NAIST-IS-MT1651218

Master's Thesis

Melancholic Ramblings on Twitter: Detecting Depression from Social Media Activity

Camille Marie Ruiz

September 13, 2018

Graduate School of Information Science
Nara Institute of Science and Technology

A Master's Thesis
submitted to Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
MASTER of SCIENCE

Camille Marie Ruiz

Thesis Committee:

Professor Yuji Matsumoto	(Supervisor)
Professor Satoshi Nakamura	(Co-supervisor)
Assistant Professor Hiroki Tanaka	(Co-supervisor)
Associate Professor Eiji Aramaki	(Co-supervisor)

Melancholic Ramblings on Twitter: Detecting Depression from Social Media Activity*

Camille Marie Ruiz

Abstract

Depression is a serious and prevalent mental illness, affecting 300 million people worldwide. One symptom of depression is rumination, wherein a person thinks about a thought repetitively. Social media sites, such as Twitter, are rich sources of data from which rumination may be detected since these are avenues where people can ramble about their thoughts and feelings. In light of the repetitive nature of rumination, this study uses Twitter data to explore whether or not coherence over a period of time can detect rumination and help in estimating depression. For this task, cosine similarity was used to calculate coherence while time window sizes were applied to account for period of time. Using different SVM kernels to estimate depression, we found that our coherence feature only produced a negligible effect in accuracy across all time window sizes. Nonetheless, the coherence feature was part of the top important features using the SVM linear kernel, indicating its potential. We also found that mental health topics for depressed users and topics on activities for healthy users were used as important features for classifying users regardless of time window size.

Keywords:

social media, depression, Twitter, NLP

*Master's Thesis, Graduate School of Information Science,
Nara Institute of Science and Technology, NAIST-IS-MT1651218, September 13, 2018.

Contents

List of Figures

1. Introduction

Depression is a serious and prevalent illness today—affecting at least 300 million people in the whole world [1]. While it is generally associated with the feeling of sadness, depression can range from being mild to severe depending on various factors and symptoms such as the existence of a cause, the length of the experience of unhappiness, and the ability to function socially or at work [2]. In its severe form, depression can cause death through suicide; this makes it even more imperative that depression is detected and treated as soon as possible.

In light of this, many efforts have been made to recognize depression early on, and one of the avenues of these efforts is through social media. Social media has become a relevant aspect in our lives today as a way to connect to others and express oneself. This makes social media a rich source of data that may be informative about the state of its users. For instance, machine learning has been used on data from social media sites like Twitter to estimate whether or not a user is at risk for depression [3–6]. In doing this, many researchers consider possible symptoms of depression like depressive language use as well as insomnia through linguistic properties of tweet texts and the time the tweets were made.

For this research, we would like to make the first step of exploring whether or not considering *rumination*–a possible symptom of depression–can aid in estimating risk of depression using social media. Rumination is a cognitive process of repetitively thinking of a certain thought, usually with negative affect [7]. In the 5th edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5), the recurrent thought of death and suicide–which is a form of rumination–is one of the key symptoms of major depressive disorder (MDD) [8]. Moreover, many studies support that rumination is a feature of MDD [9, 10].

Because of the repetitive nature of the symptom, we speculate that rumination can be detected on social media through measuring the coherence of a user's

posts. With social media as an avenue to express one's thoughts and feelings, a depressed person who ruminates may ramble as she expresses her repetitive thoughts on a social media site like Twitter. If this is so, then it is possible that depressed users tend to be more coherent on social media compared to healthy users. Assuming this is the case, coherence as a feature in estimating depression would then positively affect the results.

However, coherence of posts should be measured and compared over different periods of time. If we simply take the whole time line of the user, both depressed and healthy users would probably be incoherent since this would often include posts that are years apart. On the other hand, if we take posts in short time spans–like a day or a week–it is possible that the level of coherence of depressed and healthy users would be the same since it is intuitive to suppose that users generally talk about the same topic or event across posts in short periods of time. Considering this, we would like to measure and compare the coherence of posts across different periods of time through dividing the user's data according to different time window sizes.

The goal of this research then is to explore the effect of coherence across different periods of time (time window sizes) on the results of estimating depression using Twitter data. In pursuing this goal, we also investigate the effects of different time window sizes on identifying depression. Our tasks to reach our objectives are as follows:

1. execute and evaluate a process that identifies depressed users and healthy controls based on Twitter data

2. implement a machine learning algorithm that estimates whether the user is depressed or healthy using different time window sizes

3. apply a method that measures the coherence of a user's time line based on different time window sizes

4. apply the machine learning algorithm from 2 with coherence as an additional feature

5. compare and analyze results from 2 and 4

From the results, we found that the accuracy across time window sizes slightly increased when including the coherence feature in our classification. However, these increases are insignificant since the changes range from -0.0002 to 0.0048. With regard to effects of time window sizes, we investigated the important features used in classifying depressed users and healthy controls. We observed that for depressed users, important keywords involve mental health topics such as depression and anxiety across different time window sizes. On the other hand, the important features for healthy controls include topics on various activities such as business and travel.

The novelty of this research centers on both the use of a coherence feature and the application of time window size in estimating depression. Although small, the changes brought about by the coherence feature across time window mostly improved the accuracy of the classification. This may indicate the potential of this feature especially if our data set was made larger and robust. Investigating feature importance also supports this potential since coherence is part of the top important features in all time window sizes. Moreover, we found that time window size significantly affects the results of estimating depression.

2. Review of Related Literature

In the general usage of the term, depression is often associated with having a low mood that may be treated as a symptom of certain events or medical conditions. However, there is a form of depression that may be classified as a mental illness instead of a symptom; this is formally called major depressive disorder (MDD). According to DSM-5, a person may be diagnosed to have MDD when the person, in a 2-week period, experiences a depressed mood or a loss of interest [8]. In the same 2-week period, the person should also have at least 4 other symptoms such as insomnia, fatigue, feelings of worthlessness, and recurrent thoughts of death.

Efforts have been made to better recognize and understand depression, and a possible avenue for this is through information and communications technology (ICT). For example, a study by Wang et al. utilized smart phones and wearables to gather behavioral information about college students with depression [11]. Similarly, another study collects smartphone sensing data to predict clinical depression [12]. Exploring other sensing technologies, Alghowinem et al. builds a machine learning algorithm using features extracted by affective sensing technology from speaking behaviour, eye activity, and head pose of severely depressed patients [13].

