Factor Model of Personality (Big-5) (McCrae and John., 1992). Personality traits were measured using a 100 item scale from the open source International Personality Item Pool Goldberg et al. (2006).

6.2.4 Annotation Process

This section introduces our annotation process for affect and cognitive distortion. Samples cited in this work were modified via paraphrasing the original post. Our first task was to annotate whether the post contained positive, negative, or mixed affect [1]. Our second task was to identify cognitive distortions among posts with negative or mixed affect [2]. Five student annotators and I completed the first task. The second task was completed by me alone. A member of the supervisory team co-annotated some of the samples.

This work is an exploratory study. The annotation for the cognitive distortion task is yet to be validated by a second annotator because the funding for a co-annotator could not be secured. This issue needs to be addressed in the future work.

[1] https://github.com/luciasalar/AnnotationTasks

[2] https://github.com/luciasalar/AnnotationTasks/blob/master/CognitiveDistortion.md

6.2.4.1 Affect

We defined our annotation guideline of affect according to concepts described in (Mohammad, 2016). There were two steps in this annotation task:

1. We identified the Primary Target of Opinion (PTO). PTO refers to the entity towards which we can determine the speaker's attitude. The entity can be a person, situation or event.

2. We identified the positivity and negativity of the post related to the PTO according to the valenced words (e.g., adjectives, emoji, verbs). In contrast to other existing affect annotation task, we introduced a Mixed class to define posts containing both positive and negative affect.

Here are some examples for affect annotation:

- Positive
  - We made a snowman last night! :)
- Negative
  - I'm so bored of quarantine.
  - My cat died.
- Mixed
  - I like the movie, but the conversation between Tom and Jack makes me feel a bit sad.

Sometimes users reposted affective content, for example:

> *you have a sister who has made you laugh, punched you, stuck up for you, drove you crazy, hugged you, watched you succeed , saw you fail, picked you back up, cheered you on, made you strong, and is someone you cant live without someone you can always count on....REPOST THIS IF YOU HAVE A SISTER THAT YOU LOVE.*

We included content such as this in our annotation task because our earlier work showed that copy-and-paste repost content also reflects depressive symptom level (Chen et al., 2020b).

6.2.4.2 Cognitive distortion

We developed an annotation guideline according to the cognitive distortion checklist (Oliveira, 2014). Below we list samples of cognitive distortion from social media context (see Table 6.1). It is worthy of note that some categories of cognitive distortions are more commonly present in the social media context. However, we did not distinguish the categories in our annotation task, therefore, we were not able to provide a conclusion on which categories were more common.

6.3 Results

6.3.1 Cognitive Distortions in Social Media Text Reflect Depressive Symptoms

Of 4145 posts, 804 of them were dominated by negative affect, and 36 of them contained a mix of positive and negative affect. Among 840 posts that contained negative affect, only 41 contained cognitive distortion. We found that cognitive distortion rarely presented in the social media context, and which accounted for only 1% of the posts in this sample.

To understand whether cognitive distortions extracted from social media text were associated with depressive symptoms, we computed a cumulative cognitive distortion score for each user by counting how many of their posts contained cognitive distortions. We used the same approach to compute a cumulative negative affect score. Table 6.2 shows the statistics of the two scores and their correlations with depressive symptoms (CES-D). The p-values were adjusted by bonferroni correction. We further normalized the negative affect score and cognitive distortion score with the total post count. Table 6.3 shows the correlation between depressive symptoms, the negative affect and cognitive distortion scores.

Our results show that although cognitive distortion only accounted for 1% of this sample, it had a moderate correlation with self-reported depressive symptoms. The cumulative score of cognitive distortion was more strongly correlated with depressive symptoms than the normalized score. Cumulative cognitive distortion was also significantly correlated with life satisfaction. Both the cumulative and normalized negative affect scores were not significantly correlated with users' depressive symptom scores.

Table 6.1: Examples of Facebook Posts with Cognitive Distortion

post	cognitive distortion	comments
I feel like my life is a waste. I have no story, no influence, no particular skills that are useful. I suck at everything.	*Minimization*: Minimizing or discounting the importance of some event, trait, or circumstance. *Labeling*: Labeling oneself or others using derogatory names.	The author generates the global negative pattern based on some incidents. The actual number of such incidents is unknown. But the author fails to focus on life events that are counter to this statement. The author also labels themselves as "a waste".
I hate the past. I have to erase them from memory forever. I don't care if the memories were good or bad.	*Discounting positives and dichotomous thinking*: The tendency to view all experiences as fitting into one of two categories (e.g., positive or negative; good or bad).	The author hates everything in the past, which is all-or-nothing thinking and he/she diminishes the positive events or achievements in the past.
I keep breaking things in my life and trying to put them back together. I don't want to break anything anymore. :'(	*Magnification*: Exaggerate the importance of your errors, fears, and imperfections.	The author gives greater weight to perceived failure or weakness but fails to know the positive side of mending relationships.
I feel like I'm ready to fuck things up, piss some people off and get my way. I know it is selfish, but I don't care.	*Emotional Reasoning*: Letting your emotions direct your conclusions about yourself, others, or situations.	The author lets his/her emotions direct their behaviours but refuse to consider the consequences.

Table 6.2: Pearson Correlations between negative affect, cognitive distortions and depressive symptoms (cumulative score)

	mean	SD	CES-D		SWL	
NA	10.91	11.835	0.192,	CI(-0.03,	-0.16,	CI(-0.37, 0.60)
			0.40)			
CD	0.532	0.981	0.300*,	(CI 0.08,	-0.250*,	CI(-0.45, -0.02)
			0.49)			

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, NA: negative affect, CD: cognitive distortion, CES-D: correlation with depressive symptoms. SWL: correlation with Satisfaction with Life, $p-value$ adjusted with Bonferroni correction. CI: 90% confidence level

Table 6.3: Correlations between negative affect, cognitive distortions and depressive symptoms (normalized score)

	mean	SD	CES-D		SWL	
NA	0.182	0.197	0.123,	CI(-0.06,	-0.108,	CI(-0.29,
			0.31)		0.08)	
CD	0.009	0.016	0.261*,	CI(0.08,	-0.208,	CI(-0.41,
			0.42)		0.02)	

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, CES-D: correlation with depressive symptoms. NA: negative affect, CD: cognitive distortion, SWL: correlation with Satisfaction with Life, $p-value$ adjusted with Bonferroni correction. CI: 90% confidence level

6.3.2 Cognitive Distortion and Personality Dimensions

To observe whether cognitive distortion reflects personality dimensions, we divided the sample into two groups according to the mean cognitive distortion score. This study included 26 participants at the higher end (CES-D score > 22) and 51 of them at the lower end (CES-D score $<= 22$). In our previous work that used the same dataset, we identified most of the variables were normally distributed (Chen et al., 2020c), we compared the depressive symptoms, satisfaction with life, and personality among the two groups with independent t-tests. Our results showed that users with different levels of cognitive distortion did not differ in personality dimensions (see Table 6.4, Bonferroni correction was applied to the test results).

6.3.3 Linguistic Styles of Individuals with High Cognitive Distortions

Next, we use a quantitative approach (LIWC) to examine users' linguistic style. LIWC is software to measure the degree to which different categories of psychological pro-

Table 6.4: Personality dimensions in users with high vs. low cognitive distortions (t-test)

	All (n=77)		High CD		Low CD			
	M	SD	M	SD	M	SD	p	d
ope	4.17	0.46	4.05	0.43	4.15	0.48	0.11	-0.37
con	3.18	0.74	3.09	0.74	3.09	0.75	0.41	-0.19
ext	3.10	0.84	2.84	0.65	3.09	0.90	0.03	-0.48
agr	3.54	0.68	3.46	0.64	3.52	0.71	0.44	-0.18
neu	3.02	0.87	3.15	0.78	2.96	0.92	0.33	0.22

Note: SD: standard deviation, M: mean, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ after Bonferroni correction. Effect size: 0.8 = large(L); 0.5= moderate(M); 0.2 = small(S), CD: cognitive distortion, d: Cohen's d, num. of posts: Number of posts in two months; ope: openness; con: conscientiousness; ext: extraversion; agr: agreeableness; neu: neuroticism.

cessing related lexicons are present in the text. For each Facebook post, we extracted scores that reflect various psychological processing with LIWC. Then we aggregated the LIWC score for each user. This score represents the linguistic style of the user. Table 6.5 shows the correlation between user linguistic style and cognitive distortion score.

Similar to previous findings on depression and written language (Rude et al., 2004), we also found that users with more cognitive distortions or negative affect in their social media text tend to use more first-person pronouns and less third-person pronouns. Our result showed that there were differences in the language characteristics between users who used more negative affective words and those who expressed cognitive distortions more frequently in their Facebook posts. Users who used more negative affective words in their posts tend to use more words that reflect social status and confidence, anger and swear words across all their posts. In contrast, users who expressed cognitive distortions more frequently in their posts tend to use more words that describe feelings, perceptual processes, risk and prevention, but fewer words related to rewards. Unlike users who used more negative affective words in their account, users who expressed more cognitive distortions did not use more words associated with anger, anxiety, and negative affect in their account.

Table 6.5: Cognitive distortion, negative affect and depressive symptoms

variables	NA	CD
social status and confidence	-0.518**	
authentic		0.325**
1st per pronoun	0.369**	0.325**
3rd per singular	-0.326*	-0.249*
2nd person	-0.235*	
verb	0.244*	
AFFECT WORDS		
negative affect	0.322**	
anger	0.413***	
anxiety		
sadness		
swear	0.385***	
COGNITIVE PROCESS		
differentiation	0.323**	
PERCEPTUAL		
perceptual process		0.255*
feeling		0.301*
BIOLOGICAL		
health/illness		
CORE DRIVE		
reward focus		-0.249*
risk/prevention focus		0.312**

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, NA: negative affect, CD: cognitive distortion,

6.4 Discussion

6.4.1 Identifying Cognitive Distortion in Social Media Text

Cognitive distortions are negative thoughts that reflect an individual perceives the reality inaccurately. We found that cognitive distortion can be identified on Facebook, but they only accounted for 1% of posts in our sample. Identifying cognitive distortion expressed in social media text is a challenging task. A social media post may contain unhelpful thoughts, but a single post does not reflect a thinking pattern. For example, Ophir et al. (2017) considered the following sentence "Wow! What a bad day and what a bad mood", as catastrophizing (exaggerating negative affect and events). This sentence could be a common exclamatory statement for someone who has a bad day. If the person complains about a bad day repeatedly, this may reflect a thinking pattern to exaggerate negative daily life events. Throughout the annotation process, we need to consider if the thought is just a general complaint of life or shows signs of unhelpful thinking.

Distinguishing types of cognitive distortion was an even more difficult task. Ophir et al. (2017) found that the annotation agreement for cognitive distortion types was low. We found that some posts contained multiple categories of cognitive distortions (see table 6.1). Therefore, we only annotated whether the post includes cognitive distortion in this pilot study, but we didn't distinguish the types of cognitive distortion.

6.4.2 Cognitive Distortions and Users' Depressive Symptom Levels

Our results reaffirmed that cognitive distortion extracted from social media text is associated with users' self-reported depressive symptom level ($r = 0.30$, $p < 0.01$). This finding echoed with Ophir et al. (2017)'s finding ($r = 0.75$, $p < 0.01$) for clinically depressed users. Cognitive distortion extracted from social media text was negatively associated with satisfaction with life but it was not correlated with users' personality dimensions.

Existing literature suggests that users with more depressive symptoms tend to use more words that reflect negative affect and anger on their social media wall (Ophir et al., 2017; De Choudhury et al., 2013; Tsugawa et al., 2015). Our finding suggested that not only users who used more negative affective words on their social media account tend to have higher depressive symptoms, but those who expressed cognitive dis-

tortion more frequently in their posts also had the same tendency. However, those who expressed cognitive distortion more frequently, including posting quotes and lyrics that contain unhelpful thoughts, did not necessarily use more negative affective words on their account. Instead, they tended to use words that describe perceptual process (e.g., feels, touch), reward (e.g., earn, win) and risk (Pennebaker et al., 2001), see Table 6.5. In contrast, participants who used more negative affective words in their accounts tended to use more words reflecting anger (e.g., swearing).

6.4.3 Future Direction

Researchers recently explore the machine learning approach to annotate cognitive distortion in social media text (Simms et al., 2017). There is a potential for researchers to develop automatic systems to detect cognitive distortions and use the detected distortions to infer depressive symptom levels. However, our findings suggested posts that contain cognitive distortion were rare on Facebook. There are many challenges for training models on an extremely imbalanced data set. To construct a balanced data set, recently Shickel et al. (2020) invited MTurk workers to write a short piece of text to describe their experience with different types of cognitive distortions. They trained psychology students to also annotate a set of mental health therapy logs. Shickel et al. (2020)'s dataset was more aligned with the therapy situation, whereas Simms et al. (2017)'s dataset was in the social media context. Currently, little effort has been spent in identifying cognitive distortion in written text. Detecting cognitive distortions in the text is still an under-explored topic. Future studies can explore the automatic annotation techniques on a larger annotated social media dataset.

6.4.4 Limitation

The myPersonality dataset was collected in 2011. Social media users are becoming more careful about what they post on the wall over the years as they gain knowledge that researchers are using their data to analyze their psychological traits (Schwartz et al., 2013, 2014), political views (Trottier and Fuchs, 2014) and mental health status (De Choudhury et al., 2014; Chancellor and De Choudhury, 2020). Therefore, our finding of the percentage of posts that contain cognitive distortion may not generalize well in the current Facebook data or data from other social media platforms.

We filtered out participants who do not have sufficient Facebook updates to allow analysis, thus, our final sample is relatively small. Given the size of the significant

effects we found in the data, power calculations indicate that a well-powered study should include data from around 200 users (Schönbrodt and Perugini, 2013).

It is important to note that most of our participants were white, young adults. The insights discussed here were drawn from a sample that contains only a small amount of Facebook posts identified as cognitive distortions. Our results have not yet been validated by large-scaled data. Therefore, our findings may not generalize well on another sample due to biases in the dataset.

6.5 Conclusion

In this study, we explored the possibility of identifying cognitive distortions from social media text. We developed a guideline to annotate cognitive distortions expressed in the social media text. We examined cognitive distortions expressed in social media text and their relationship with social media users' self-reported depressive symptoms, life satisfaction and personality dimensions. We studied the language characteristics of users who posted cognitive distortions more often and those who used more negative affective words. Our result suggested that although social media users rarely express cognitive distortion in their posts, compared with the amount of negative affective words they use, cognitive distortions have a stronger correlation with users' self-reported depressive symptom levels.

Chapter 7

Detecting Suicidal Ideations

In the previous chapters, we examine approaches to extract cognitive distortions and represent affective processes from social media text. We found these psychological processes extracted from social media text are associated with users' self-reported depression symptoms. Literature supported that severe and long-term depression would lead to suicidal ideations. Can we combine affective symptoms with behavioral symptoms to assess suicidal ideations? This chapter describes our system submission for the CLPsych 2019 shared task B on suicide risk assessment (Zirikly et al., 2019). We approached the problem with three separate models: a behavior model, a language model, and a hybrid model. For the behavioral model approach (support vector machine model), we model each user's behavior and thoughts with four groups of features: posting behavior, sentiment, motivation, and content of the user's posting. We used domain expertise to compile lists of vocabularies associated with suicidal ideations. These vocabularies turn out to be important features for the prediction model. For the language model approach, we trained a language model for each risk level using all the posts from the users as the training corpora. Then, we computed the perplexity of each user's posts to determine how likely his/her posts were to belong to each risk level. Finally, we built a hybrid model that combines both the language model and the behavioral model, demonstrating the best performance in detecting the suicide risk level.

7.1 Introduction

Every year, there are over 800,000 people who die of suicide (WHO, 2019). Although health care systems play a major role in assessing suicide risk, clinicians cannot assess

all the risk factors thoroughly given limited time. One of the most important warning signs for suicide is the expression of suicidal thoughts. The standard practice of clinicians asking people about suicidal thoughts cannot effectively predict and prevent suicide because most patients who died of suicide did not report any suicidal thoughts when asked by a doctor (McHugh et al., 2019; Chan et al., 2016). Therefore, many of them were assessed to have a low or moderate risk before their suicide attempts (Powell et al., 2000).