Another form of ICT used by researchers in identifying depression is social media. There are many advantages in using social media data. Social media is rich in information since it is widely used worldwide. Moreover, collection of this data is sometimes made accessible through API. Finally, gathering social media data can be less costly compared to data from other ICTs such as sensing technologies because less hardware is involved.

In fact, the use of social media data for public health has been explored by a multitude of studies in the past years. In tandem with machine learning techniques, this often involves detecting certain outbreaks or epidemics like in-

fluenza [14–16], dengue [17], and ebola [18]. Mental health has also been a target for surveillance through social media. Focusing on different illnesses like major depressive disorder, attention deficit hyperactivity disorder (ADHD), and post-traumatic stress disorder (PTSD), studies range from analyzing the language used by the afflicted [19, 20] to using machine learning algorithms for detecting mental disorders [3, 5].

Multiple studies use social media data to aid in identifying depression and gaining insights on the behavior of the depressed [3,5,6]. However, there are many studies similar to depression classification. For example, estimating depression on social media is closely related to sentiment analysis, especially since depression mostly involves negative emotion. However, it is important to note that the topics of the two tasks are different. Although depression involves negative emotion such as sadness, having negative emotions is not a disorder since it is a common experience of having a low mood for a specific time of event. In light of this, most sentiment analysis research on social media classify individual posts, such as a single tweet on Twitter [21]. On the other hand, studies on predicting depression classify by user by looking at sets of posts or tweets. Another similar field of research is bot detection [22]. This is because bot detection, like depression estimation, delineates on the level of users, not individual posts. Nonetheless, these two differ because the nature of their targets are different. The behavior of depressed user is dynamic since depression has many levels and depressive episodes may be spread out over time. In light of this, only a subset of tweets would indicate that the user is depressed. In contrast, bots are more consistent and rigid. If a user time line is completely generated by a bot, each tweet would display signs that this was computer generated.

Identifying and exploring depression through social media consists of multiple challenges. An example of this would be data collection. Collecting and identifying afflicted and non-afflicted users is usually costly since this often involves employing crowdsourcing and surveys. For instance, De Choudhury et al. utilized Amazon's Mechanical Turk service for individuals to take a clinical depression survey and opt to provide access to their public tweets [3]. In light of this, Coppersmith et al. proposed and evaluated a cheap method to gather data by collecting users who self-reported to have mental illnesses [4]. In the study,

they validated this method by showing that there is a statistical and significant difference between the language use of the control group and the diagnosed group. While their novel data collection model works well with conditions like depression and PTSD, the researchers admit that the procedure may not work as well with less common illnesses and that the method is best treated more as a complementary approach rather than a replacement.

Another challenge is to execute a method for identifying depression. Taking into account multiple aspects available from the Twitter API, the approach of De Choudhury et al. was to introduce various measures that are used as features for supervised learning models such as the Support Vector Machine (SVM) [3]. This includes metrics for insomnia, emotional state, and linguistic styles. On the other hand, Resnik et al. uses a different procedure by exploring the use of topic models to predict depression [6]. More specifically, they compare and examine the use of basic LDA and extensions of LDA (supervised topic models) to analyze the linguistic signals for detecting depression.

The studies discussed approached the problem of detecting depression through social media in light of certain known behaviors of the depressed such as sleeping patterns (insomnia) and language use. Another behavior that can be considered is rumination. Rumination is the repetitive thinking of a certain thought, usually in a negative manner [7]. Although not unique to the disorder, rumination is often seen as a symptom or a key feature of depression [9, 10]. Since rumination can also occur in other mental illnesses, DSM-5 only considers a specific form of rumination as a possible symptom of MDD–namely, ruminating about death or suicide [8]. Nonetheless, many studies suggest that rumination plays a bigger role in depression. For example, studies support the idea that rumination can maintain and aggravate depression [23–25]. Rumination can also be used to predict the onset of depression [26].

For our research, we explore a possible measure of rumination in social media and its efficacy as a feature in detecting depression. In light of the repetitive nature in the content of rumination, we experiment with a measure for coherence. We speculate that users who are depressed tend to be coherent since social media is an avenue to express one's thoughts. However, from the best of our knowledge, we did not find studies that explored coherence and depression.

Nonetheless, coherence is explored in other symptoms related to mental health such as psychosis. Bedi et al. analyzed interviews to evaluate measures of semantic coherence and syntactic markers as features to predict psychosis [27]. To compute for semantic coherence, words were represented as vectors using Latent Semantic Analysis [28] and the cosine of vectors from adjacent phrases were calculated. Along with features from syntactic markers of speech complexity, the statistics from semantic coherence (minimum, mean, median, and s.d.) were fed into a cross-validated classifier. The classifier was able to predict later psychosis onset with 100% accuracy, showing the efficacy of the features of semantic coherence and syntactic markers in predicting psychosis.

Learning from previous studies, we approach our goal of exploring the effect of coherence in estimating depression using some of the methods and procedures discussed. For data collection, Coppersmith et al.'s process [4] is used and evaluated. Similar to Bedi et al. calculation of coherence similarity [27], we would like to use cosine of TF-IDF vectors (not LSA) to compute for similarity between tweets. However, instead of computing coherence of adjacent tweets, we consider multiple pairs of tweets that are within a specified time window size. Finally, like De Choudhury et al. [3], supervised learning models were used with different kernels of SVM.

3. Material

3.1. Data Collection

The general goal of this research is to develop an algorithm that can estimate whether or not a Twitter user is at risk of depression. In light of this, two sets of data are needed. The first is a set of tweets from users who are classified as depressed (DP). The second is a set of tweets from users who are classified as the healthy control (HC). We collected a total of 210 users where 105 users are classified as DP and the other 105 users are classified as HC. After filtering out users with less than 500 tweets, we were left with 85 depressed users and 85 healthy control (see Table 3.1).

Each dataset has its own challenges in terms of collection. Collecting DP requires a method that makes it likely that the users have depression. On the other hand, the collection of HC requires a method where we identify healthy or normal users. In this section, we discuss how we tackled each of these challenges in the following sections: (1) collecting depressed users and (2) collecting healthy control.

Table 3.1.: Statistics of User Collection: 170 users amounting to around 300,000 tweets compose the data set after preprocessing. The mean, median, min, and max are based on the total tweets per user.

	No. of users	Total tweets	Mean	Median	Min	Max
DP	85	150,643	1,772.27	1,715	511	3187
HC	85	163,494	1,952.30	1,874	519	3200

3.1.1. Collecting Depressed Users

The challenge with collecting depressed users is to find a reliable method that can identify a user as depressed. For instance, simply checking whether a user tweets "I feel depressed" is not enough to classify that user as depressed because this is a common saying even among people who do not have depression. Moreover, being depressed is a symptom of multiple mental illnesses. In light of this, we target users who report that they have a specific mental illness, particularly major depressive disorder.

For this, we refer to the method used for collecting depressed users from by Coppersmith et al. [4] which was used in the CLPsych 2015 Shared Task [5]. Similar to Coppersmith's method, we collected tweets that mention "major depressive disorder." Note that when a tweet is collected, the data consists of the text, the user, time stamp, and location (if available) of the tweet. A human annotator then selects tweets that indicate that the user has major depressive disorder such as "I was diagnosed with major depressive disorder" or "I have major depressive major disorder" (see Table 3.2). This includes filtering out tweets such as jokes, quotes, news reports, and tweets that refer to other people. 830 tweets were collected with the keyword "major depressive disorder" over a period of a month (2018 January to 2018 February). Only 105 tweets passed our criteria of indicating that the user has major depressive disorder, from which we get a list of users who we consider to have depression. Note that some users in this list may have other disorders in addition to MDD.

Given this user list, we then collect up to 3,200 tweets of each user. We removed all users with less than 500 tweets, leaving 85 users. We also cleaned the tweets by converting the text to lowercase, removing punctuations from the texts, and replacing mentions with "somementioneduser" and links with "somelink." The tweets of these users then compose the dataset of depressed users.

3.1.2. Collecting Healthy Controls

For collecting healthy controls, the challenge is to estimate whether a user is normal in relation to the Twitter community. There are two parts to this challenge. The first concerns the limitations of the Twitter API [29], while the second is

Table 3.2.: MDD Tweets: examples of tweets found using the "major depressive disorder" keyword

Tweets Qualified as Diagnosis
<somementioneduser>As someone with ptsd, generalized anxiety, and major depressive disorder who is also. . .
As a visual artist who also suffers from Major Depressive Disorder this creeps up on me from time
My diagnosis so far: Major Depressive disorder, condition severe General Anxiety Disorder, condition severe Panic...
today i was diagnosed with major depressive disorder anyway has anyone here taken cipralex antidepressants...
Tweets Disqualified as Diagnosis
Predicting treatment outcomes of major depressive disorder by early improvement in painful physical symptoms <someuser>
So, my doctor thinks I may have Bipolar II instead of major depressive disorder.
Yeah problems like depression need to be tackle for what it is as a major depressive disorder and not ignored
Jose is a survivor of political torture..ptsd..& major depressive disorder. Here is the Law Statue of Jose's...

about the measure used to classify the user as part of the healthy control.

The first part of this endeavor is confronting the limitation the Twitter API gives in terms of collecting tweets. When using the API, at least one of the following should be given: a location or a keyword. In other words, a researcher cannot simply collect a random tweet from a random location. Since collecting from a single location will not represent what is standard, we collected 1300 tweets from 8 different locations, namely: Beijing, Capetown, Dublin, London, New York, Quezon City, Singapore, and Tokyo.

The second part of this challenge is to use a measure that can identify healthy controls. Simply collecting random tweets from various locations is not enough because there is still a large chance we collect a depressed user and label this user as healthy. Coppersmith et al. generated their control group by taking a random list of users from a previous collection and filtered the users based on number of tweets and language (should mostly be English) [4]. In the study, they admit of the possibility that their control set is contaminated by depressed users and made no attempt to remove them from the list. Nonetheless, they assert that their collection method is promising using various experiments for validation.

Although the research of Coppersmith et al. yielded promising results, we hope to lower the chance of including a depressed user by adding a filter based on word frequency. Our assumption is that a normal user's tweets would contain words that are commonly used by most users, while non-typical users would tend to deviate from using common words. The filter is applied through the use of the Spearman's rank correlation coefficient, or Spearman's Rho, to estimate for a healthy user according to word frequency.

Considering these challenges, we first collected a total of 1300 English tweets from unique users with no keyword from Beijing, Capetown, Dublin, London, New York, Quezon City, Singapore, and Tokyo. We then collect up to 3200 tweets of each unique user, which is the maximum number of tweets per user made available by the Twitter API [29]. For each user, a score was calculated using the Spearman's rho on word frequency rankings. Users under healthy controls are then taken from the set of users with scores that fall within a certain score range. To identify the score range, we generate the distribution of the users based on the scores and find that the range of 0.70 to 0.80 is where most users fall. In light

of this, 105 random users from with scores within 0.70 to 0.80 were randomly selected.

Only users with more than 500 tweets were retained. Similar to the tweets the depressed set, texts were converted to lowercase and punctuations were removed. Mentions and links were also replaced with "somementioneduser" and "somelink", respectively. After preprocessing, we are left with 85 users whose tweets compose the healthy controls dataset.

Spearman's Rank Correlation Coefficient on Word Frequency Ranking

To generate a score for each user using the Spearman's Rho, we first calculate for the top 100 words for all collected users. For each user, the word frequency ranking is computed in order to extract the user's ranking for each word in the top 100 words. Each user would then have their unique ranking for the top 100 words. For example, the word "you" ranks as the 5th in the top 100 words for all users. We then get the word frequency ranking for each user and extract the rank of the word "you." If "you" ranks as the 10th for a user, the 5th point in the user's ranking has the value 10 (see Fig. 3.1). The Spearman's Rho is then applied to the overall top 100 word ranking (which is always 1 to 100) against each user's ranking for the top 100 words–producing a score for each user (see Table 3.3 for sample of tweets from users).

3.2. Validation of User Collection

3.2.1. Human Check

To support the validity of the user collection, we employed a basic method of checking the capability of humans to differentiate the two groups based on tweets. While Coppersmith et al. used various procedures to validate their data set, their process did not involve an examination of the raw data by human beings [4]. We addressed this by having individuals guess whether or not the user is from the depressed set or the healthy controls using a small sample of the data.