The CLpsych 2019 shared task B (Zirikly et al., 2019) attempts to address the challenge of automatic suicide risk assessment using people's forum postings. The task aims to distinguish the levels of suicide risks among users who posted any content in the suicide watch (SW) subreddit. The dataset includes all the posts (N = 31,553) in any subreddit from 621 users who had posted on SW. One of the four risk levels ranging from "No Risk" to "Severe Risk" was assigned to each user according to their SW posts. The annotation process is described in (Shing et al., 2018).

We treat the task as a multi-classification problem. We approach it with three models: a behavioral model (BM), a suicide language model (SLM) and a hybrid model (HM_{BM_SLM}) that combines the (BM) and (SLM) models. The SLM offers good classification accuracy, but it does not provide any human interpretable reason for its classification decisions. Hence, we define a collection of features to better capture users' posting behaviors and thoughts. Then we use these features in the BM. The overall results show that the hybrid model (HM_{BM_SLM}) performs the best in identifying the risk level with a f1-score 38% for the CLPsych task B.

7.2 Related Work

Suicide is a complex behavior involving biological, psychological and social factors. For psychological factors, a large amount of literature suggests that a history of psychiatric disorders, especially affective disorders, is a strong predictor of suicide (Angst et al., 2002; Brent et al., 1993; Bostwick and Pankratz, 2002). Another important precursor of suicide is self-harm or previous attempt. Biological and social factors that contribute to suicide include: substance abuse (Vijayakumar et al., 2011; Hawton et al., 2012; Bergen et al., 2012; Chan et al., 2016; Joiner, 2007), gender (males have a higher suicide risk) and living alone(Joiner, 2007).

The suicidal behavior model by (Wilson et al., 2005; Cukrowicz et al., 2011) proposed that the unmet need of belonging (e.g., relationship breakup) and the self-

perceived burden were the major motivations for suicidal behaviors (Trout, 1980). Other motivations include: having a negative self-image, hopelessness (Kovacs and Garrison, 1985), and having a plan of the suicidal attempt. The duration, intensity, and frequency of the suicidal desires also indicate the pertinacity of the attempt.

The majority of the prior work on the suicide risk detection focuses on manually generated (BoW) features centering only around the textual cues of the user's post (Varathan and Talib, 2014; O'Dea et al., 2015), such as the LIWC pre-trained word embeddings (Husseini Orabi et al., 2018) or supervised learning topics (e.g., latent Dirichlet allocation) (Ji et al., 2018). Unlike these studies, we design a model that leverages user's behavioral data in combination with a suicide language model to detect the suicide risk level. Our features intend to capture the language and behavioral characteristics proposed by clinical literature as suicide risk factors. For example, we develop a feature vector that represents suicide motivations. Examining the validity of these features in our experimental model provides us a way to understand the prevalence of these characteristics in people with different suicide risk levels.

7.3 Suicide Risk Prediction Models

In this study, we propose three models to measure suicide risk levels. BM uses user's posting behaviors and manual selected language characteristics to predict suicidal risk level. SLM learns the language characteristics of each risk level. The hybrid model (HM_{BM_SLM}) combines the advantages of the BM and SLM models.

7.3.1 Behavioral Model

Most of the existing studies focus on the language used in expressing suicide thoughts, and only a small number of them examine the behavioral and thought patterns on social media. For instance, (Colombo et al., 2016) use Twitter followers, friends, and number of retweets to represent the connectivity between users having suicide ideas. Based on the clinical literature, we engineer four sets of features that capture user behaviors and thoughts for the Behavioral model (BM), including posting behavior, sentiment, content, and motivation for suicide. Posting behaviors consist of users' posting patterns in SW, mental health related subreddits and all the other subreddits. Sentiment features consist of a sentiment profile for each user and the user's sentiment towards selected topics (e.g., friends and family). Content features consist of Linguistic

Inquiry and Word Count (LIWC) (Pennebaker et al., 2001), EMPATH (Fast et al., 2016) and count vectors normalized by TF-IDF. For the motivation features, we use a word count approach to define whether the user has suggested any motivations.

Some of these features were constructed using Suicide Watch (SW) posts only, while others were constructed using all the Reddit posts from the users. Although many of these posts might not be directly related to suicide thoughts, we hypothesized that using irrelevant posts to define a user's interaction behavior and emotional magnitude would help identify the users' virtual community with suicide risk.

7.3.1.1 Sentiment

Sentiment Profile. The sentiment of each user's previous posting was used to identify the similarity between users' postings. This set of features is represented as a sentiment value vector corresponding to a user's previous posting. Then, we use the Levenshtein Distance to compute the similarity between two such vectors (Yujian and Bo, 2007).

Topic Sentiment. We inspect the sentiment of specific topics in the SW posts. We extract the sentences containing keywords related to family members (e.g., mom, dad), partners (e.g., boyfriend), and self (e.g., myself). We then use SentiStrength (Thelwall et al., 2010) to detect the sentiment values of these sentences and aggregate the topic sentiment at a user level.

7.3.1.2 Posting Behaviors

Frequency of posting We use the number of posts, word count in each post, whether and when a user posts more frequently as features. To check whether a user has recently started posting more frequently, we define a posting frequency vector by computing the average posting time interval between any two posts from a user. We use a sliding window from the head to the tail of the frequency vector to identify which time interval(s) is at least one standard deviation below the mean of all intervals. Users are highly likely to post more frequently if the last window is one standard deviation below the mean. The frequency of posting is inspected in the SW posts, all user posts, and posts involving mental illnesses and drug use. To extract the posts involving mental illnesses and drugs use, we compile a dictionary of mental illnesses' names and symptoms. Posts that contain words from this dictionary are selected. Meanwhile, posts from subreddits that are associated with mental illnesses self-help groups (e.g., self-harm, TwoXADHD) are also extracted.

7.3.1.3 Motivation Factors

Financial problems, drug use, mental illness history, relationship break up, hopelessness, suicide tools and self-harm have been found to be predictive to suicidal behaviors (Kessler et al., 1999). In our study, we compile dictionaries for each of the motivation factors. Terms in drug use, mental illness and suicide tools dictionaries are extracted from websites using the web scraping techniques.

7.3.1.4 Content Feature

We use both the open and closed BoW approaches to generate the content feature. For the open vocabulary approach, we counted the term frequency and normalized it with tf-idf. For the closed vocabulary approach, we used LIWC and Empath. Both tools are used to count words from predefined psychologically meaningful categories.

7.3.1.5 Clustering

We use model-based clustering (Banfield and Raftery, 1993) to group sentiment, posting behavior and motivation factors. Model-based clustering assumes that the data are formed by multiple Gaussians. The clustering algorithm tries to recover the models that generate the data. The best model is selected according to the Bayesian information criterion (BIC). We adopt five clusters as our solution.

7.3.2 Suicide Language Model

The behavioral model (BM) enables us to observe the behavioral and thought differences among individuals with various suicide risk levels. However, one disadvantage of the BM approach is that we might miss some relevant cases that do not contain words in the manually selected dictionary or include irrelevant cases but contain the dictionary words.

With this challenge in mind, we also tackle the suicide risk classification problem with suicide language modeling (SLM). Language modeling is used in domains such as machine translation, speech recognition and text classification (McCallum et al., 1998; Brants et al., 2007; Coppersmith et al., 2014). The principle of language modeling is to compute a probability distribution over words to determine how likely a specific language model is to generate a given document. In our case, we train one model for

each risk level. Then, we calculate a document's likelihood (perplexity) for all the models, and select the model with the best score.

7.4 Dataset and Experiment Setup

The dataset used for training the models is provided by the CLPsych shared task B (Zirikly et al., 2019). It contains 621 Reddit users who had posted on SW with an overall of 31,553 posts. The users are labeled as "no risk" (class A), "low risk" (class B), "moderate risk" (class C), and "severe risk" (class D). Dataset statistics is presented in table 7.1. The training set shows that nearly half of the posts were labeled as "severe risk", class B only accounts for less than 10% of the posts. Nearly half of the posts in both the training and testing sets did not have any contents in the post body.

Table 7.1: Distribution of posts and users in the training and test set

Training	P (%)	WC	U	P/U	SW/U	emP
A	10662 (34%)	52	127	84	1.28	6070
B	2715 (9%)	101	50	54	1.18	984
C	5726 (18%)	79	113	51	1.36	2556
D	12450 (39)	72	206	60	2.64	5344
Test	9610	63	125	77	1.49	4704

Note: A:no risk, B:mild risk, C: moderate risk, D: severe risk. P: number of posts. WC: average word count in posts. U: users. P/U: posts per user. SW/U: SuicideWatch post per user. emP: posts without content in the post body.

7.4.1 Suicide Language Model Setup

We train the (SLM) language model with the minimally processed data (raw text), and tokenized and truncated data. For the raw text model, the data are preprocessed as follows: Sentences are split by the NLTK sentence splitter and then spaces are inserted around each full stop to make sure misspeled cases are parsed correctly. For example, "tomorrow.And today" is processed as "tomorrow . And today". For the tokenized and truecased model, we apply the tokenizer from the Moses machine translation toolkit (Koehn et al., 2007).

The language model is trained with KenLM's default settings (modified Kneser-Nay smoothing) (Heafield et al., 2013). In each model, all the posts from a Redditor

and annotated with a specific risk level are used as the training corpora. All the posts from a Redditor are treated as a single document. To assign a risk level to the document, we calculate its perplexity for each language model and assign the document's class based on the language model that produces the lowest perplexity score. We experiment with the context windows of 3 to 6-gram and find that 4-gram works the best.

7.5 Experiments

In the SLM, for each document, the model with the lowest perplexity is assigned to the document. Perplexity is the inverse probability of a test set, normalized by the number of words, a low perplexity indicates that the probability distribution is good at predicting the sentence (Sennrich, 2012). Given a sample test, we calculate its likelihood for all the models, and select the model with the best score.

In the BM, we used feature set 4 (see Table 7.4 in our final prediction model). We validate our BM features on the multi-classification problem using support vector machines (SVM) in scikitlearn [1]. We use the 5-fold cross-validation on training data and grid-search parameters to explore both the kernels and margin of the hyperplane (C parameter).

Furthermore, we construct a hybrid model based on our observations on the prediction results from the SLM and the BM. In the training process, we observe the BM is weak in distinguishing classes B and C, but the SLM is better in identifying class B. Therefore, we adopt the class B results from the SLM. We also find that some posts in class A are suicide experiences from someone associated with the authors, but not the authors themselves. The BM is better than the language model in identifying these cases, so we use the BM for class A. However, if the confidence score is lower than 0.4, the SLM can better identify class A. Therefore, we replace the results with confidence score lower than 0.4 with those from the SLM model.

7.6 Results

Table 7.2 shows the test set results of the three models. Table 7.3 shows f1 for flagged vs. non-flagged and urgent vs. non-urgent. Flagged vs. non-flagged distinguished class A from the rest of the classes. Urgent vs. non-urgent distinguished classes A, B with

[1] https://scikit-learn.org/stable/

classes C, D. The hybrid model had the best average f1 macro in the risk assessment task.

Table 7.2: Results for risk assessment task

Model	Risk level	P	R	F
BM	A	53	78	63
	B	22	15	18
	C	14	14	14
	D	55	42	48
	$F1_{AVG}$			36
SLM	A	73	25	37
	B	27	23	25
	C	12	7	9
	D	49	83	62
	$F1_{AVG}$			33
HM_{BM_SLM}	A	56	72	63
	B	25	39	30
	C	12	11	11
	D	55	42	48
	$F1_{AVG}$			**38**

P: precision (%), R: recall (%), F: f1 macro average (%). $F1_{AVG}$: f1 (%) macro average of four classes.

Table 7.3: Results for flagged and urgent cases

	Flagged			Urgent		
	P	R	F	P	R	F
BM	91	76	83	80	69	74
SLM	79	97	87	69	89	78
HM_{BM_SLM}	89	81	85	81	65	72

P: precision (%), R: recall (%), F: f1 macro average (%).

In our test set result, we find that SLM is over-fitting. SLM classifies most of the posts to class D in the testing set. Whereas the BM has consistently good performances on classes A and D, but poor performances on classes B and C.

7.7 Conclusion

Our results demonstrate that suicide risk can be gauged by user's posting behaviors. Suicide risk factors identified by clinical literature are useful in the automatic detection of suicide risks. Suicide language can be modeled by statistical language model, especially for risk level B and D, in which cases it surpasses the behavioral model. Hence, a combination of the two models results in a more accurate user classification. As future work, further analysis of each feature would gauge its contribution towards identifying suicide risk levels.

7.8 Important features

In this section, we show our model performances on the training set with different feature groups. We divided the training set into a train and validation set with a ratio of 0.3: 0.7. Table 7.4 shows the model result on the train and validation set. Our model on the published work adopted feature set 4. We found that adding the manually defined topics have dramatically improved the prediction results. These topics were manually defined dictionaries that reflect certain themes, for example, financial difficulties (e.g., bankrupt), drug abuse, mental health, relationship problems (e.g., break up), suicide (e.g., kill, cut) and hopelessness (e.g., no hope). The vocabularies of the dictionaries were documented on Github [2]. Existing lexicon-based topics analysis (e.g., LIWC and Empath) and unsupervised learning topic modeling (e.g., LDA) have been widely adopted as tools to generate features for systems that identify suicidal ideations and mental illness symptoms. These tools were also adopted in Chapter 4. Our feature analysis results suggested that expert knowledge is critical for feature constructions of the machine learning system.

7.9 Review and Next Step

Suicidal behavior is a transdiagnostic outcome for many mental illnesses. It is a complex but preventable health problem. A majority of the suicide decedents consult with their physicians days to weeks before committing suicide (Juurlink et al., 2004). Early detection of suicidal risk can be an intervention opportunity. Application of machine learning, especially to electronic medical records, yields promising results (Sanderson

[2]https://github.com/luciasalar/suicideDetection/tree/master/dictionaries

Table 7.4: Model results on training and validation set

feature group	risk level	precision	recall	f1-score
set 1: count vector, psychological processes	A	81.4	57.8	67.8
	B	100	6.7	12.5
	C	20.8	14.7	17.2
	D	54.6	85.5	66.7
	$F1_{AVG}$	41.0		
set2: count vector, psychological processes, PF health, methods	A	81.4	57.8	67.8
	B	100	6.7	12.5
	C	20.8	14.7	17.2
	D	54.6	85.5	66.7
	$F1_{AVG}$	41.0		
set 3: count vector, psychological processes, PF health, methods, PF SW	A	74.1	60.5	66.7
	B	100	6.7	12.5
	C	22.2	17.6	19.7
	D	55.6	80.6	65.8
	$F1_{AVG}$	41.2		
set 4: count vector, psychological processes, PF SW, methods, manual topics	A	82.1	60.5	69.7
	B	100	6.7	12.5
	C	31.4	32.4	31.9
	D	58.8	80.6	65.8
	$F1_{AVG}$	45.5		
set 5: count vector, psychological processes , PF health, methods, manual topics, sentiment	A	79.3	60.5	68.7
	B	100	6.7	12.5
	C	26.5	26.5	26.5
	D	56.5	77.4	65.3
	$F1_{AVG}$	43.2		
set 6: count vector, psychological processes, PF health, methods, manual topics, sentiment, LDA	A	71.8	60.5	65.7
	B	0	0	0
	C	24.1	20.6	22.2
	D	56.8	80.6	66.7
	$F1_{AVG}$	38.6		

F1: weighted average f1 score. count vector: 300 n-gram count vector ($n = 1,2,3$). psychological processes: LIWC (Pennebaker et al., 2001), EMPATH (Fast et al., 2016). PF health: posting frequency in health related subreddits. Method: suicidal methods or tools mentioned in the posts. PF SW: posting frequency in r/SuicideWatch.

et al., 2020; Zhong et al., 2018). In addition to this, detecting suicidal thoughts in social media platforms is a potential promising screening technology because studies found that young people are likely to disclose suicidal ideations and suicidal risk on social media platforms (Roy et al., 2020).