The issue of the user collection method is the possibility that users are labeled

Tweets From User With 76 Spearman's Rho Score
'<somementioneduser>Lmaooo I really don't know what she was expecting they made it seem like it was a leak I told her th... <somelink>',
'Fixed,<somelink>',
'So I'm at work and they call me and asked me to fix this water fountain . I dead laughted at the lady so hard she g... <somelink>',
'<somementioneduser>I guess you could of been around for that if you early 90s . pizzerias was full of them',
'<somementioneduser>I guess you right or what ever',
'<somementioneduser>I would if they was a quarter and had drinks like barcade nothing like'
Tweets From User With 60 Spearman's Rho Score
'Studies of the benefits of massage demonstrate that it is an effective treatment for reducing stress, pain and musc... <somelink>',
'Plan on getting naked this winter? Has your summer tan disappeared? Drop by Pleasuredrome and use our stand up tann... <somelink>',
'Happy New Year from the Pleasuredrome team. Beat the winter blues and jump into London's largest spa pool. We also... <somelink>',
'Studies of the benefits of massage demonstrate that it is an effective treatment for reducing stress, pain and musc... <somelink>'

Table 3.3.: Sample Tweets Based On Spearman's Rho: the user with score of 76 had more natural tweets compared to user with score of 60, whose tweets seem structured and repetitive

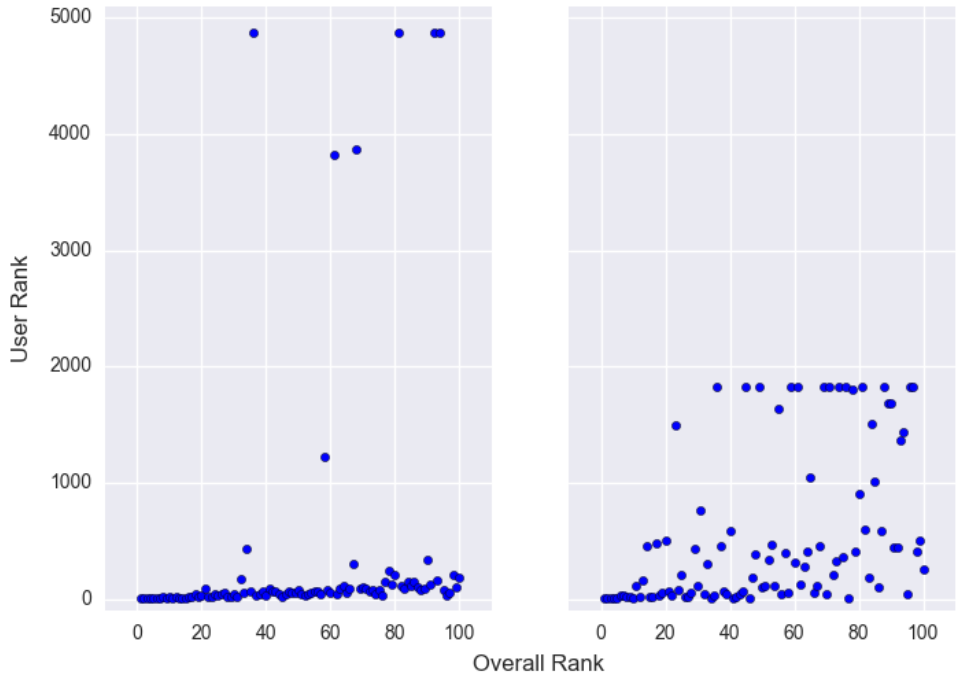

Figure 3.1.: User Rank vs Overall Rank: a user whose Spearman's Rho score is 0.76 (left) has a more correlated ranking with the overall rank than a user with a score of 0.60 (right)

incorrectly. For example, a user labeled as depressed may have lied when stating in a tweet that they have major depressive disorder. It is also possible that a user who has major depressive disorder was by chance included in the healthy controls. If there is a large amount of incorrect labels, the two groups would be difficult to differentiate and there will be little common characteristics to signal the identity of the group.

In light of this, the aim is to find indications that the depressed set and the healthy controls have significant differences and certain characteristics unique to each group. Coppersmith explored this for their method by building classifiers and analyzing the language used that replicates previous findings [4]. Building on this, our task considers the ability of human judgment to delineate between the two groups. While not always correct, humans will have an idea of what characteristics to look for to recognize depression. If human labeling mostly does not match with the grouping in the user collection, this would indicate that the users in the sets are often mislabeled. However, if the accuracy of human labeling is decent and if different individuals generally agree on the labels, this would suggest that there are significant differences between the sets and that the sets have characteristics that most humans commonly expect.

We first generate a sample for the testers to analyze. For extracting a sample from the user collection, 10 depressed users and 10 healthy controls were randomly chosen from the data set. For each of these users, 30 tweets were taken randomly to generate a set of tweets per user. This amounts to 20 sets where each set represents 30 random tweets from a specific user. After shuffling the sets, an answer key was produced by creating a list that indicates whether a certain set came from the depressed users or the healthy controls.

Copies of the sample extracted from the user collection were given to 6 individuals. Three of these individuals are fluent English speakers for both casual and formal situations, with one of them having an undergraduate degree on psychology. The other three have a professional working proficiency in English such that they have a broad vocabulary as well as the ability to engage in formal conversations. These testers were tasked to label each user either as depressed or healthy based on the tweets from the set of the user. From this, 6 answer sets were produced with each corresponding to a tester.

Sample of Successful DP User's Tweets From Human Test
'when sitting in the car for long amounts of time makes your ribs hurt',
'i swear this cbd better help',
'i had a dream i could run',
'when there's rain that means pain',
'<somementioneduser>so do i',
'so much pain running out of meds',
'my ribs are out of place so it hurts to breathe'
Sample of Successful HC User's Tweets From Human Test
'because daiyan is here if we lose tonight it will make my night x100 times worse'
'always believed smes contributed significantly to this but i understand why they do it wasn't aware it would be mo... <somelink>'
'nah i was bussin up throughout that movie it was hilarious <somelink>'

Table 3.4.: Sample Tweets From Human Test: the tweets above came from a DP user and HC user that all testers guessed correctly

Table 3.5.: Human Check Scores: list of scores for every tester

Tester	Raw Score (out of 20)	Percent Accuracy
P1	15	75
P2	15	75
P3	12	60
P4	15	75
P5	14	70
P6	16	80
Average	14.5	72.5

We analyze the answer sets through two measures. First, the score for each tester was computed based on the percentage of correctly labeled users Table 3.5. This is a basic measure to produce the accuracy of the human testers. Second, Cohen's kappa statistic was applied for each pair of answer sets using a function from scikit-learn [30]. The Kappa value is a measure for the consensus of 2 raters, while taking into account the random chance of agreement [31].