In the CLpsy shared tasks, users posted on r/SuicideWatch were annotated suicidal risk according to their thoughts, including explicit ideation of suicide, the thinking pattern that they self-perceived as a burden to others, lack of hope for things to get better, a sense of impulsivity, talk about methods of suicide, previous attempts, life-changing events or isolation from friends and family. We constructed features that matched with the risk factors suggested by suicide prevention experts. For example, we used a keyword approach to capture motivation for suicide, the sentiment of sentences related to family, friends and self, lexicons for drug use and suicidal tools. It turned out that these topics are more important features than sentiment and topics learnt by unsupervised learning approach.

Zirikly et al. (2019) summarize the models for this shared task, most of the papers use pre-trained embeddings, LIWC, n-gram features, posting time and LDA topic modelings as features. Our work also identifies keywords linked to motivations and suicide-related unique identifiers. The submissions of the task include traditional machine learning and deep learning models. SVM and logistics regressions are frequently used ML models. We only submitted the behavioral model published in our work for evaluation, and this model ranked No.4 among all the submissions. However, our hybrid model, combining the behavioral model with the language model yields better performance.

For early detection and screening for suicidal risk, it is important for a model to have high recall so that fewer high-risk cases would be missed out. Deep learning techniques are good at achieving a high model performance. Therefore, many popular deep learning architectures were adopted in the CLpsy share task. For example, Hevia et al. (2019) used a GRU-based RNN. Morales et al. (2019) used CNN, LSTM, Matero et al. (2019) purposed attentive RNN and BERT.

Among all the submissions in this task, we were the only team that attempted to use language models to capture psychological factors across various constructs, such as stress, loneliness, burdensomeness and hopelessness. We trained the language model using all the Reddit users' posts, not just the posts from r/SuicideWatch. Our language model yields a slightly worse result than the behavioral model. Future work can consider only using relevant subreddits for training. Our training sample was relatively

small (around 100,000 lines). Considering that many language models usually involve millions of lines as input (Devlin et al., 2019). In this work, we only explore the statistical language model due to a small training example. Future work can try the deep learning language model on a bigger training set.

For future direction in suicidal detection technology, researchers can explore models using data from different social media platforms. The CLpsy2019 shared task models are not generalized to other social media platforms because users have different self-disclosure levels and language characteristics in various social media platforms. It is important to model suicidal language on other social media platforms as well. For example, Roy et al. (2020) used ML approach to identify suicidal ideations in tweets.

Chapter 8

Conclusion

In the preceding chapters, our overall goal was to explore methods to using social media signals to represent the psychological processes underlying depression. In this chapter, we summarize our findings, connect our results to theories of psychopathology, and discuss the limitations of the research, ethical concerns, and future directions.

8.1 Summary of Contribution

In order to answer our research question, we have proposed three sub-goals: monitoring affective patterns, identifying cognitive distortion, and identifying topics related to risky behavior (see Figure 1.1 in Chapter 1). In this section, we summarize the findings for each sub-goal as below:

8.1.1 Monitoring Affective Pattern

The first goal of the research (see Figure 1.1 in Chapter 1) was to examine how affective patterns manifested in the social media text. We represented the magnitude, categories, variation and transitions of affect using signals extracted from social media text. These dimensions of affect are documented in the affective theories in psychology. Before we started the research on Chapter 3 and Chapter 4, we found that most of the studies in this research line aggregated affective language over a long period. However, such a simple representation of affective language did not align with the structure of mood described in the psychology literature. Mood variation, magnitude and transitions, are important for understanding the development of depression in the clinical context (Akiskal, 1996; Davidson, 1998). Throughout our work (Chapter 3 and

4), we quantified the affective patterns extracted from social media text according to dimensions and concepts of mood defined in the literature on affect (see Section 4.2.1 in Chapter 4).

By experimenting with various structures to represent different aspects of mood, we found that mood patterns extracted from social media text were associated with personality and depressive symptom levels. These findings echoed with findings from psychopathology and personality literature (Akiskal and Van Valkenburg, 1994). Our studies connected the design of machine learning experiments and their results with theories in affect (Akiskal, 1996; Watson, 2000). Finally, we found that not only user-generated content reflects users' mood, but copy-and-paste in-text lyrics and quotes also reflect users' depressive symptom levels (Chapter 5).

8.1.2 Cognitive Distortion

Cognitive distortion is an irrational thought pattern involved in the onset or perpetuation of depression and other psychopathological states. The second goal of our research was to identify cognitive distortion in social media text (see Figure 1.1 in Chapter 1). In Chapter 6, using a sample of more than 4,145 Facebook posts from hundreds of users, we identified 41 posts showing cognitive distortions.

The main contribution of Chapter 6 included an annotation guideline to identify cognitive distortion from social media text. We reaffirmed that cognitive distortion could be identified in a general population of Facebook users. The amount of cognitive distortion in one's social media text was slightly positively associated with social media users' self-reported symptom levels. This finding may provide insights on whether cognitive distortion on social media posts can be another important indicator for depressive symptoms besides affect, social network, posting time and topics.

8.1.3 Topics specific to risky behavior

For the last goal, we aimed to use topics extracted from social media text to evaluate users' suicidal risk levels. In the study "Similar Minds Post Alike: Assessment of Suicide Risk Using a Hybrid Model" (Chen et al., 2020b), we compiled topic dictionaries that are relevant to suicidal risk according to psychopathology literature, including the motivation (e.g., financial crisis), drug use, somatic complaints, and so on. This work's major contribution was that we found that these topic features are associated with users' suicidal risk levels. They are more associated with the self-reported

symptom levels than sentiment and topics generated with an unsupervised learning approach. We also attempted to use a statistical language model to classify risks and we found the statistical language model captured the language characteristics of those who had suicidal ideations.

8.2 Connecting Experiment Design and Results to Theories of Psychopathology

This section explains how we mapped our experimental design and the resulting implications to existing theories of psychopathology.

8.2.1 Connecting Affective Patterns from Social Media Data to Theories of Affect

We linked the affect representation to the concept of mood. Literature defines that mood as a form of affective experience that runs in the background (Akiskal and Van Valkenburg, 1994). Therefore, we used averaged sentiment or the most frequent sentiment over a short period to represent mood. Mood is also dynamic. To define a dynamic mood, we split a large period into many small time windows. Unlike the existing studies, we tested the effect of time window size and step size on the affect representations.

Literature on affect also suggested that mood intensity, duration, variability (frequent and extreme changes in mood or emotion over time) are important aspects for affective disorders (Larsen, 1987; Akiskal, 1996). In Chapter 3 and Chapter 4, we included mood magnitude and alternations in the mood representations. Chapter 4 quantified mood fluctuation using parameters in the Gaussian process regression. We found that users who had a high level of depressive symptoms showed more instability in their mood pattern extracted from social media data. This finding aligned with the psychopathology literature that suggested mood instability and irritability were related to depression, although they were not core symptoms (Balbuena et al., 2016).

Literature on personality and mood suggested a structural convergence of mood and personality. For example, mood levels were related to extraversion and negative affect was strongly related to neuroticism (Hepburn and Eysenck, 1989). In Chapter 3, we found that users who scored high in neuroticism tended to post content with

negative affect continuously. They did not simply post more negative content (Chapter 3). Chapter 3 measured sentiment on consecutive days, which was more aligned with the structure of mood. This finding is different from most of the existing findings that suggested social media users with high depressive symptom levels tended to express more negative sentiment in their posts on average.

8.2.2 Lyrics Indicate Mood Regulation

It is well known that music can evoke a wide range of feelings. Music is especially charming and pleasurable when it deals with sadness (Sachs et al., 2015). People often choose music that is in congruence with their mood. Listening to songs centered around hurt, pain, and grief is part of the mood regulation process for coping with stressful life events (Hamilton et al., 2013; Gladding et al., 2008). Lyrics have also been found to be associated with one's mental state. Studies have found a relationship between various music and people's behaviors, even vulnerability to suicide. For example, "emo" music is related to girls' mental state (Baker and Bor, 2008).

We unveiled that social media users with more depressive symptoms tend to post more in-text copy-and-paste lyrics, especially lyrics with negative sentiment (see follow-up study in Chapter 5). We believe that posting song lyrics and quotes on social media walls may indicate mood regulation behavior, especially if sad lyrics are posted.

8.2.3 Linguistics Style from Social Media Users Posted Cognitive Distortion

This chapter found that cognitive distortions extracted from social media posts were associated with users' self-reported depressive symptom levels. However, compared with Ophir et al. (2017)'s finding, our correlation result had a smaller effect size. We speculate the difference is due to sampling differences. Ophir et al. (2017)'s sample only contained participants who received therapy treatment.

We also found that users with more depressive symptoms tend to display more cognitive distortions, but they did not necessarily use more negative affective language in their account. Existing literature on mental health predictive techniques often focuses on using affect, topics and social networks as features. This approach may miss out on high depressive symptom participants who are not using more negative affective language. Our finding on cognitive distortion suggests these participants may be identified by the cognitive distortions expressed on their posts.

8.2.4 Topics markers indicate suicidal risk

In the paper "Similar Minds Post Alike: Assessment of Suicide Risk Using a Hybrid Model" (Chen et al., 2019), we attempted to use language that reflects factors that contribute to suicidal ideations to classify suicidal risk. It is worth noting that the clinicians' evaluation of risk levels according to users' Reddit posts did not reflect if the user will commit a suicidal act. During our research, we identified multiple themes defined in the psychology literature, for example, that severe depression was one of the leading causes of suicide. Therefore, we identified topics related to somatic symptoms, biological processing and medications because these were important features to infer users' depressive symptoms in the social media context (see Chapter 2).

We were aware that the reasons for an individual to commit suicide are complex. Other factors such as socioeconomic status, stressful life events and low self-esteem also expose one to a greater risk of committing suicide. We found that users who posted in Reddit r/SuicidalWatch using keywords associated with the themes listed above had a higher risk of committing suicide. Nevertheless, those who had a more negative sentiment towards "self-worth" and "family and friends" were also at higher risk. Our findings suggest that suicidal risk factors defined in psychology literature can be detected by examining these topics in their posts using a machine learning algorithm.

8.3 Limitations

Similar to existing studies using social media data to infer mental health status, there are challenges of validity and sample bias.

8.3.1 Validity

Validity refers to whether we measure what it is claimed to measure. There are concerns about the validity in studies using social media data to predict mental health status.

- **Construct validity** Existing literature, included our studies, has often used self-reported mental health status as the "gold standard", however, self-reported status does not reflect a diagnosis. When people seek help for mental health issues in a clinical context, they should receive a detailed assessment from a specialist. The assessment considers the symptoms, feelings, thoughts, physical health,

employment, financial circumstances, social and family relationships, sexuality, drug and other issues that may affect mental health (NHS, 2020).

- **External validity** Effect observed on a particular social media platform may manifest differently on other platforms due to the specific functionality of the platform, as well as, cultural and demographic differences of the users. For example, anonymity and social ties increase self-disclosure level and content of negative emotions (Ma et al., 2016).

8.3.2 Data Quality

Data quality can be measured by data accessibility, the quantity of data, data believability, completeness and 12 other dimensions (Pipino et al., 2002; Immonen et al., 2015). Data from social media platforms is characterized by its high availability. Researchers can retrieve a large amount of data in a short period. However, social media data is poor with regard to its truthfulness and its credibility. For example, it is difficult to ascertain whether the profile information about a person's age and gender is true. Social media data is also full of spelling errors and culturally specific jargon, making the interpretation of the language and symbols challenging for some of the algorithms. Its sparsity is also a well-known shortcoming that hinders researchers from obtaining a complete picture of the users' lives.

8.3.3 Sample Biases

The samples used in the first part of the thesis mainly contained hundreds of participants from developed countries, and the majority of them were of white ethnicity. On top of this, users on social media platforms are mainly young adults. For example, around 90% of Redditors are under the age of 35, with a mean age of 25 years (Bogers and Wernersen, 2014). Mislove et al. (2011) analyzed a set of Twitter users which represented 1% of the U.S. population by estimating their gender and race according to names on the profile. They found that Twitter users were predominantly male (72%) and Caucasian (86%). Users also significantly over-represent the densely populated regions of the U.S. Facebook tends to have a more age-balanced group of users. Ribeiro et al. (2020) recently analyzed 230 million Facebook users and found that there were 16% of people range from 18 to 24, 25% from 25 to 34, 19% from 35 to 44, 15% from 45 to 54 and 12% from 55 to 64. However, 65% of the users were Caucasian.

Therefore, it is not possible to have a balanced sample using data sources from social media. Older people are in general underrepresented across platforms. Moreover, studies using social media data sources only involve those who are willing to share their data. In our study, we only involved those who completed a mental health assessment questionnaire.

Chapter 2 summarized findings from more than 100 papers in this field, and some findings were common across different samples. Findings supported by multiple existing studies are more applicable to other samples. In contrast, our findings in Chapter 3 and Chapter 4 were based on a single, isolated dataset and were not validated by other studies. These findings may not be generalized well to another dataset due to different self-disclosure levels and demographic characteristics in the new sample. To tackle sample biases, we need more future studies to replicate our findings.

Another limitation was that our studies only focused on English language data. Findings from our studies may be different from samples from other cultural backgrounds. According to our overview in Chapter 2, only around 10% of the studies examined non-English data, among which half studied Chinese language data. More studies need to be conducted in different languages to gather clinical insights across cultures. Nevertheless, multiple of our studies in this thesis used a Facebook dataset collected during 2011-2013. The language on social media platforms users' motivation to use the platform evolves over the years (Eisenstein et al., 2014; Nadkarni and Hofmann, 2012). For example, Facebook users are motivated by two primary needs: 1) The need to belong and (2) the need for self-presentation (Nadkarni and Hofmann, 2012). In a recent study, Kuru et al. (2017) found that the use of Facebook, especially on a mobile device, is mainly driven by habitual behavior. Users tend to use the platform to browse information nowadays rather than to maintain self-presentation.

We are aware that our "gold standard" label does not represent a diagnosis. Our approaches in this work did not imply a screening or diagnostic technology. Our result cannot be generalized to other social media platforms or users from different cultural backgrounds than those in our datasets. Here we discuss these limitations in more detail.

8.3.4 Small Sample Size

The scale of sample size posed constraints on the techniques we could try. Some of the methods we tried require a larger sample for training. For example, we attempted to

build language models for suicidal language in Chapter 7. We used all the Reddit posts (N=31,553) from 621 users. This approach produced promising results, but a larger dataset of millions of lines is needed to achieve satisfactory results using a statistical language model. An even larger sample is required to try the state-of-art deep learning model.

8.3.5 Sentiment Detection Techniques

In our published papers reported in Chapter 3 and Chapter 4, we used a dictionary matching approach to detect sentiment in the text. Although the algorithm we used, SentiStrength (Thelwall, 2017), has been validated in many studies, we found that using a pre-trained deep learning model (huggingface, 2021) to identify sentiment in the text demonstrated more accurate results. Algorithms that used dictionary matching approaches, such as LIWC, EMPATH, and ANEW, were popular approaches for sentiment extraction in mental health status detection techniques. We suggest future studies explore deep learning approaches for extracting sentiment features.

8.3.6 Detecting Emotions in Lyrics and Quotes

In the study "It's Not Just About Sad Songs: The Effect of Depression on Posting Lyrics and Quotes" (Chen et al., 2020b), we identified the positive, negative and neutral affect in lyrics. Literature about music and depression often found people listening to sad songs as a coping mechanism for mood regulation. Our study conducted sentiment analysis on each post that contains lyrics, but we did not perform emotion analysis on each lyric. Sentiment analysis only involved positive and negative valence. Emotions are more fine-grained. For example, Ekman (1999) identified six basic emotions: anger, disgust, fear, happiness, sadness and surprise. Cowen and Keltner (2017) identified 27 emotions using self-reported methods. We manually analyzed the emotion words by observing the most frequent topic keywords. This approach allowed us to approximate the emotions in the lyrics in general, but not on a document level. Future studies should use algorithmic approaches (e.g., Emotxt (Calefato et al., 2017)) to extract emotions from song lyrics. These emotions may be congruent with those the user was experiencing when they posted the status updates.

8.3.7 Annotation for Cognitive Distortions

In the study "Examining Cognitive Distortions in Social Media Text", we annotated whether a status update contained cognitive distortion. Different types of distortions were not distinguished because we encountered a problem similar to Ophir et al. (2017). We found that a status update may contain more than one type of cognitive distortions and it may be difficult to reach an agreement between raters. On top of this, we did not have a second-rater to annotate cognitive distortions due to funding restraints. We plan to have a second annotator to co-annotate this task in our future work.

8.4 Contribution to real-world intervention

The contribution of this work mainly lies in the techniques of representing the social media signals that include various aspects of mood. Content such as lyrics and quotations are also suggested to be included in the representations. We adopted a classifying approach to establish links between the representations we constructed in self-reported mental health status. The limitations of the dataset prevent us from using these classifiers in a real-world context. All the existing studies using social media data to detect mental health status have similar limitations. Social media data alone does not contain all the necessary information for diagnosis. Therefore, systems providing a risk level classification based on social media data alone are not very meaningful for clinicians to understand a patient's condition. Our contribution to real-world intervention is to allow clinicians to observe fine-grained documentation of psychological processes, such as mood and distorted thinking patterns. This information can enable clinicians to obtain a more comprehensive picture of the trajectory of psychological processes underlying various psychological disorders if combined with other available sources. It may also allow patients to have a better recall of their life records during a therapy session.

Mikal et al. (2016) pointed out that some participants envision they would benefit from a computer system that tracks their mood using social media data, especially that it may provide extra information to therapists. However, studies about human interaction with the system are still rare. Future research should consider conducting surveys to understand the public's opinion on using this technology.

8.5 Ethical Challenges

This section highlighted our research ethics and discussed the ethical challenges of using social media data to study users' mental health status. Our studies adopted rigorous procedures to safeguard user privacy and followed the best practice in research ethics.

8.5.1 User Privacy

Social media users often consider themselves responsible for protecting their privacy (Mikal et al., 2016; Golder et al., 2019). However, many social media users have misconceptions about the social media platform's functionality. For example, deleting posts doesn't mean the data is not available. Having a small number of followers does not imply that people cannot find your account. Due to misunderstandings regarding data reach and permanence, researchers need to safeguard users' data.

Many studies published social media datasets to encourage reproducible experiments. Zimmer raised the ethical concern of publishing Facebook data (Zimmer, 2010). Although published social media datasets are often anonymized, some posts contain contact information, personal identifiers, and health information (Honey and Herring, 2009; Naaman et al., 2010). Yet, datasets with personal information pose potential privacy threats to social media users. Our studies from Chapter 3 - 6 used the MyPersonality dataset. MyPersonality stopped sharing data with other scholars in 2018 due to the founders' lack of resources for maintaining the database and responding to inquiries (Kosinski, 2021). I was granted to use the MyPersonality in 2017 for studying social media users' satisfaction with life (Chen et al., 2017) and depressive symptoms. We were not able to publish the annotated dataset of MyPersonality, because this would have violated the terms of use of the data set.

8.5.2 Research Ethics

Our research complies with General Data Protection Regulation (GDPR)'s definition of pseudonymization. Article 4(5) GDPR defines pseudonymization as:

Pseudonymisation *The processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person.*

According to GDPR's definition of pseudonymization, the data must be modified to prevent direct identification and be protected against re-identification (Mourby et al., 2018). The myPersonality dataset and the CLpsy Reddit dataset used in our studies have gone through at least one standardized anonymous procedure – a unique identifier that does not reveal users' "real world" identity was assigned to each user. There was no direct way to identify users who wrote the posts without combining external sources. For the myPersonality dataset, we extracted thousands of posts for the sentiment and cognitive distortion annotation tasks. Before we distributed the sample for annotation, human names were removed from the select posts to prevent re-identification.

Our studies using the myPersonality dataset obtained the research ethics approval from the School of Informatics at the University of Edinburgh. The studies involving human participants were reviewed and approved by Self-Certification according to the procedure of the School of Informatics, University of Edinburgh. The secondary analysis of this data set was reviewed and approved by the Ethics Committee of the School of Informatics, University of Edinburgh, Reference Number 72771. For the CLpsy Reddit dataset, we applied to use the dataset from the Shared Task organizers (CLpsy, 2019) and we obtained ethical approval from the Educational Testing Service. All the datasets for our research were stored on university servers. Only I had access to all the data. Co-authors could access data samples shared on a platform hosted on the university server. In addition, we hosted an annotation tool on a server located in the UK for the annotation task.

8.5.3 Social Media Users' Opinion on Social Media Mental Data Health Predictive Techniques

Social media users recognize that techniques to predict mental health status from social media have potential for social good, although some of them expressed concern regarding privacy (Mikal et al., 2016; Golder et al., 2019). One primary concern was that users were afraid that the tools used to pinpoint individuals' mental health status could identify and stigmatize them. For example, in Mikal et al. (2016)'s study, some interviewees expressed that they would unfollow a user if they knew the person had depression. Techniques for classifying depression may cause harm in terms of misleading results and risk of stigmatization.

Although there were concerns about privacy and potential harm, many respon-

dents had a positive opinion about tools that can accurately assess their moods. People suggested that their social media record might be able to help them respond to their therapists' question (Mikal et al., 2016). Many respondents also felt that automated social media tracking could allow mental health practitioners to observe them through a broader window. It could provide them with some objective evidence of mood swings and duration (Mikal et al., 2016). Our thesis's goal echoed with respondents' positive opinion about using social media data to monitor mood pattern Mikal et al. (2016). We shifted our focus from the classification of mental health status to developing approaches to improve the representations of social media signals. We believe our findings contribute to the understanding of psychiatric illnesses.

Future researchers of mental health predictive techniques need to think more carefully about the benefits of developing classification techniques. Is the benefit greater than the risks? The potential benefit of the research is the most influential factor in determining whether participants would give researchers consent to collect their data (Golder et al., 2019).

Therefore, more studies need to be conducted with a broader range of samples to understand the potential benefits of research in this field.

8.6 Future Work and Implications

Interest in using social media data to infer users' mental health status has been growing over the last few years. In this section, we briefly discuss the directions for future explorations.

8.6.1 Reducing Sample Biases

The greatest limitation of using social media data to infer mental health status was that existing findings were often based on single, separated and biased datasets. There was no benchmark dataset to compare the performance of the algorithms. The small sample size also limited the approaches we could explore on the optimization. For example, in the CLpsy suicide risk assessment task, our statistical language model did not perform very well in some classes, mainly due to the small training sample.

Moreover, most current mental health status findings using social media data did not generalize well to different platforms and users from various cultural backgrounds. For example, while our literature overview in Chapter 2 summarized that users with

more depressive symptoms used more negative affective words, Chen et al. (2020a) found that Chinese social media users with more depressive symptoms tend to use fewer words that were positive affective. In the future, researchers should examine if the existing findings can be replicated in samples of different demographic groups.

Future work should focus on collecting a high-quality dataset with participants from a more varied demographic background to tackle the challenge of sample bias. For example, the OurDataHelps project from the University of Maryland (UMD, 2021) is collecting a large-scale dataset to study how language usage and language changes may be connected to psychiatric symptoms.

8.6.2 Using Social Media Signals to Provide Insights for Psychopathology Research

The previous section (8.5.3) explained social media users' mixed attitudes towards using their data to predict their mental health status. Users prefer technology that can assist their therapy treatment (Mikal et al., 2016). However, existing research mainly focused on framing the problem as a classification task, and the goal was to optimize functional algorithms. Few studies explored approaches to extract social media signals to provide insights to users or mental health practitioners.

Future studies can continue exploring approaches to extract social media signals to allow mental health practitioners to observe more information about the patients' lives. For example, how can we accurately retrieve adverse life events from status updates? Do users' emotions toward significant life events, family and friends on social media text provide insights for researchers and mental health practitioners? Do mood swings recorded in social media text help patients retrieve memories important to the treatment? Future research can explore many more exciting questions with social media data. These questions provide more meaningful insights to researchers than optimizing classification problems.

Chapter 9

Appendix

.1 EXPERIMENT AND RESULT DETAILS for Chapter 4

The following supplementary material details what is required to reproduce our results as closely as possible.

.1.1 MODEL TRAINING

Grid searches of the following pairings of parameter spaces and Scikit-Learn implementations of algorithms were carried out:

- Feature Extraction
  - number of n-gram: 1000, 1500, 3000, 4000, 5000, 6000
  - number of topics: 10, 20, 30
- HMM:
  - Initial transition probability: [0.5, 0.5], [0.5, 0.5]
  - Initial transition probability: [0.2, 0.3, 0,2, 0.3], [0.2, 0.2, 0.3, 0.3]
  - Number of iteration: 10
- Support Vector Machine
  - Inverse of regularization strength: 0.5, 0.7, 1.0, 1.5, 2.0, 2.5
  - Kernel: linear, poly, rbf, sigmoid
  - Kernel coefficient: 0.01, 0.001, 0.0005

- Extra Trees
  - Number of Estimators: 100, 300, 500, 1000
  - Maximum Tree Depth: 20, 50, 100, 200
  - Maximum number of features: sqrt, log2
- Logistic Regression:
  - Penalty: l1, l2
  - Inverse of regularization strength: 0.1, 0.3, 0.5, 0.7, 0.9, 1.0, 1.5, 2.0

.1.2 Using HMM hidden states to predict symptoms

Table 4.5 in this paper shows part of the results that we used HMM hidden states to predict depressive symptom level. Table 1 in this section shows all the x (high symptom states) and y (number of days before participants completed the CESD scale) in our experiments. Figure 4.4 in this paper is plotted based on the result in table 1. We can see that using the condition "counting the number of symptom states in the past two weeks" to classifier users' depressive symptom level yields more reasonable precision and recall.

.2 Tables and Figures for Chapter 5

.2.1 Topic Modeling

The LDA model has three important parameters: the number of topics (n), alpha (a), Beta (b). a represents document-topic density, and a higher alpha means the documents are made up of more topics; b represents topic-word density, higher beta means the topics are made up of most of the words in the corpus. We performed a grid search $n \in N := \{5, 10, 15\}$, $a \in A := \{0.1, 0.3, 0.5, 0.7, 0.9\}$, $b \in B := \{0.1, 0.3, 0.5, 0.7, 0.9\}$ in order to optimize the coherence score. The coherence score retrieves co-occurrence counts for the given words using a sliding window. The counts are used to calculated the normalized pointwise mutual information of every top word to every other top word Mimno et al. (2011); Syed and Spruit (2017). The grid search approach shows that 15 topics yield higher coherence scores than 5 or 10.

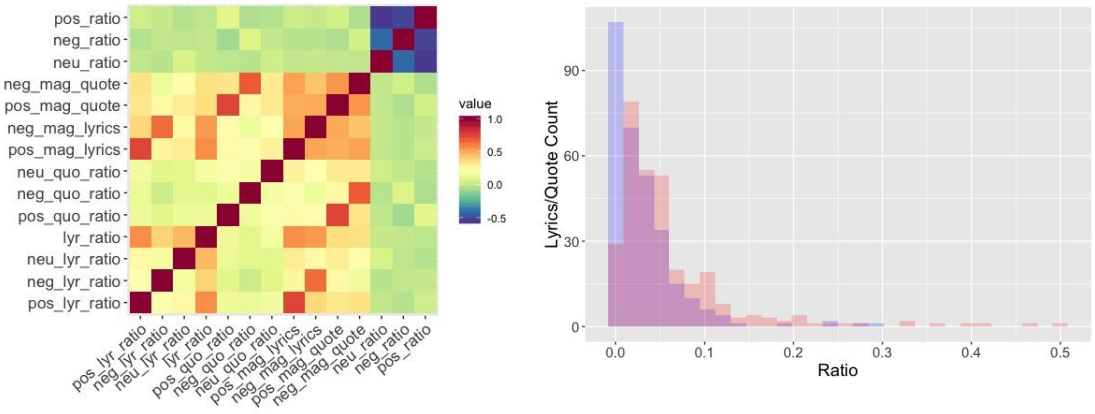
Table 1: Predicting depressive symptom with hidden states (extension of table) 4.5

window size (y)	day (x)	precision	recall
7	1	61.5	45.2
7	2	62.9	25.1
7	3	64.6	12.7
7	4	65.9	7.3
7	5	71.2	3.0
7	6	75.9	1.0
7	7	100	0.2
14	1	60.3	58.1
14	2	66.7	38.7
14	3	66.3	27.1
14	4	67.3	20.2
14	5	65.0	14.2
14	6	71.2	10.8
14	7	100	8.6

Note: Recall: recall of high symptom class, Precision: precision of high symptom class, window: size of the time window, day: days before participants completed the CESD scale. Assumption: participants assigned high symptom level if they have x high symptom states within y days before completing the CESD scale.

Table 2: Demographic Information of the 781 Participants

Ethnicity	No.	%	Marital Status	No.	%
White	511	65.3	Single	574	73.8
Asian	110	14.1	Divorced	28	3.5
Black	38	4.3	Married	27	3.4
Native American	13	1.6	Married with Children	38	4
Middle Eastern	13	1.7	Not specified	36	4.5
Not Specified	96	12.2			



A. Variable correlations

B. Quotations and Lyrics Ratio

Figure 1: Variable statistics. graph A: $p < 0.001$ for all correlations, graph B: blue:NL-quotes; red:lyrics; ratio: lyrics or quotation ratio to all post count.

.2.2 Tables and Figures

Table 3: Quotes Topics

		Lyrics		
#	theme	docs	top 10 keywords	example
0	overwhelming	245	$love_H$, got_H, $want_H$, $thing_H$, $feel_H$, $need_H$, $make_H$, $come_H$, $think_H$, say_H	example 1: You traded in your wings for everything freedom brings. You never left me.
3	self-empowerment	305	$love_L$, see_L, $know_L$, $make_L$, $feel_L$, let_L, $time_L$, go_L, got_L, $life_L$	example 1: If you're trying to turn me into someone else. Its easy to see I'm not down with that. I'm not nobody's fool.
5	self-empowerment	129	$know_H$, $hold_H$, $wait_H$, $want_H$, $tell_H$, day_H, $love_H$, $heart_H$, $dark_H$, $live_H$	example 1: So, so you think you can tell heaven from Hell blue skies from pain, can you tell a green field
7	introspection	130	$take_H$, say_H, $good_H$, $feel_H$, got_H, $time_H$, $sleep_H$, see_H, $know_H$, $change_H$	example 1: I feel angry. I feel helpless. Want to change the world. I feel violent. I feel alone. Don't try to change my mind
12	self-empowerment	264	go_H, $know_H$, let_H, $love_H$, $time_H$, day_H, $come_H$, $fall_H$, $make_H$, gon_H	example 1: we all got holes to fill, them holes are all that's real some fall on you like a storm, sometimes you dig your own,the choice is yours to make, time is yours to take
		non-lyrics quotes		
13		1432	$love_H$, $life_H$, go_H, $know_H$, day_H, $make_H$, $thing_H$, $time_H$, $feel_H$, $people_H$	example 1: Gratitude unlocks the fullness of life. It turns what we have into enough, and more. example 2: As a girl you see the world as a giant candy store filled with sweet candy and such. But one day you look around and you see a prison and you're on death row.
9		343	$life_L$, $love_L$, $thing_L$, day_L, go_L, $time_L$, $come_L$, see_L, $think_L$, $want_L$	example 1: If you can't make it good, at least make it look good. example 2: Would you dare? Would you dare to believe that you still have a reason to sing? Cause the pain that you've been feeling can't compare to the joy that's coming. So hold on.

Notes: *H* and *L* indicate whether the word comes from high or low symptoms users.

Bibliography

Henry David Abraham and Maurizio Fava. Order of onset of substance abuse and depression in a sample of depressed outpatients. *Comprehensive Psychiatry*, 40(1): 44–50, 1999.

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 30–38, 2011.

Hagop S Akiskal. The temperamental foundations of affective disorders. *Interpersonal factors in the origin and course of affective disorders*, pages 3–30, 1996.