The results support that the user collection is reasonable, but not perfect. The average score of the testers was 72.5 %, where fluent speakers scored 75 % in average and the rest averaged with the score of 70 %. The highest score of 80 % was attained by the fluent English speaker with an educational background in psychology. Moreover, the average Kappa value was 0.45 (moderate agreement). These scores indicate that the depressed set and healthy controls set are different enough such that humans can delineate between the two with decent accuracy. This is impressive especially because the testers were limited by the small amount of information provided. The average Kappa value also shows that there is moderate consensus–most testers have the same answer for certain users. This means that testers may see the same signals which may be characteristics that aid in identifying the users. It is still important to note that this does not guarantee that all users from the collection are classified correctly. Nonetheless, the results reinforce Coppersmith's validation of the method especially given the limitations of the sample.

Figure 3.2.: Botometer Score vs. Spearman's Rho Score: each point in the plot represents a user from the 1000 users we collected. Users with low Botometer scores tend to have high Spearman values.

3.2.2. Evaluating the Use of Spearman's Rho

Another way to see if the Spearman rho is reliable in determining the healthy control is to test if it is able to distinguish bots from being a human user. To do this, we used Botometer to evaluate whether or not the user is a bot. Botometer is a tool from a joint project of Indiana University Network Science Institute (IUNI) and the Center for Complex Networks and Systems Research (CNetS) that gives a score to a Twitter account based on how likely it is to be a bot [32]. After obtaining the scores per user from Botometer, we plot the Botometer score against the Spearman's rho score of each user to enable us to evaluate our use of the Spearman's rho method to detect standard users. Botometer scores and Spearman's rho user scores were compared through a scatter plot (see Fig. 3.2). Note that the higher the Botometer score, the more likely the user is a bot. The plot shows that most users with high Spearman's rho scores have low Botometer scores–showing that the use of the scores at least filter's out bots.

4. Method

Three main tasks were taken to explore the efficacy of coherence over different periods of time in estimating depression. These tasks involve (1) time window size application, the (2) coherence function, and (3) classification. The first task was to setup the input according to periods of time by applying time window sizes to the user's data. Next, a coherence function was designed in order to generate coherence features from the input. Finally, two sets of classifiers were built and ran with different feature extraction methods: one set uses a binary vectorizer, and the second set uses both a binary vectorizer and the coherence function.

4.1. Time Window Size Application

To structure the data such that period of time is incorporated, time window size is applied by dividing each user's data into time slices. Time slices are used to organize tweets of each user according to each tweet's time stamp. A time span of a user is a set of tweets that are within the range of the earliest time stamp of the user to the latest time stamp. The size, or the number of days, of this range is called span length.

The user's time span is divided into different time slices. A time slice is a set of tweets that fall the within range of a certain time window size or period. For this paper, we use five types of time slices: 1 week, 1 month, 3 months, 6 months, and 1 year. Measured in days, time window sizes (periods) of the different types of time slices can be seen in table 4.1. Note that if span length s of time span S is larger than the period p of the time slice T, then time slice T is a subset of time span S. On the other hand, if $s \leq p$, time span S and time slice T would contain the same set of tweets.

A time span partitioned (or sliced) according to a period generates a set of

Table 4.1.: Period Per Time Slice: the size of time slices are measured in days

	Period (in days)
1 Week	7
1 Month	30
3 Months	90
6 Months	180
1 Year	365

Table 4.2.: Time Slice Statistics: the mean, standard deviation, minimum, and maximum are calculated from all of the users' tweets according to time slice

	Mean	STD	Min	Max
1 week	22.88	52.73	0	1,005
1 month	95.97	198.65	0	2,608
3 months	272.97	470.19	0	3,155
6 months	503.08	683.50	0	3,155
1 year	857.70	864.64	0	3,155

time slices (see Fig. 4.1 for an example). To be precise, let n be the total number of tweets for user u, S be the time span of user u, T_j be a time slice with period p, x_i be a tweet from user u, and t_i be the time stamp of tweet x_i. Note that the set of tweets are ordered according to the time stamp of each tweet. That is, the sets $[t_0, t_1, ..., t_{n-1}]$ and $[x_0, x_1, ..., x_{n-1}]$ are in ascending order according to the values of t_i. Moreover, every user would have $\lceil \frac{s}{p} \rceil$ time slices per type. Time slice T_j is then generated as a set of tweets that includes x_i where $t_0 + p*j \leq t_i \leq t_0 + p*(j+1)$ and $0 \leq j < \lceil \frac{s}{p} \rceil$.

4.2. Coherence Function

Coherence features are extracted from time slices using a coherence function. This function computes for the average coherence scores between all pairs of tweets in

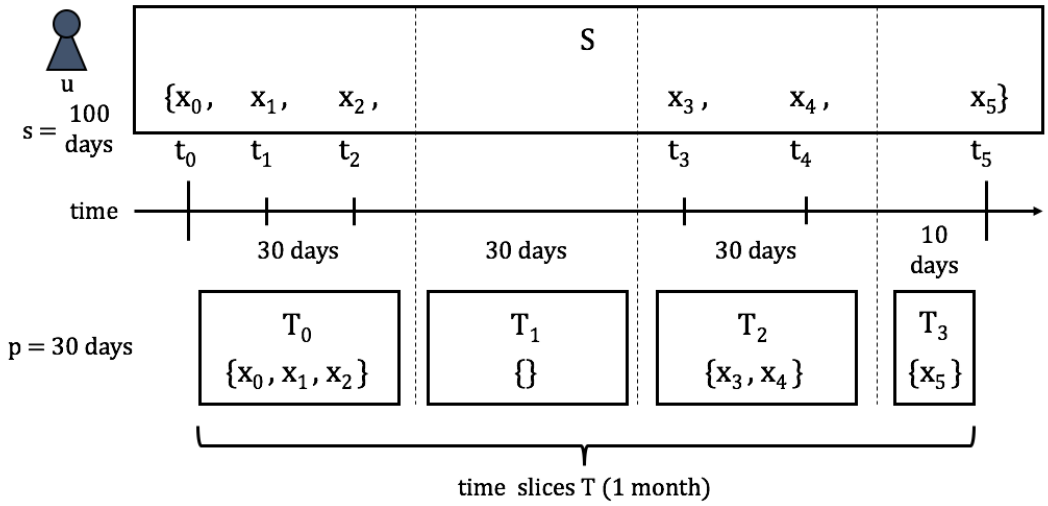

Figure 4.1.: Time Slice Generation: an example of 1 month time slicing with a user whose time line consists of 6 tweets for 100 days. Note that x_i represents a tweet and t_i is the time stamp of tweet x_i.

the time slice. To calculate the coherence score between two tweets, we use cosine similarity and TFIDF. Cosine similarity estimates the similarity of two vectors by computing for the cosine of the angle in between the two vectors. We can get this value by the following function

$$\cos(x, y) = \frac{x \cdot y}{||x|| \cdot ||y||}$$

where x and y are vectors.