Hagop S Akiskal and Charles Van Valkenburg. Mood disorders. In *Diagnostic interviewing*, pages 79–107. Springer, 1994.

Kareen K Akiskal and Hagop S Akiskal. The theoretical underpinnings of affective temperaments: implications for evolutionary foundations of bipolar disorder and human nature. *Journal of affective disorders*, 85(1-2):231–239, 2005.

Maryam Mohammed Aldarwish and Hafiz Farooq Ahmad. Predicting depression levels using social media posts. In *2017 IEEE 13th International Symposium on Principles of Autonomous Decentralized System (ISADS)*, pages 277–280, 2017. doi: 10.1109/ISADS.2017.41. URL `http://dx.doi.org/10.1109/ISADS.2017.41`.

Hayda Almeida, Antoine Briand, and Marie-Jean Meurs. Detecting early risk of depression from social media user-generated content. In *Conference and Labs of the Evaluation Forum (Working Notes)*, 2017.

Nazanin Andalibi, Pinar Ozturk, and Andrea Forte. Sensitive Self-disclosures, Responses, and Social Support on Instagram: The Case of #Depression. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pages 1485–1500, New York, NY, USA, 2017.

ACM. ISBN 978-1-4503-4335-0. doi: 10.1145/2998181.2998243. URL http://doi.acm.org/10.1145/2998181.2998243.

Monica Anderson, Jingjing Jiang, et al. Teens, social media & technology 2018. *Pew Research Center*, 31(2018):1673–1689, 2018.

F Angst, H. H Stassen, P. J Clayton, and J Angst. Mortality of patients with mood disorders: follow-up over 34–38 years. *Journal of Affective Disorders*, 68(2):167–181, 2002. ISSN 0165-0327. doi: 10.1016/S0165-0327(01)00377-9. URL http://www.sciencedirect.com/science/article/pii/S0165032701003779.

APA et al. *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub, 2013.

Ella Arensman, Nicole Koburger, Celine Larkin, Gillian Karwig, Claire Coffey, Margaret Maxwell, Fiona Harris, Christine Rummel-Kluge, Chantal Van Audenhove, Merike Sisask, et al. Depression awareness and self-management through the internet: protocol for an internationally standardized approach. *JMIR research protocols*, 4(3):e99, 2015.

Yoram Bachrach, Michal Kosinski, Thore Graepel, Pushmeet Kohli, and David Stillwell. Personality and patterns of facebook usage. In *Proceedings of the 4th annual ACM web science conference*, pages 24–32. ACM, 2012.

Felicity Baker and William Bor. Can music preference indicate mental health status in young people? *Australasian psychiatry*, 16(4):284–288, 2008.

Alexandra Balahur, Ralf Steinberger, Erik van der Goot, Bruno Pouliquen, and Mijail Kabadjov. Opinion mining on newspaper quotations. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology-Volume 03*, pages 523–526. IEEE Computer Society, 2009.

Lloyd Balbuena, Rudy Bowen, Marilyn Baetz, and Steven Marwaha. Mood instability and irritability as core symptoms of major depression: an exploration using rasch analysis. *Frontiers in psychiatry*, 7:174, 2016.

Jeffrey D Banfield and Adrian E Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, pages 803–821, 1993.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.148. URL https://www.aclweb.org/anthology/2020.findings-emnlp.148.

Bara Bataineh, Rehab Duwairi, and Malak Abdullah. Ardep: An arabic lexicon for detecting depression. In *Proceedings of the 2019 3rd International Conference on Advances in Artificial Intelligence*, pages 146–151, 2019.

Krishna C Bathina, Marijn ten Thij, Lorenzo Lorenzo-Luaces, Lauren A Rutter, and Johan Bollen. Depressed individuals express more distorted thinking on social media. *arXiv preprint arXiv:2002.02800*, 2020.

C Daniel Batson, Laura L Shaw, and Kathryn C Oleson. Differentiating affect, mood, and emotion: Toward functionally based conceptual distinctions. *Review of personality and social psychology*, 13:p294–326, 1992.

Natalya N Bazarova, Yoon Hyung Choi, Victoria Schwanda Sosik, Dan Cosley, and Janis Whitlock. Social sharing of emotions on facebook: Channel differences, satisfaction, and replies. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, pages 154–164. ACM, 2015.

Matthew J Beal, Zoubin Ghahramani, and Carl E Rasmussen. The infinite hidden markov model. In *Advances in neural information processing systems*, pages 577–584, 2002.

Margaret Beasley, Ted Thompson, and John Davidson. Resilience in response to life stress: the effects of coping style and cognitive hardiness. *Personality and Individual differences*, 34(1):77–95, 2003.

Aaron T Beck. Cognitive therapy: A 30-year retrospective. *American psychologist*, 46 (4):368, 1991.

Aaron T Beck. The evolution of the cognitive model of depression and its neurobiological correlates. *American Journal of Psychiatry*, 165(8):969–977, 2008.

Aaron T Beck. A 60-year evolution of cognitive theory and therapy. *Perspectives on Psychological Science*, 14(1):16–20, 2019.

Farah Benamara, Véronique Moriceau, Josiane Mothe, Faneva Ramiandrisoa, and Zhaolong He. Automatic Detection of Depressive Users in Social Media. In *Conférence francophone en Recherche d'Information et Applications (CORIA)*, 2018.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. Multi-task learning for mental health using social media text. *arXiv preprint arXiv:1712.03538*, 2017.

Helen Bergen, Keith Hawton, Keith Waters, Jennifer Ness, Jayne Cooper, Sarah Steeg, and Navneet Kapur. Premature death after self-harm: a multicentre cohort study. *The Lancet*, 380(9853):1568–1574, 2012. ISSN 0140-6736. doi: 10.1016/S0140-6736(12)61141-6. URL http://www.sciencedirect.com/science/article/pii/S0140673612611416.

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

Toine Bogers and Rasmus Nordenhoff Wernersen. How'social'are social news sites? exploring the motivations for using reddit. com. In *iConference 2014: Breaking Down Walls: Culture-Context-Computing*, pages 329–344. iSchools, 2014.

Johan Bollen, Huina Mao, and Alberto Pepe. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

John Michael Bostwick and V. Shane Pankratz. Affective disorders and suicide risk: A reexamination. *American Journal of Psychiatry*, 157(12):1925–1932, 2002. ISSN 0002-953X. doi: 10.1176/appi.ajp.157.12.1925. URL https://ajp.psychiatryonline.org/doi/full/10.1176/appi.ajp.157.12.1925.

John Bowlby. *Loss: Sadness and depression*. 3. Random House, 1998.

Thorsten Brants, Ashok C Popat, Peng Xu, Franz J Och, and Jeffrey Dean. Large language models in machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007.

David A. Brent, Joshua A. Perper, Grace Moritz, Chris Allman, Amy Friend, Claudia Roth, Joy Schweers, Lisa Balach, and Marinanne Baugher. Psychiatric risk factors

for adolescent suicide: A case-control study. *Journal of the American Academy of Child & Adolescent Psychiatry*, 32(3):521–529, 1993. ISSN 0890-8567. doi: 10.1097/00004583-199305000-00006. URL http://www.sciencedirect.com/science/article/pii/S0890856709652606.

Jonathon D Brown and Judith M Siegel. Attributions for negative life events and depression: The role of perceived control. *Journal of Personality and Social Psychology*, 54(2):316, 1988.

Lauren M Bylsma, Bethany H Morris, and Jonathan Rottenberg. A meta-analysis of emotional reactivity in major depressive disorder. *Clinical psychology review*, 28 (4):676–691, 2008.

Lauren M Bylsma, April Taylor-Clift, and Jonathan Rottenberg. Emotional reactivity to daily events in major and minor depression. *Journal of abnormal psychology*, 120 (1):155, 2011.

Fidel Cacheda, Diego Fernandez, Francisco J. Novoa, and Victor Carneiro. Early Detection of Depression: Social Network Analysis and Random Forest Techniques. *J Med Internet Res*, 21(6):e12554, 2019. doi: 10.2196/12554.

Carole Cadwalladr and Emma Graham-Harrison. Revealed: 50 million facebook profiles harvested for cambridge analytica in major data breach. *The guardian*, 17:22, 2018.

Clara Caldeira, Yu Chen, Lesley Chan, Vivian Pham, Yunan Chen, and Kai Zheng. Mobile apps for mood tracking: an analysis of features and user reviews. In *AMIA Annual Symposium Proceedings*, volume 2017, page 495. American Medical Informatics Association, 2017.

Fabio Calefato, Filippo Lanubile, and Nicole Novielli. Emotxt: a toolkit for emotion recognition from text. In *2017 seventh international conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 79–80. IEEE, 2017.

Rafael A Calvo, David N Milne, M Sazzad Hussain, and Helen Christensen. Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, 23(5):649–685, 2017.

Gustavo Carlo, Maria Vicenta Mestre, Meredith M McGinley, Paula Samper, Ana Tur, and Deanna Sandman. The interplay of emotional instability, empathy, and coping on prosocial and aggressive behaviors. *Personality and Individual Differences*, 53 (5):675–680, 2012.

Patricia A Cavazos-Rehg, Melissa J Krauss, Shaina Sowles, Sarah Connolly, Carlos Rosas, Meghana Bharadwaj, and Laura J Bierut. A content analysis of depression-related tweets. *Computers in human behavior*, 54:351–357, 2016.

Fabio Celli, Arindam Ghosh, Firoj Alam, and Giuseppe Riccardi. In the mood for sharing contents: Emotions, personality and interaction styles in the diffusion of news. *Information Processing & Management*, 52(1):93–98, 2016.

Krzysztof Chalupka, Christopher KI Williams, and Iain Murray. A framework for evaluating approximation methods for gaussian process regression. *Journal of Machine Learning Research*, 14(Feb):333–350, 2013.

Melissa KY Chan, Henna Bhatti, Nick Meader, Sarah Stockton, Jonathan Evans, Rory C O'Connor, Nav Kapur, and Tim Kendall. Predicting suicide following self-harm: systematic review of risk factors and risk scales. *The British Journal of Psychiatry*, 209(4):277–283, 2016.

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

Lucia Lushi Chen, Christopher HK Cheng, and Tao Gong. Inspecting vulnerability to depression from social media affect. *Frontiers in Psychiatry*, 11:54, 2020a.

Lucia Lushi Chen, Walid Magdy, Heather Whalley, and Maria Wolters. It's not just about sad songs: The effect of depression on posting lyrics and quotes. In *International Conference on Social Informatics*, pages 58–66. Springer, 2020b.

Lucia Lushi Chen, Walid Magdy, and Maria K Wolters. The effect of user psychology on the content of social media posts: Originality and transitions matter. *Frontiers in Psychology*, 11:526, 2020c.

Lushi Chen, Tao Gong, Michal Kosinski, David Stillwell, and Robert L Davidson. Building a profile of subjective well-being for social media users. *PloS one*, 12(11): e0187278, 2017.

Lushi Chen, Abeer Aldayel, Nikolay Bogoychev, and Tao Gong. Similar minds post alike: Assessment of suicide risk using a hybrid model. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 152–157, 2019.

Lushi Chen, Walid Magdy, Heather Whalley, and Maria Klara Wolters. Examining the role of mood patterns in predicting self-reported depressive symptoms. In *12th ACM Conference on Web Science*, pages 164–173, 2020d.

Xuetong Chen, Martin D Sykora, Thomas W Jackson, and Suzanne Elayan. What about mood swings: Identifying depression on twitter with temporal measures of emotions. In *Companion Proceedings of the The Web Conference 2018*, pages 1653–1660, 2018.

Paula Glenda Ferrer Cheng, Roann Munoz Ramos, Jó Ágila Bitsch, Stephan Michael Jonas, Tim Ix, Portia Lynn Quetulio See, and Klaus Wehrle. Psychologist in a pocket: lexicon development and content validation of a mobile-based app for depression screening. *JMIR mHealth and uHealth*, 4(3):e88, 2016.

Chun Yueh Chiu, Hsien Yuan Lane, Jia Ling Koh, and Arbee LP Chen. Multimodal depression detection on instagram considering time interval of posts. *Journal of Intelligent Information Systems*, pages 1–23, 2020.

David A Clark, Aaron T Beck, Brad A Alford, Peter J Bieling, and Zindel V Segal. Scientific foundations of cognitive theory and therapy of depression, 2000.

CLpsy. Suicide risk prediction sharedtask, 2019. URL `https://clpsych.org/shared-task-2019-2/`.

Gualtiero B Colombo, Pete Burnap, Andrei Hodorog, and Jonathan Scourfield. Analysing the connectivity and communication of suicidal users on twitter. *Computer communications*, 73:291–300, 2016.

Glen Coppersmith, Mark Dredze, and Craig Harman. Quantifying mental health signals in twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 51–60, 2014.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. Clpsych 2015 shared task: Depression and ptsd on twitter. In *Proceedings*

of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pages 31–39, 2015.

Paul T Costa and Robert R McCrae. *Neo Pi-R*. Psychological Assessment Resources Odessa, FL, 1992.

Alan S Cowen and Dacher Keltner. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, 114(38):E7900–E7909, 2017.

Kelly C Cukrowicz, Jennifer S Cheavens, Kimberly A Van Orden, R Michael Ragain, and Ronald L Cook. Perceived burdensomeness and suicide ideation in older adults. *Psychology and aging*, 26(2):331, 2011.

Richard J Davidson. Affective style and affective disorders: Perspectives from affective neuroscience. *Cognition & Emotion*, 12(3):307–330, 1998.

Munmun De Choudhury and Sushovan De. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Eighth International AAAI Conference on Weblogs and Social Media*, pages 21–30, 2014.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. In *Seventh international AAAI conference on weblogs and social media*, pages 170–185, 2013.

Munmun De Choudhury, Scott Counts, Eric J Horvitz, and Aaron Hoff. Characterizing and predicting postpartum depression from shared facebook data. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 626–638, 2014.

Irismar Reis De Oliveira, Camila Seixas, Flávia L Osório, José Alexandre S Crippa, José Neander De Abreu, Igor Gomes Menezes, Aileen Pidgeon, Donna Sudak, and Amy Wenzel. Evaluation of the psychometric properties of the cognitive distortions questionnaire (cd-quest) in a sample of undergraduate students. *Innovations in clinical neuroscience*, 12(7-8):20, 2015.

M. Deshpande and V. Rao. Depression detection using emotion artificial intelligence. pages 858–862, 2017. doi: 10.1109/ISS1.2017.8389299.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

Barbara Di Eugenio, Nick Green, and Rajen Subba. Detecting life events in feeds from twitter. In *2013 IEEE Seventh International Conference on Semantic Computing*, pages 274–277. Ieee, 2013.

E. D. Diener, Robert A. Emmons, Randy J. Larsen, and Sharon Griffin. The satisfaction with life scale. *Journal of personality assessment*, 49(1):71–75, 1985.

Ed Diener and Robert A Emmons. The independence of positive and negative affect. *Journal of personality and social psychology*, 47(5):1105, 1984.

Ed Diener, Eunkook M Suh, Richard E Lucas, and Heidi L Smith. Subjective well-being: Three decades of progress. *Psychological bulletin*, 125(2):276, 1999.

John M Digman. Personality structure: Emergence of the five-factor model. *Annual review of psychology*, 41(1):417–440, 1990.

Seth G Disner, Christopher G Beevers, Emily AP Haigh, and Aaron T Beck. Neural mechanisms of the cognitive model of depression. *Nature Reviews Neuroscience*, 12(8):467–477, 2011.

Bryan Dosono, Yasmeen Rashidi, Taslima Akter, Bryan Semaan, and Apu Kapadia. Challenges in Transitioning from Civil to Military Culture: Hyper-Selective Disclosure Through ICTs. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW):41:1–41:23, December 2017. ISSN 2573-0142. doi: 10.1145/3134676. URL http://doi.acm.org/10.1145/3134676.

Afsoon Eftekhari, Lori A Zoellner, and Shree A Vigil. Patterns of emotion regulation and psychopathology. *Anxiety, Stress, & Coping*, 22(5):571–586, 2009.

Samuel E. Ehrenreich and Marion K. Underwood. Adolescents' internalizing symptoms as predictors of the content of their Facebook communication and responses

received from peers. *Translational Issues in Psychological Science*, 2(3):227–237, 2016. doi: 10.1037/tps0000077. Place: US Publisher: Educational Publishing Foundation.

Johannes C Eichstaedt, Robert J Smith, Raina M Merchant, Lyle H Ungar, Patrick Crutchley, Daniel Preoţiuc-Pietro, David A Asch, and H Andrew Schwartz. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44):11203–11208, 2018.

Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. Diffusion of lexical change in social media. *PloS one*, 9(11):e113114, 2014.

Paul Ekman. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16, 1999.

Sindhu Kiranmai Ernala, Michael L Birnbaum, Kristin A Candan, Asra F Rizvi, William A Sterling, John M Kane, and Munmun De Choudhury. Methodological gaps in predicting mental health states from social media: triangulating diagnostic signals. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2019.

Golnoosh Farnadi, Susana Zoghbi, Marie-Francine Moens, and Martine De Cock. Recognising personality traits using facebook status updates. In *Seventh International AAAI Conference on Weblogs and Social Media*, pages 154–164, 2013.

Ethan Fast, Binbin Chen, and Michael S Bernstein. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4647–4657, 2016.

Iram Fatima, Hamid Mukhtar, Hafiz Farooq Ahmad, and Kashif Rajpoot. Analysis of user-generated content from online social communities to characterise and predict depression degree. *Journal of Information Science*, 44(5):683–695, 2018.

Iram Fatima, Burhan Ud Din Abbasi, Sharifullah Khan, Majed Al-Saeed, Hafiz Farooq Ahmad, and Rafia Mumtaz. Prediction of postpartum depression using machine learning techniques from social media text. *Expert Systems*, 36(4):e12409, 2019. doi: 10.1111/exsy.12409.

Johannes Feldhege, Markus Moessner, and Stephanie Bauer. Who says what? content and participation characteristics in an online depression community. *Journal of Affective Disorders*, 263:521–527, 2020.

Susan T Fiske and Robert M Hauser. Protecting human research participants in the age of big data, 2014.

Elizabeth Ford, Keegan Curlewis, Akkapon Wongkoblap, and Vasa Curcin. Public opinions on using social media content to identify users with depression and target mental health care advertising: Mixed methods survey. *JMIR mental health*, 6(11): e12942, 2019.

Barbara L Fredrickson. What good are positive emotions? *Review of general psychology*, 2(3):300–319, 1998.

Jan-Philipp Freudenstein, Christoph Strauch, Patrick Mussel, and Matthias Ziegler. Four personality types may be neither robust nor exhaustive. *Nature human behaviour*, 3(10):1045–1046, 2019.

Nico H Frijda. Moods, emotion episodes, and emotions. *Handbook of emotions*, 12 (2):155, 1993.

Frank Fujita, Ed Diener, and Ed Sandvik. Gender differences in negative affect and well-being: the case for emotional intensity. *Journal of personality and social psychology*, 61(3):427, 1991.

Matthias Gamer, Jim Lemon, Ian Fellows, and Puspendra Singh. irr: Various coefficients of interrater reliability and agreement. r package version 0.84. *Internet resource: http://CRAN. R-project. org/package= irr](Verified April 10, 2013)*, 2012.

Zhong-Ke Gao, Michael Small, and Juergen Kurths. Complex network analysis of time series. *EPL (Europhysics Letters)*, 116(5):50001, 2017.

Steven T Garren. Permutation tests for nonparametric statistics using r. *Asian Research Journal of Mathematics*, pages 1–8, 2017.

Samuel T Gladding, Deborah Newsome, Erin Binkley, and Donna A Henderson. The lyrics of hurting and healing: Finding words that are revealing. *Journal of Creativity in Mental Health*, 3(3):212–219, 2008.

Coppersmith Glen, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. Clpsych 2015 shared task: Depression and ptsd on twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 361–364, 2015.

Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. Predicting personality from twitter. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, pages 149–156. IEEE, 2011.

L R Goldberg, J A Johnson, H W Eber, R Hogan, M C Ashton, C R Cloninger, and H G Gough. The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, 40:84–96, 2006.

Su Golder, Arabella Scantlebury, and Helen Christmas. Understanding public attitudes toward researchers using social media for detecting and monitoring adverse events data: multi methods study. *Journal of medical Internet research*, 21(8):e7081, 2019.

Susan Gore, Robert H Aseltine Jr, and Mary Ellen Colten. Gender, social-relationship involvement, and depression. *Journal of research on adolescence*, 3(2):101–125, 1993.

James J Gross. Emotion regulation: Past, present, future. *Cognition & emotion*, 13(5): 551–573, 1999.

James J Gross, Steven K Sutton, and Timothy Ketelaar. Relations between affect and personality: Support for the affect-level and affective-reactivity views. *Personality and social psychology bulletin*, 24(3):279–288, 1998.

Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49, 2017.

Sharath Chandra Guntuku, Daniel Preotiuc-Pietro, Johannes C Eichstaedt, and Lyle H Ungar. What twitter profile and posted images reveal about depression and anxiety. In *Proceedings of the international AAAI conference on web and social media*, volume 13, pages 236–246, 2019.

Lei Guo, Enhua Tan, Songqing Chen, Xiaodong Zhang, and Yihong Zhao. Analyzing patterns of user content generation in online social networks. In *Proceedings of*

the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 369–378, 2009.

Brian W Haas, Kazufumi Omura, R Todd Constable, and Turhan Canli. Is automatic emotion regulation associated with agreeableness? a perspective using a social neuroscience approach. *Psychological Science*, 18(2):130–132, 2007.

Jill B Hamilton, Margarete Sandelowski, Angelo D Moore, Mansi Agarwal, and Harold G Koenig. "you need a song to bring you through": The use of religious songs to manage stressful life events. *The Gerontologist*, 53(1):26–38, 2013.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. On the state of social media data for mental health research. In *Proceedings of the 7th Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, 2021.

Allison G Harvey. *Cognitive behavioural processes across psychological disorders: A transdiagnostic approach to research and treatment*. Oxford University Press, USA, 2004.

Keith Hawton, Kate EA Saunders, and Rory C O'Connor. Self-harm and suicide in adolescents. *The Lancet*, 379(9834):2373–2382, 2012. ISSN 0140-6736. doi: 10.1016/S0140-6736(12)60322-5. URL `http://www.sciencedirect.com/science/article/pii/S0140673612603225`.

Bruce Headey, Jonathan Kelley, and Alex Wearing. Dimensions of mental health: Life satisfaction, positive affect, anxiety and depression. *Social indicators research*, 29 (1):63–82, 1993.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H Clark, and Philipp Koehn. Scalable modified kneser-ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 690–696, 2013.

Lesley Hepburn and Michael W Eysenck. Personality, average mood and mood variability. *Personality and Individual Differences*, 10(9):975–983, 1989.

Alejandro González Hevia, Rebeca Cerezo Menéndez, and Daniel Gayo-Avello. Analyzing the use of existing systems for the clpsych 2019 shared task. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 148–151, 2019.

J.J. Higgins. *Introduction to Modern Non-Parametric Statistics*. Brooks/Cole Pacific Grove, 2003.

Craig A Hill. Seeking emotional support: The influence of affiliative need and partner warmth. *Journal of Personality and Social Psychology*, 60(1):112, 1991.

Steven D Hollon and Aaron T Beck. Cognitive therapy of depression. *Cognitive-behavioral interventions: Theory, research, and procedures*, pages 153–203, 1979.

Courtenay Honey and Susan C Herring. Beyond microblogging: Conversation and collaboration via twitter. In *2009 42nd Hawaii International Conference on System Sciences*, pages 1–10. Ieee, 2009.

Matthew Honnibal and Mark Johnson. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal, September 2015. Association for Computational Linguistics. URL `https://aclweb.org/anthology/D/D15/D15-1162`.

Seyed Habib Hosseini-Saravani, Sara Besharati, Hiram Calvo, and Alexander Gelbukh. Depression Detection in Social Media Using a Psychoanalytical Technique for Feature Extraction and a Cognitive Based Classifier. In *Mexican International Conference on Artificial Intelligence*, pages 282–292. Springer, 2020.

Marlies Houben, Wim Van Den Noortgate, and Peter Kuppens. The relation between short-term emotion dynamics and psychological well-being: A meta-analysis. *Psychological bulletin*, 141(4):901, 2015.

Hsiao-Wei Hu, Kai-Shyang Hsu, Connie Lee, Hung-Lin Hu, Cheng-Yen Hsu, Wen-Han Yang, Ling-Yun Wang, and Ting-An Chen. Keyword-Driven Depressive Tendency Model for Social Media Posts. In *Business Information Systems, Bis 2019, Pt Ii*, volume 354, pages 14–22. Springer, 2019. Journal Abbreviation: Business Information Systems, Bis 2019, Pt Ii.

Quan Hu, Ang Li, Fei Heng, Jianpeng Li, and Tingshao Zhu. Predicting Depression of Social Media User on Different Observation Windows. In *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, 2015.

Yu-Ching Huang, Chieh-Feng Chiang, and Arbee LP Chen. Predicting Depression Tendency based on Image, Text and Behavior Data from Instagram. In *Proceedings of the 8th International Conference on Data Science, Technology and Applications (DATA 2019)*, pages 32–40, 2019.

huggingface. huggingface models, 2021. URL `https://huggingface.co/models`.

Patrick G Hunter, E Glenn Schellenberg, and Andrew T Griffith. Misery loves company: Mood-congruent emotional responding to music. *Emotion*, 11(5):1068, 2011.

Mohammad Shahid Husain. Social Media Analytics to Predict Depression Level in the Users. In *Early Detection of Neurological Disorders Using Machine Learning Systems*, pages 199–215. IGI Global, 2019. Journal Abbreviation: Early Detection of Neurological Disorders Using Machine Learning Systems.

Jamil Hussain, Fahad Ahmed Satti, Muhammad Afzal, Wajahat Ali Khan, Hafiz Syed Muhammad Bilal, Muhammad Zaki Ansaar, Hafiz Farooq Ahmad, Taeho Hur, Jaehun Bang, Jee-In Kim, Gwang Hoon Park, Hyonwoo Seung, and Sungyoung Lee. Exploring the dominant features of social media for depression detection. *Journal of Information Science*, 46(6):739–759, 2020. doi: 10.1177/0165551519860469.

Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. Deep learning for depression detection of twitter users. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 88–97. Association for Computational Linguistics, 2018. URL `http://www.aclweb.org/anthology/W18-0609`.

Anne Immonen, Pekka Pääkkönen, and Eila Ovaska. Evaluating the quality of social media data in big data architecture. *Ieee Access*, 3:2028–2043, 2015.

Md Rafiqul Islam, Muhammad Ashad Kabir, Ashir Ahmed, Abu Raihan M Kamal, Hua Wang, and Anwaar Ulhaq. Depression detection from social network data using machine learning techniques. *Health information science and systems*, 6(1):8, 2018.

Zunaira Jamil. *Monitoring tweets for depression to detect at-risk users*. PhD thesis, Université d'Ottawa/University of Ottawa, 2017.

Shaoxiong Ji, Celina Ping Yu, Sai-fu Fung, Shirui Pan, and Guodong Long. Supervised learning for suicidal ideation detection in online user content. *Complexity*, 2018, 2018.

Mathias Johansson and Tomas Olofsson. Bayesian model selection for markov, hidden markov, and multinomial models. *IEEE signal processing letters*, 14(2):129–132, 2007.

Grace J Johnson and Paul J Ambrose. Neo-tribes: The power and potential of online communities in health care. *Communications of the ACM*, 49(1):107–113, 2006.

Thomas Joiner. *Why people die by suicide*. Harvard University Press, 2007.

Patrik N Juslin and Petri Laukka. Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. *Journal of new music research*, 33(3):217–238, 2004.

David N Juurlink, Nathan Herrmann, John P Szalai, Alexander Kopp, and Donald A Redelmeier. Medical illness and the risk of suicide in the elderly. *Archives of internal medicine*, 164(11):1179–1184, 2004.

Atreyi Kankanhalli, Bernard CY Tan, and Kwok-Kee Wei. Contributing knowledge to electronic knowledge repositories: An empirical investigation. *MIS quarterly*, pages 113–143, 2005.

Simona C Kaplan, Amanda S Morrison, Philippe R Goldin, Thomas M Olino, Richard G Heimberg, and James J Gross. The cognitive distortions questionnaire (cd-quest): Validation in a sample of adults with social anxiety disorder. *Cognitive therapy and research*, 41(4):576–587, 2017.

Christian Karmen, Robert C. Hsiung, and Thomas Wetter. Screening Internet forum participants for depression symptoms by assembling and enhancing multiple NLP methods. *Computer Methods and Programs in Biomedicine*, 120(1):27–36, 2015. doi: 10.1016/j.cmpb.2015.03.008.

Ronald C Kessler, Guilherme Borges, and Ellen E Walters. Prevalence of and risk factors for lifetime suicide attempts in the national comorbidity survey. *Archives of general psychiatry*, 56(7):617–626, 1999.

Terrence A Ketter et al. Diagnostic features, prevalence, and impact of bipolar disorder. *J Clin Psychiatry*, 71(6):e14, 2010.

Jan H Kietzmann, Kristopher Hermkens, Ian P McCarthy, and Bruno S Silvestre. Social media? get serious! understanding the functional building blocks of social media. *Business horizons*, 54(3):241–251, 2011.

Junghyun Kim and Jong-Eun Roselyn Lee. The facebook paths to happiness: Effects of the number of facebook friends and self-presentation on subjective well-being. *CyberPsychology, behavior, and social networking*, 14(6):359–364, 2011.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180, 2007.

Kosinski. Mypersonality, 2021. URL `https://sites.google.com/michalkosinski.com/mypersonality`.

Maria Kovacs and Aaron T Beck. Maladaptive cognitive structures in depression. *Essential papers on depression*, pages 240–258, 1986.

Maria Kovacs and Betsy Garrison. Hopelessness and eventual suicide: a 10-year prospective study of patients hospitalized with suicidal ideation. *American journal of Psychiatry*, 1(42):559–563, 1985.

Ozan Kuru, Joseph Bayer, Josh Pasek, and Scott W Campbell. Understanding and measuring mobile facebook use: Who, why, and how? *Mobile Media & Communication*, 5(1):102–120, 2017.

E. Megan Lachmar, Andrea K. Wittenborn, Katherine W. Bogen, and Heather L. McCauley. #MyDepressionLooksLike: Examining Public Discourse About Depression on Twitter. *JMIR Ment Health*, 4(4):e43, 2017. doi: 10.2196/mental.8141.

Randy J Larsen. The stability of mood variability: a spectral analytic approach to daily mood assessments. *Journal of personality and social psychology*, 52(6):1195, 1987.

Mark F Lefebvre. Cognitive distortion and cognitive errors in depressed psychiatric and low back pain patients. *Journal of consulting and clinical psychology*, 49(4): 517, 1981.

Angela Leis, Francesco Ronzano, Miguel A. Mayer, Laura I. Furlong, and Ferran Sanz. Detecting Signs of Depression in Tweets in Spanish: Behavioral and Linguistic Analysis. *Journal of Medical Internet Research*, 21(6):e14199, 2019. doi: 10.2196/14199.

Janice C Levy and Eva Y Deykin. Suicidality, depression, and substance abuse in adolescence. *The American journal of psychiatry*, 1989.

Yong Li, Mengsi Cai, Shuo Qin, and Xin Lu. Depressive Emotion Detection and Behavior Analysis of Men Who Have Sex With MenviaSocial Media. *Frontiers in Psychiatry*, 11:830, 2020. doi: 10.3389/fpsyt.2020.00830.

Grace Y Lim, Wilson W Tam, Yanxia Lu, Cyrus S Ho, Melvyn W Zhang, and Roger C Ho. Prevalence of depression in the community from 30 countries between 1994 and 2014. *Scientific reports*, 8(1):1–10, 2018.

Huijie Lin, Jia Jia, Quan Guo, Yuanyuan Xue, Qi Li, Jie Huang, Lianhong Cai, and Ling Feng. User-level psychological stress detection from social media using deep neural network. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 507–516, 2014.