When the two vectors x and y are normalized, their magnitude is equal to 1. This way, we can also compute the cosine similarity of two normalized vectors by getting their dot product. In other words, we can change the cosine function for normalized vectors as follows:

$$\cos(x', y') = x' \cdot y'$$

where x' and y' are normalized vectors.

One way to convert text into a normalized vector is through TFIDF. For this research, we find the cosine similarity of two tweets by vectorizing them through

TFIDF then applying the dot product to the two vectors. In light of this, the coherence score of a time slice is extracted by first computing the cosine similarity of every pair of tweets. This gives a set of cosine similarity values; the average of this set is the coherence score for the time slice.

The rest of the subsection discusses the coherence function in detail. We first explain the TFIDF function and how it represents time slices. An elaboration is then made on how this is used with cosine similarity in computing for the coherence score of a time slice.

4.2.1. Representing Time Slices using TFIDF

TFIDF is a method that represents a text as vector as it estimates the importance of words in a set of text. There are two parts to TFIDF: the first is term frequency, and the second is inverse document frequency. A way to compute for term frequency is to get how many times each word in the document appears. Note that time slice T is a set of documents. Let D be a set of documents, d_i be a document in D, and w_{ik} be a word in d_i. This way, we can represent term frequency as

$$tf(w_{ik}, d_i) = f_{w_{ik},d_i}$$

We can get the log normalization of the function as follows:

$$tf'(w_{ik}, d_i) = \log(1 + f_{w_{ik},d_i})$$

On the other hand, inverse document frequency estimates how common or rare a word is in a set of documents. This is done by dividing the total number of documents by the total number of documents a word appears in then taking the logarithm of the result. Inverse document frequency can be represented as

$$idf(w, D) = \log \frac{|D|}{|d \in D : w \in d|}$$

Taking the product of term frequency and inverse document frequency produces TFIDF. TFIDF estimates the importance of a word in a set of texts, usually filtering out common words. TFIDF can then be represented as

$$tfidf(w_{ik}, d_i, D) = tf'(w_{ik}, d_i) \cdot idf(w_{ik}, D)$$

4.2.2. Coherence Function

This previous function returns a single score for w_{ik}. Recall that we need to compare two tweets–two documents–from time slice T. We then need function that returns a vector that represents a document. Let the function be

$$tfidfvec(d_i) =< tfidf(w_{i1}, d_i, D), tfidf(w_{i2}, d_i, D), ..., tfidf(w_{i_{m_i}}, d_i, D) >$$

where m_i is the number of words in the document d_i.

With this function we can compute the cosine similarity of two documents (or tweets). This means we can get the average of all the cosine similarity score of every pair of tweets in time slice T. In other words, we can then use this function to compute for the coherence score of time slice T which can be represented by the following equation

$$\begin{aligned} coherence(T) &= \frac{\sum\limits_{a,b \in T} cos(tfidfvec(a), tfidfvec(b))}{\binom{|T|}{2}} \\ &= \frac{\sum\limits_{a,b \in T} tfidfvec(a) \cdot tfidfvec(b)}{\binom{|T|}{2}} \end{aligned}$$

4.3. Classification

Both using SVM kernels, two pipelines were built which only differ on whether or not the coherence function is applied. The first acts as a baseline such that only a binary vectorizer is used to extract features from the input. On the other hand, the second uses both the binary vectorizer and coherence function for feature extraction. The results of estimating depression without the coherence feature and estimating depression with coherence are put side by side in a table to allow us to compare the accuracies produced by the estimations (see Table 5.1).

4.3.1. Feature Extraction

To generate features from the time slices, we use a custom binary vectorizer that utilizes the top 1000 words of our whole corpus. The top 1000 words are

determined by the frequency of each word in the corpus, where we take the top 1000 words with the highest frequency count. We then represent these top words with a vector of length 1000 with each integer representing one word. When the vectorizer is called, the vector is initialized with all 0 values. It then takes in a list of word lists (tokenized time slice) and checks if a word corresponds to one of the 1000 top words. If the list of words lists contains a top word, the value for that top word is set to 1.

4.3.2. Baseline Classification

For estimating depression, we use time slices as input in a straightforward pipeline that first vectorizes tweets and then classifies them through a SVM kernel. This is done with Stratified KFold where $k = 5$.

A set of shuffled users are split into training and test sets by scikit-learn's Stratified Kfold [30], and the data of these users are then divided into time slices (see Fig. 4.2).

Tweets that are tokenized using stop words from the Natural Language Toolkit (NLTK) [33] are extracted from each user to generate time slices. Each time slice is used as a data point for training/testing and is paired with the label of the user. For example, if there are 10 users and each user for a type of time slice would have 50 time slices, a list with 500 time slices and a list of 500 labels will serve as input to the pipeline.

Each time slice is represented as a feature vector. The feature vector generated from the custom binary vectorizer then serves as the input to the SVM classifier. We experimented with multiple settings and used 3 classifiers: linear, rbf, and polynomial 2. With 5 types of time slices (week, 3 months, 6 months, and 1 year) and 3 kinds of SVM classifiers, we produce 15 different results for estimating depression.