Junjie Lin, Wenji Mao, and Daniel D Zeng. Personality-based refinement for sentiment classification in microblog. *Knowledge-Based Systems*, 132:204–214, 2017.

Geoffrey G Lloyd and William A Lishman. Effect of depression on the speed of recall of pleasant and unpleasant experiences. *Psychological medicine*, 5(2):173–180, 1975.

D Lumb. Scientists release personal data for 70,000 okcupid profiles. *Available at engt. co/2b4NnQ0. Accessed August*, 7:2016, 2016.

Olivier Luminet IV, Patrick Bouts, Frédérique Delie, Antony SR Manstead, and Bernard Rimé. Social sharing of emotion following exposure to a negatively valenced situation. *Cognition & Emotion*, 14(5):661–688, 2000.

Xiao Ma, Jeff Hancock, and Mor Naaman. Anonymity, intimacy and self-disclosure in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 3857–3869, 2016.

Michael J Mahoney. *Cognition and psychotherapy*. Springer Science & Business Media, 2013.

Daniel Maier, Annie Waldherr, Peter Miltner, Gregor Wiedemann, Andreas Niekler, Alexa Keinert, Barbara Pfetsch, Gerhard Heyer, Ueli Reber, Thomas Häussler, et al. Applying lda topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, 12(2-3):93–118, 2018.

Paulo Mann, Aline Paes, and Elton H Matsushima. See and Read: Detecting Depression Symptoms in Higher Education Students Using Multimodal Social Media Data. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 440–451, 2020.

Luciana Marinelarena-Dondena, Edgardo Ferretti, Manolis Maragoudakis, Maximiliano Sapino, and Marcelo Luis Errecalde. Predicting Depression: a comparative study of machine learning approaches based on language usage. *Cuadernos De Neuropsicologia-Panamerican Journal of Neuropsychology*, 11(3):42–54, 2017. doi: 10.7714/CNPS/11.3.201.

Maryanne Martin. Neuroticism as predisposition toward depression: A cognitive mechanism. *Personality and Individual Differences*, 6(3):353–365, 1985.

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

Andrew McCallum, Kamal Nigam, et al. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Citeseer, 1998.

Robert R. McCrae and Juri Allik, editors. *The five-factor model of personality across cultures*. Springer Science and Business Media, 2002.

Robert R. McCrae and Oliver P. John. An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215, 1992.

James McDonald, Kate Lockwood Harris, and Jessica Ramirez. Revealing and Concealing Difference: A Critical Approach to Disclosure and an Intersectional Theory of "Closeting". *Communication Theory*, 2019. doi: 10.1093/ct/qtz017. URL `https://academic.oup.com/ct/advance-article/doi/10.1093/ct/qtz017/5625990`.

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

Soraya Mehdizadeh. Self-presentation 2.0: Narcissism and self-esteem on facebook. *Cyberpsychology, behavior, and social networking*, 13(4):357–364, 2010.

Brian P Meier, Michael D Robinson, and Benjamin M Wilkowski. Turning the other cheek: Agreeableness and the regulation of aggression-related primes. *Psychological Science*, 17(2):136–142, 2006.

Raina M. Merchant, David A. Asch, Patrick Crutchley, Lyle H. Ungar, Sharath C. Guntuku, Johannes C. Eichstaedt, Shawndra Hill, Kevin Padrez, Robert J. Smith, and H. Andrew Schwartz. Evaluating the predictability of medical conditions from social media posts. *PLoS One*, 14(6):e0215476, 2019. doi: 10.1371/journal.pone.0215476.

Ivan Mervielde, Veerle Buyst, and Filip De Fruyt. The validity of the big-five as a model for teachers' ratings of individual differences among children aged 4–12 years. *Personality and Individual Differences*, 18(4):525–534, 1995.

Minas Michikyan. Depression symptoms and negative online disclosure among young adults in college: a mixed-methods approach. *J Ment Health*, 29(4):392–400, 2020. doi: 10.1080/09638237.2019.1581357.

Jude Mikal, Samantha Hurst, and Mike Conway. Ethical issues in using twitter for population-level depression monitoring: a qualitative study. *BMC medical ethics*, 17(1):22, 2016.

David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the conference on empirical methods in natural language processing*, pages 262–272. Association for Computational Linguistics, 2011.

Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and James Rosenquist. Understanding the demographics of twitter users. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5 of *1*, 2011.

Saif Mohammad. A practical guide to sentiment annotation: Challenges and solutions. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 174–179, 2016.

Karo Moilanen and Stephen Pulman. Sentiment composition. In *Proceedings of the Recent Advances in Natural Language Processing International Conference*, pages 378–382, 2007.

Scott M Monroe and Kate L Harkness. Life stress, the" kindling" hypothesis, and the recurrence of depression: considerations from a life stress perspective. *Psychological review*, 112(2):417, 2005.

Scott M Monroe, George M Slavich, Katholiki Georgiades, et al. The social environment and life stress in depression. *Handbook of depression*, 2(1):340–60, 2009.

Michelle Morales, Prajjalita Dey, Thomas Theisen, Daniel Belitz, and Natalia Chernova. An investigation of deep learning systems for suicide risk assessment. In *Proceedings of the sixth workshop on computational linguistics and clinical psychology*, pages 177–181, 2019.

William N Morris. *Mood: The frame of mind*. Springer Science & Business Media, 2012.

Amanda S Morrison, Carrie M Potter, Matthew M Carper, Dina G Kinner, Dane Jensen, Laura Bruce, Judy Wong, Irismar Reis de Oliveira, Donna M Sudak, and Richard G Heimberg. The cognitive distortions questionnaire (cd-quest): Psychometric properties and exploratory factor analysis. *International Journal of Cognitive Therapy*, 8(4):287–305, 2015.

Miranda Mourby, Elaine Mackey, Mark Elliot, Heather Gowans, Susan E Wallace, Jessica Bell, Hannah Smith, Stergios Aidinlis, and Jane Kaye. Are 'pseudonymised'data always personal data? implications of the gdpr for administrative data research in the uk. *Computer Law & Security Review*, 34(2):222–233, 2018.

Raza Ul Mustafa, Noman Ashraf, Fahad Shabbir Ahmed, Javed Ferzund, Basit Shahzad, and Alexander Gelbukh. A Multiclass Depression Detection in Social Media Based on Sentiment Analysis. In *7th International Conference on Information Technology–New Generations (ITNG 2020)*, pages 659–662. Springer, 2020.

Inez Myin-Germeys, Zuzana Kasanova, Thomas Vaessen, Hugo Vachon, Olivia Kirtley, Wolfgang Viechtbauer, and Ulrich Reininghaus. Experience sampling methodology in mental health research: new insights and technical developments. *World*

Psychiatry, 17(2):123–132, 2018. doi: 10.1002/wps.20513. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/wps.20513.

Mor Naaman, Jeffrey Boase, and Chih-Hui Lai. Is it really about me? message content in social awareness streams. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 189–192, 2010.

Moin Nadeem. Identifying depression on twitter. *arXiv preprint arXiv:1607.07384*, 2016.

Ashwini Nadkarni and Stefan G Hofmann. Why do people use facebook? *Personality and individual differences*, 52(3):243–249, 2012.

Priya Nambisan, Zhihui Luo, Akshat Kapoor, Timothy B Patrick, and Ron A Cisler. Social media, big data, and public health informatics: Ruminating behavior of depression revealed through twitter. In *2015 48th Hawaii International Conference on System Sciences*, pages 2906–2913. IEEE, 2015.

Sonya Negriff. Depressive Symptoms Predict Characteristics of Online Social Networks. *J Adolesc Health*, 65(1):101–106, 2019. doi: 10.1016/j.jadohealth.2019.01.026.

Cory F Newman. *Cognitive Restructuring/Cognitive Therapy*. OxfordPress, christine maguth nezu and arthur m. nezu edition, 2015. doi: 10.1093/oxfordhb/9780199733255.013.22.

David Newman, Chaitanya Chemudugunta, and Padhraic Smyth. Statistical entity-topic models. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 680–686, 2006.

Thin Nguyen, Dinh Phung, Bo Dao, Svetha Venkatesh, and Michael Berk. Affective and content analysis of online depression communities. *IEEE Transactions on Affective Computing*, 5(3):217–226, 2014.

NHS. Mental health assessments. https://www.nhs.uk/using-the-nhs/nhs-services/mental-health-services/mental-health-assessments/, 2020. Accessed: 2020-12-14.

William H Norman, Ivan W Miller, and Steven H Klee. Assessment of cognitive distortion in a clinically depressed population. *Cognitive Therapy and Research*, 7 (2):133–140, 1983.

Bridianne O'Dea, Stephen Wan, Philip J. Batterham, Alison L. Calear, Cecile Paris, and Helen Christensen. Detecting suicidality on twitter. *Internet Interventions*, 2(2): 183–188, 2015. ISSN 2214-7829. doi: 10.1016/j.invent.2015.03.005. URL `http://www.sciencedirect.com/science/article/pii/S2214782915000160`.

Bridianne O'Dea, Tjeerd W. Boonstra, Mark E. Larsen, Thin Nguyen, Svetha Venkatesh, and Helen Christensen. The Relationship Between Linguistic Expression And Symptoms Of Depression, Anxiety, And Suicidal Thoughts: A Longitudinal Study Of Blog Content. *Zenodo*, 2018. doi: 10.5281/ZENODO.1476492.

Irismar Reis De Oliveira. *Trial-based cognitive therapy: a manual for clinicians*. Routledge, 2014.

Yaakov Ophir, Christa SC Asterhan, and Baruch B Schwarz. Unfolding the notes from the walls: Adolescents' depression manifestations on facebook. *Computers in Human Behavior*, 72:96–107, 2017.

Yaakov Ophir, Christa S. C. Asterhan, and Baruch B. Schwarz. The digital footprints of adolescent depression, social rejection and victimization of bullying on Facebook. *Computers in Human Behavior*, 91:62–71, 2019. doi: 10.1016/j.chb.2018.09.025.

John G. Orme, Janet Reis, and Elicia J. Herz. Factorial and discriminant validity of the center for epidemiological studies depression (ces-d) scale. *Journal of clinical psychology*, 42(1):28–33, 1986.

Laura Orsolini, Roberto Latini, Maurizio Pompili, Gianluca Serafini, Umberto Volpe, Federica Vellante, Michele Fornaro, Alessandro Valchera, Carmine Tomasetti, Silvia Fraticelli, et al. Understanding the complex of suicide in depression: from research to clinics. *Psychiatry investigation*, 17(3):207, 2020.

Bo Pang, Lillian Lee, et al. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135, 2008.

Gregory Park, H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Michal Kosinski, David J Stillwell, Lyle H Ungar, and Martin EP Seligman. Automatic personality assessment through social media language. *Journal of personality and social psychology*, 108(6):934, 2015.

Minsu Park, Chiyoung Cha, and Meeyoung Cha. Depressive moods of users portrayed in twitter. In *Proceedings of the ACM SIGKDD Workshop on healthcare informatics (HI-KDD)*, volume 2012, pages 1–8. ACM New York, NY, 2012.

William Pavot and Ed Diener. Review of the satisfaction with life scale. In *Assessing well-being*, pages 101–117. Springer, 2009.

Zhichao Peng, Qinghua Hu, and Jianwu Dang. Multi-kernel SVM based depression recognition using social media data. *International Journal of Machine Learning and Cybernetics*, 10(1):43–57, 2019.

James W Pennebaker and Laura A King. Linguistic styles: Language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296, 1999.

James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001, 2001.

Leo L Pipino, Yang W Lee, and Richard Y Wang. Data quality assessment. *Communications of the ACM*, 45(4):211–218, 2002.

Nooshin Pishva, Maryam Ghalehban, Afsane Moradi, and Leila Hoseini. Personality and happiness. *Procedia-Social and Behavioral Sciences*, 30:429–432, 2011.

Sara Poletti, Cristina Colombo, and Francesco Benedetti. Adverse childhood experiences worsen cognitive distortion during adult bipolar depression. *Comprehensive psychiatry*, 55(8):1803–1808, 2014.

John Powell, John Geddes, Jonathan Deeks, Michael Goldacre, and Keith Hawton. Suicide in psychiatric hospital in-patients: risk factors and their predictive power. *The British Journal of Psychiatry*, 176(3):266–272, 2000.

Nurul F Praharso, Morgan J Tear, and Tegan Cruwys. Stressful life transitions and wellbeing: A comparison of the stress buffering hypothesis and the social identity model of identity change. *Psychiatry research*, 247:265–275, 2017.

Liju Qian, Li Liu, Min Chen, Shanmei Wang, Zhongchang Cao, and Ning Zhang. Reliability and validity of the chinese version of the cognitive distortions questionnaire (cd-quest) in college students. *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research*, 26:e926786–1, 2020.

Joaquin Quiñonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate gaussian process regression. *Journal of Machine Learning Research*, 6 (Dec):1939–1959, 2005.

Lenore Sawyer Radloff. The ces-d scale: A self-report depression scale for research in the general population. *Applied psychological measurement*, 1(3):385–401, 1977.

Faneva Ramiandrisoa and Josiane Mothe. Early Detection of Depression and Anorexia from Social Media: A Machine Learning Approach. In *Circle 2020*, volume 2621, 2020.

Andrew G Reece and Christopher M Danforth. Instagram photos reveal predictive markers of depression. *EPJ Data Science*, 6(1):1–12, 2017.

Andrew G. Reece, Andrew J. Reagan, Katharina L. M. Lix, Peter Sheridan Dodds, Christopher M. Danforth, and Ellen J. Langer. Forecasting the onset and course of mental illness with Twitter data. *Sci Rep*, 7(1):13006, 2017. doi: 10.1038/s41598-017-12961-9.

Philip Resnik, William Armstrong, Leonardo Claudino, Thang Nguyen, Viet-An Nguyen, and Jordan Boyd-Graber. Beyond lda: exploring supervised topic modeling for depression-related language in twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 99–107, 2015.

Seyed Mahdi Rezaeinia, Rouhollah Rahmani, Ali Ghodsi, and Hadi Veisi. Sentiment analysis based on improved pre-trained word embeddings. *Expert Systems with Applications*, 117:139–147, 2019.

Filipe N Ribeiro, Fabrício Benevenuto, and Emilio Zagheni. How biased is the population of facebook users? comparing the demographics of facebook users with census data to generate correction factors. In *12th ACM Conference on Web Science*, pages 325–334, 2020.

Benjamin J Ricard, Lisa A Marsch, Benjamin Crosier, and Saeed Hassanpour. Exploring the utility of community-generated social media content for detecting depression: an analytical study on Instagram. *Journal of medical Internet research*, 20 (12):e11817, 2018.

Robert E. Roberts. Reliability of the ces-d scale in different ethnic contexts. *Psychiatry research*, 2(2):125–134, 1980.

Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif Mohammad, Alan Ritter, and Veselin Stoyanov. Semeval-2015 task 10: Sentiment analysis in twitter. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 451–463, 2015.

Mary K Rothbart, Stephan A Ahadi, and David E Evans. Temperament and personality: origins and outcomes. *Journal of personality and social psychology*, 78(1):122, 2000.

Jonathan Rottenberg. Mood and emotion in major depression. *Current Directions in Psychological Science*, 14(3):167–170, 2005.

Jonathan Rottenberg and James J Gross. When emotion goes wrong: Realizing the promise of affective science. *Clinical Psychology: Science and Practice*, 10(2): 227–232, 2003.

Clay Routledge, Tim Wildschut, Constantine Sedikides, Jacob Juhl, and Jamie Arndt. The power of the past: Nostalgia as a meaning-making resource. *Memory*, 20(5): 452–460, 2012.

Arunima Roy, Katerina Nikolitch, Rachel McGinn, Safiya Jinah, William Klement, and Zachary A Kaminsky. A machine learning approach predicts future risk to suicidal ideation from social media data. *NPJ Digital Medicine*, 3(1):1–12, 2020.

Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8): 1121–1133, 2004.

José A Ruiz-Caballero and José Bermúdez. Neuroticism, mood, and retrieval of negative personal memories. *The Journal of general psychology*, 122(1):29–35, 1995.