4.3.3. Classification with Coherence Feature

The main pipeline follows the baseline but additionally uses the coherence function for feature extraction (see Fig. 4.3). Recall that the function $coherence(T)$ estimates the coherence of a time slice T by taking the mean of cosine similar-

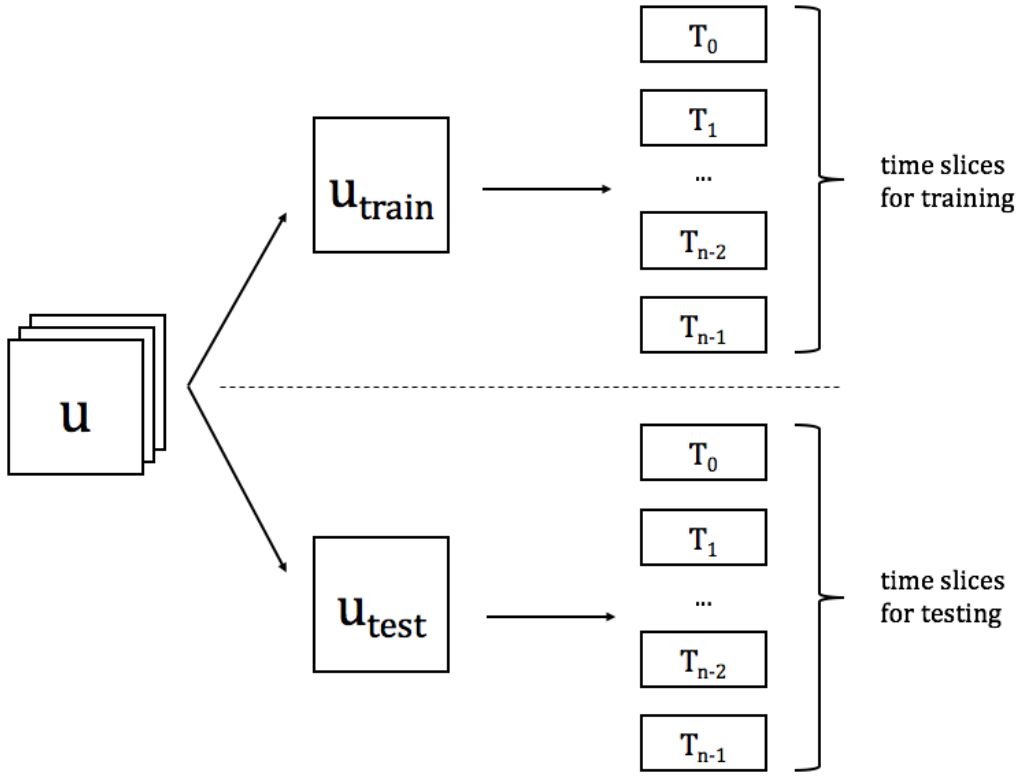

Figure 4.2.: Generating Training and Testing Sets: users are first randomly split into training and testing sets. From these users, time slices are generated and are used as input for training/testing.

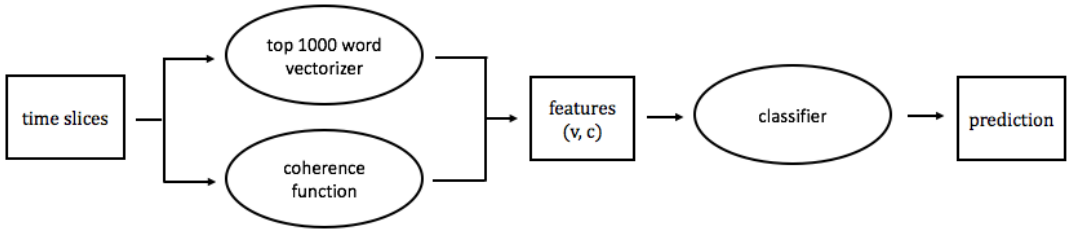

Figure 4.3.: Pipeline with Coherence Feature Extraction: this pipeline uses the coherence function to generate additional features from a time slice.

ity values of each pair of tweets in T. Using this, we can estimate depression with coherence score as an additional feature. The standard deviation, minimum, maximum of the pairwise cosine similarity values were also extracted from T. With these values, the additional feature is a vector with the mean, the standard deviation, the minimum score, and the maximum score of cosine similarity values of each pair of tweets in the time slice. The function to generate the vector is then combined with the top 1000 words vectorizer through scikit-learn's FeatureUnion [30].

Like the baseline, the output of the feature extraction is fed into a classifier. The classifiers used were the same three SVM kernels: linear, rbf, and polynomial 2. 15 results are also produced by experimenting with 5 types of time slices and 3 SVM kernels.

5. Results

5.1. Depression Estimation

With the exception of scores from the RBF kernel, the accuracies produced tend to increase as the time window size increases (see Table 5.1). Consistent with this, the highest accuracy for the linear and polynomial 2 kernels without coherence fall under time window size of a year: 76.65% and 67.80%, respectively. Although the RBF kernel's highest accuracy without coherence also falls under the period of a year (75.10%), the scores for RBF rise until 3 months, slightly falls for 6 months, and rises again for 1 year.

The effect of the coherence feature on the accuracies are negligible (see Table 5.1). Although the accuracies are higher for most of the kernels and time window sizes, the differences are insignificant, as seen in Fig. 5.1, Fig. 5.2, and Fig. 5.3. The largest difference is 0.86%, coming from linear kernel with time window size of a week (see Fig. 5.1).

5.2. User Coherence

To analyze the results generated by the coherence function, box plots of the scores were created (see Fig. 5.4). Most of the scores from all time window sizes are small; these tend to fall below 0.10 for both DP and HC.

	linear		rbf		poly2	
	no coh	with coh	no coh	with coh	no coh	with coh
1 week	0.6540	0.6582	0.6926	0.6955	0.5483	0.5517
1 month	0.6516	0.6573	0.7284	0.7304	0.6121	0.6169
3 months	0.6640	0.6674	0.7416	0.7436	0.6307	0.6353
6 months	0.7083	0.7128	0.7385	0.7398	0.6504	0.6508
1 year	0.7665	0.7686	0.7510	0.7509	0.6780	0.6777

Table 5.1.: SVM Mean ROC AUC: accuracies using the coherence feature (coh) are only negligibly higher than accuracies without the coherence feature

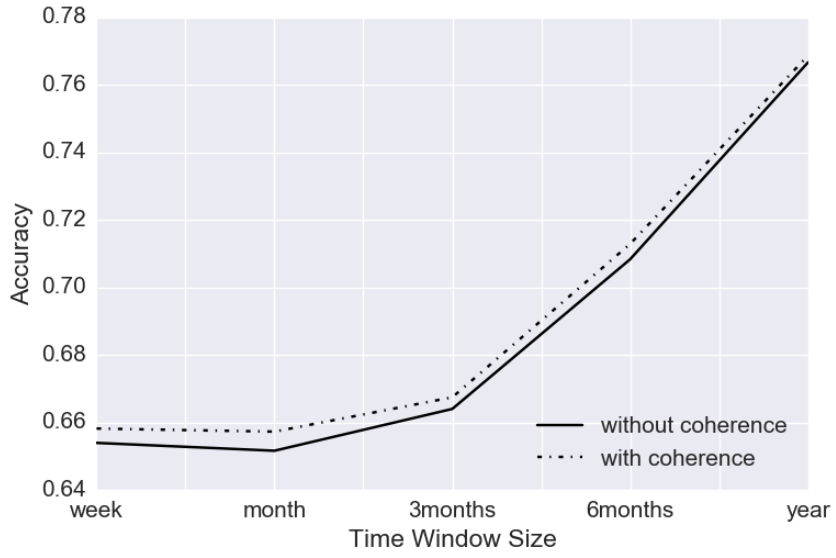

Figure 5.1.: Linear Kernel Scores: the difference between the accuracies of the linear kernel with and without coherence are small