James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

James A Russell. Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145, 2003.

Cheryl L Rusting. Personality, mood, and cognitive processing of emotional information: three conceptual frameworks. *Psychological bulletin*, 124(2):165, 1998.

Cheryl L Rusting and Randy J Larsen. Moods as sources of stimulation: Relationships between personality and desired mood states. *Personality and individual differences*, 18(3):321–329, 1995.

Richard M Ryan and Edward L Deci. On happiness and human potentials: A review of research on hedonic and eudaimonic well-being. *Annual review of psychology*, 52(1):141–166, 2001.

Matthew E Sachs, Antonio Damasio, and Assal Habibi. The pleasures of sad music: a systematic review. *Frontiers in human neuroscience*, 9:404, 2015.

Sukanta Saha, David Chant, Joy Welham, and John McGrath. A systematic review of the prevalence of schizophrenia. *PLoS Med*, 2(5):e141, 2005.

Michael Sanderson, Andrew GM Bulloch, JianLi Wang, Kimberly G Williams, Tyler Williamson, and Scott B Patten. Predicting death by suicide following an emergency department visit for parasuicide with administrative health care system data and machine learning. *EClinicalMedicine*, page 100281, 2020.

Elvis Saravia, Chun-Hao Chang, Renaud Jollet De Lorenzo, and Yi-Shin Chen. Midas: Mental illness detection and analysis via social media. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 1418–1421. IEEE, 2016.

Klaus R Scherer. Which emotions can be induced by music? what are the underlying mechanisms? and how can we measure them? *Journal of new music research*, 33 (3):239–251, 2004.

Klaus R Scherer, Marcel R Zentner, et al. Emotional effects of music: Production rules. *Music and emotion: Theory and research*, 361(2001):392, 2001.

Ulrich Schimmack, Shigehiro Oishi, and Ed Diener. Cultural influences on the relation between pleasant emotions and unpleasant emotions: Asian dialectic philosophies or individualism-collectivism? *Cognition & Emotion*, 16(6):705–719, 2002.

Kimberly T Schneider and Nathan J Carpenter. Sharing# metoo on twitter: incidents, coping responses, and social reactions. *Equality, Diversity and Inclusion: An International Journal*, 2019.

Felix D Schönbrodt and Marco Perugini. At what sample size do correlations stabilize? *Journal of Research in Personality*, 47(5):609–612, 2013.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):234, 2013.

H Andrew Schwartz, Johannes Eichstaedt, Margaret Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. Towards assessing changes in degree of depression through facebook. In *Proceedings of the workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, pages 118–125, 2014.

Shalom H Schwartz. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in experimental social psychology*, volume 25, pages 1–65. Elsevier, 1992.

Elizabeth M Seabrook, Margaret L Kern, Ben D Fulcher, and Nikki S Rickard. Predicting depression from language-based emotion dynamics: longitudinal analysis of facebook and twitter status updates. *Journal of medical Internet research*, 20(5): e168, 2018.

Rico Sennrich. Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 539–549. Association for Computational Linguistics, 2012.

F. M. Shah, F. Ahmed, S. K. Saha Joy, S. Ahmed, S. Sadek, R. Shil, and M. H. Kabir. Early Depression Detection from Social Network Using Deep Learning Techniques. In *2020 IEEE Region 10 Symposium (TENSYMP)*, pages 823–826, 2020. doi: 10.1109/TENSYMP50017.2020.9231008.

Dipti Sharma, Munish Sabharwal, Vinay Goyal, and Mohit Vij. Sentiment analysis techniques for social media data: A review. In *First International Conference on Sustainable Technologies for Computational Intelligence*, pages 75–90. Springer, 2020.

Adrian B. R. Shatte, Delyse M. Hutchinson, Matthew Fuller-Tyszkiewicz, and Samantha J. Teague. Social Media Markers to Identify Fathers at Risk of Postpartum Depression: A Machine Learning Approach. *Cyberpsychology Behavior and Social Networking*, 23(9):611–618, 2020. doi: 10.1089/cyber.2019.0746.

Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, Tat-Seng Chua, and Wenwu Zhu. Depression Detection via Harvesting Social Media: A Multimodal Dictionary Learning Solution. In *IJCAI*, pages 3838–3844, 2017.

Tiancheng Shen, Jia Jia, Guangyao Shen, Fuli Feng, Xiangnan He, Huanbo Luan, Jie Tang, Thanassis Tiropanis, Tat Seng Chua, and Wendy Hall. Cross-domain depression detection via harvesting social media. In *International Joint Conferences on Artificial Intelligence*, 2018.

Yu-Chun Shen, Tsung-Ting Kuo, I-Ning Yeh, Tzu-Ting Chen, and Shou-De Lin. Exploiting temporal information in a two-stage classification framework for content-based depression detection. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 276–288. Springer, 2013.

Gal Sheppes, Gaurav Suri, and James J Gross. Emotion regulation and psychopathology. *Annual review of clinical psychology*, 11:379–405, 2015.

Benjamin Shickel, Scott Siegel, Martin Heesacker, Sherry Benton, and Parisa Rashidi. Automatic detection and classification of cognitive distortions in mental health text. In *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 275–280. IEEE, 2020.

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

David H Silvera, Anne M Lavack, and Fredric Kropp. Impulse buying: the role of affect, social influence, and subjective wellbeing. *Journal of Consumer Marketing*, 25(1):23–33, 2008.

T Simms, C Ramstedt, M Rich, M Richards, T Martinez, and C Giraud-Carrier. Detecting cognitive distortions through machine learning text analytics. In *2017 IEEE*

international conference on healthcare informatics (ICHI), pages 508–512. IEEE, 2017.

Kamlesh Singh and Shalini Duggal Jha. Positive and negative affect, and grit as predictors of happiness and life satisfaction. *Journal of the Indian Academy of Applied Psychology*, 34(2):40–45, 2008. doi: 10.1038/s41746-020-0233-7.

Ruba Skaik and Diana Inkpen. Using social media for mental health surveillance: A review. *ACM Computing Surveys (CSUR)*, 53(6):1–31, 2020.

George M Slavich, Aoife O'Donovan, Elissa S Epel, and Margaret E Kemeny. Black sheep get the blues: A psychobiological model of social rejection and depression. *Neuroscience & Biobehavioral Reviews*, 35(1):39–45, 2010.

Robert J Smith, Patrick Crutchley, H Andrew Schwartz, Lyle Ungar, Frances Shofer, Kevin A Padrez, and Raina M Merchant. Variations in facebook posting patterns across validated patient health conditions: a prospective cohort study. *Journal of medical Internet research*, 19(1):e7, 2017.

Timothy W Smith, Alan J Christensen, Judith R Peck, and John R Ward. Cognitive distortion, helplessness, and depressed mood in rheumatoid arthritis: a four-year longitudinal analysis. *Health Psychology*, 13(3):213, 1994.

Shamsah B Sonawalla and Maurizio Fava. Severe depression. *CNS drugs*, 15(10): 765–776, 2001.

Xiangguo Sun, Bo Liu, Qing Meng, Jiuxin Cao, Junzhou Luo, and Hongzhi Yin. Group-level personality detection based on text generated networks. *World Wide Web*, pages 1–20, 2019.

Angelina R Sutin, Antonio Terracciano, Yuri Milaneschi, Yang An, Luigi Ferrucci, and Alan B Zonderman. The trajectory of depressive symptoms across the adult life span. *JAMA psychiatry*, 70(8):803–811, 2013.

Johan AK Suykens and Joos Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999.

Shaheen Syed and Marco Spruit. Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. In *2017 IEEE International conference on data science and advanced analytics (DSAA)*, pages 165–174. IEEE, 2017.

M. M. Tadesse, H. Lin, B. Xu, and L. Yang. Detection of Depression-Related Posts in Reddit Social Media Forum. *IEEE Access*, 7:44883–44893, 2019. doi: 10.1109/ACCESS.2019.2909180.

C. Tai, Z. Tan, Y. Lin, and Y. Chang. Mental Disorder Detection and Measurement Using Latent Dirichlet Allocation and SentiWordNet. In *2015 IEEE International Conference on Systems, Man, and Cybernetics*, pages 1215–1220, 2015. doi: 10.1109/SMC.2015.217.

Liila Taruffi and Stefan Koelsch. The paradox of music-evoked sadness: an online survey. *PLoS One*, 9(10), 2014.

Yla R Tausczik and James W Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010.

Simone Teufel. *Argumentative Zoning: Information Extraction from Scientific Articles*. PhD thesis, Centre for Cognitive Science, University of Edinburgh, 1999.

Mike Thelwall. The heart and soul of the web? sentiment strength detection in the social web with sentistrength. In *Cyberemotions*, pages 119–134. Springer, 2017.

Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558, 2010.

Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. Sentiment in twitter events. *Journal of the American Society for Information Science and Technology*, 62(2):406–418, 2011.

Renee J Thompson, Jutta Mata, Susanne M Jaeggi, Martin Buschkuehl, John Jonides, and Ian H Gotlib. The everyday emotional experience of adults with major depressive disorder: Examining emotional instability, inertia, and reactivity. *Journal of abnormal psychology*, 121(4):819, 2012.

Robert Thorstad and Phillip Wolff. Predicting future mental illness from social media: A big-data approach. *Behav Res Methods*, 51(4):1586–1600, 2019. doi: 10.3758/s13428-019-01235-z.

Xianyun Tian, Philip Batterham, Shuang Song, Xiaoxu Yao, and Guang Yu. Characterizing Depression Issues on Sina Weibo. *Int J Environ Res Public Health*, 15(4), 2018. doi: 10.3390/ijerph15040764.

M. Tlachac and E. Rundensteiner. Screening For Depression With Retrospectively Harvested Private Versus Public Text. *IEEE Journal of Biomedical and Health Informatics*, 24(11):3326–3332, 2020. doi: 10.1109/JBHI.2020.2983035.

Lei Tong, Qianni Zhang, Abdul Sadka, Ling Li, Huiyu Zhou, et al. Inverse boosting pruning trees for depression detection on twitter. *arXiv preprint arXiv:1906.00398*, 2019.

Olivier Toubia and Andrew T Stephen. Intrinsic vs. image-related utility in social media: Why do people contribute content to twitter? *Marketing Science*, 32(3): 368–392, 2013.

Madhukar H Trivedi. The link between depression and physical symptoms. *Primary care companion to the Journal of clinical psychiatry*, 6(suppl 1):12, 2004.

Daniel Trottier and Christian Fuchs. *Social media, politics and the state: protests, revolutions, riots, crime and policing in the age of Facebook, Twitter and YouTube*. Routledge, 2014.

Deborah L Trout. The role of social isolation in suicide. *Suicide and Life-Threatening Behavior*, 10(1):10–23, 1980.

Sho Tsugawa, Yusuke Kikuchi, Fumio Kishino, Kosuke Nakajima, Yuichi Itoh, and Hiroyuki Ohsaki. Recognizing depression from twitter activity. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 3187–3196. ACM, 2015.

UMD. Ourdatahelps project, 2021. `https://umd.ourdatahelps.org/`.

K. D. Varathan and N. Talib. Suicide detection system based on twitter. In *2014 Science and Information Conference*, pages 785–788, 2014. doi: 10.1109/SAI.2014.6918275.

Nikhita Vedula and Srinivasan Parthasarathy. Emotional and Linguistic Cues of Depression from Social Media. In *Proceedings of the 2017 International Conference on Digital Health*, pages 127–136. Association for Computing Machin-

ery, 2017. doi: 10.1145/3079452.3079465. URL https://doi.org/10.1145/3079452.3079465.

Lakshmi Vijayakumar, M Suresh Kumar, and Vinayak Vijayakumar. Substance use and suicide:. *Current Opinion in Psychiatry*, 24(3):197–202, 2011. ISSN 0951-7367. doi: 10.1097/YCO.0b013e3283459242. URL https://insights.ovid.com/crossref?an=00001504-201105000-00005.

Erin A Vogel, Jason P Rose, Lindsay R Roberts, and Katheryn Eckles. Social comparison, social media, and self-esteem. *Psychology of Popular Media Culture*, 3(4): 206, 2014.

Bairong Wang and Jun Zhuang. Crisis information distribution on twitter: a content analysis of tweets during hurricane sandy. *Natural hazards*, 89(1):161–181, 2017.

Xiaofeng Wang, Shuai Chen, Tao Li, Wanting Li, Yejie Zhou, Jie Zheng, Qingcai Chen, Jun Yan, and Buzhou Tang. Depression Risk Prediction for Chinese Microblogs via Deep-Learning Methods: Content Analysis. *JMIR Med Inform*, 8(7): e17958, 2020. doi: 10.2196/17958.

Xinyu Wang, Chunhong Zhang, and Li Sun. An improved model for depression detection in micro-blog social network. In *2013 IEEE 13th International Conference on Data Mining Workshops*, pages 80–87. IEEE, 2013a.

Xinyu Wang, Chunhong Zhang, and Li Sun. An improved model for depression detection in micro-blog social network. In *2013 IEEE 13th International Conference on Data Mining Workshops*, pages 80–87. IEEE, 2013b.

Molly McLure Wasko and Samer Faraj. Why should i share? examining social capital and knowledge contribution in electronic networks of practice. *MIS quarterly*, pages 35–57, 2005.

David Watson. *Mood and temperament*. Guilford Press, 2000.

David Watson and Lee Anna Clark. Extraversion and its positive emotional core. In *Handbook of personality psychology*, pages 767–793. Elsevier, 1997.

David Watson, Lee Anna Clark, and Auke Tellegen. Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of personality and social psychology*, 54(6):1063, 1988.

WHO. Who.int, 2019. URL https://www.who.int/news-room/fact-sheets/detail/suicide.

WHO. Depression. https://www.who.int/news-room/fact-sheets/detail/depression, June 2020.

Keith G. Wilson, Dorothyann Curran, and Christine J. McPherson. A burden to others: A common source of distress for the terminally ill. *Cognitive Behaviour Therapy*, 34(2):115–123, 2005. ISSN 1650-6073. doi: 10.1080/16506070510008461. URL https://doi.org/10.1080/16506070510008461.

Akkapon Wongkoblap, Miguel A Vadillo, and Vasa Curcin. A multilevel predictive model for detecting social network users with depression. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 130–135. IEEE, 2018.

PinHua Wu, JiaLing Koh, and Arbee LP Chen. Event detection for exploring emotional upheavals of depressive people. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, pages 2086–2095, 2019.

Xingwei Yang, Rhonda McEwen, Liza Robee Ong, and Morteza Zihayat. A big data analytics framework for detecting user-level depression from social networks. *International Journal of Information Management*, 54:102141, 2020. doi: 10.1016/j.ijinfomgt.2020.102141.

Tal Yarkoni. Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of research in personality*, 44(3):363–373, 2010.

Wu Youyou, Michal Kosinski, and David Stillwell. Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112(4):1036–1040, 2015.

Li Yujian and Liu Bo. A normalized levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1091–1095, 2007.

Carrie L Yurica and Robert A DiTomasso. Cognitive distortions. In *Encyclopedia of cognitive behavior therapy*, pages 117–122. Springer, 2005.

Justine Zhang, William Hamilton, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. Community identity and user engagement in a multi-community landscape. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11 of *1*, 2017.

Yong Zhang, Hengfen Li, and Shaohong Zou. Association between cognitive distortion, type d personality, family environment, and depression in chinese adolescents. *Depression research and treatment*, 2011, 2011.

Qiu-Yue Zhong, Elizabeth W Karlson, Bizu Gelaye, Sean Finan, Paul Avillach, Jordan W Smoller, Tianxi Cai, and Michelle A Williams. Screening pregnant women for suicidal behavior in electronic medical records: diagnostic codes vs. clinical notes processed by natural language processing. *BMC medical informatics and decision making*, 18(1):30, 2018.

Michael Zimmer. "but the data is already public": on the ethics of research in facebook. *Ethics and information technology*, 12(4):313–325, 2010.

Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, 2019.

Hamad Zogan, Xianzhi Wang, Shoaib Jameel, and Guandong Xu. Depression detection with multi-modalities using a hybrid deep learning model on social media. *arXiv preprint arXiv:2007.02847*, 2020.