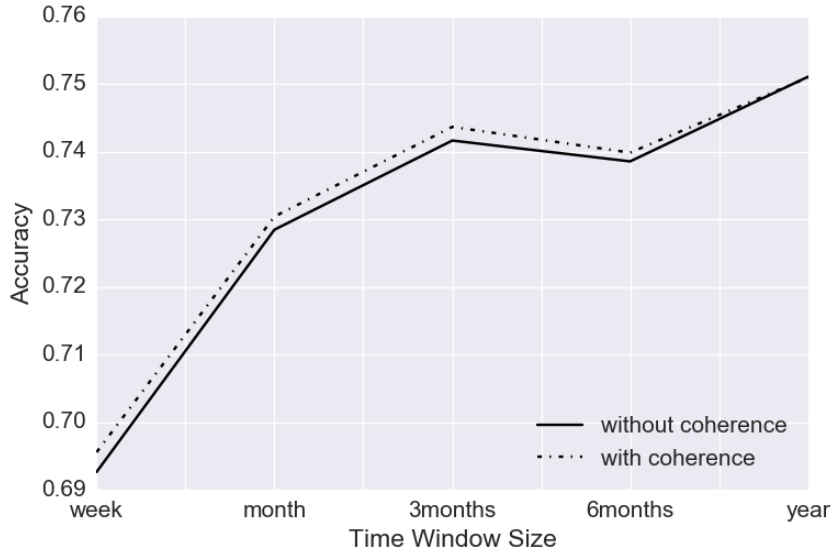

Figure 5.2.: RBF Kernel Scores: like the other kernels, there is only a negligible increase in the accuracy when using the coherence feature with RBF kernel

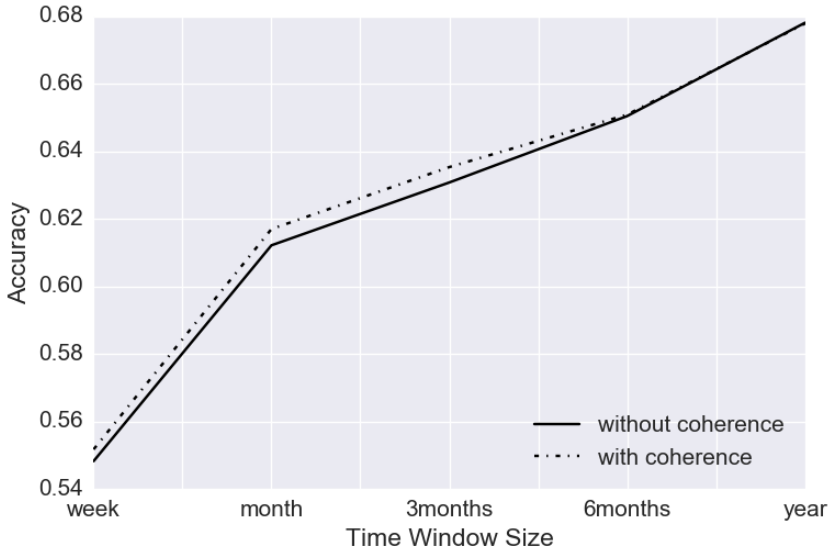

Figure 5.3.: Polynomial 2 Kernel Scores: while these are generally lower than those of the other kernels, the accuracies when using the polynomial kernel follow the same behaviour such that there is only a slight increase when using the coherence feature

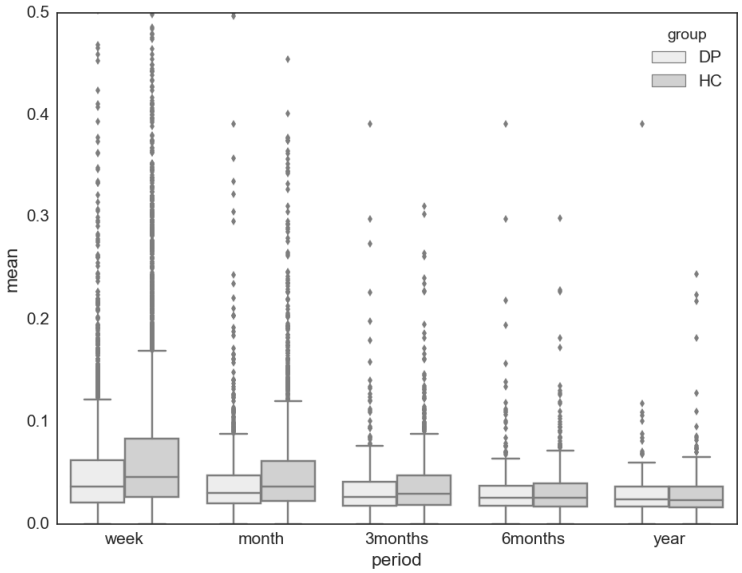

Figure 5.4.: Time Slice Coherence Scores: the values from the coherence function are small for both DP and HC

6. Discussion

6.1. Coherence Feature

The results on estimating depression show that the additional coherence feature does not significantly improve the accuracy of the algorithm across all periods (refer to Table 5.1). Originally, we hypothesized that depressed users would be more coherent over time. In light of this, we expected that the accuracy would significantly improve at least for some time window sizes of a month or three months. While there is some improvement from adding the coherence feature, the difference is too small to make the conclusion that our coherence measure is a strong feature that the classifier can use to delineate depressed users from standard users.

The possible reasons why our results show that there is negligible improvement from the coherence function are both technical and nontechnical. The technical reasons concern the dataset size and the coherence function's efficacy. Since we only have data for 170 users, the dataset size may be too small to test whether an additional feature is useful in estimating depression. Another technical reason comes from the potency of cosine similarity to detect coherence. The function does not take into account semantic similarities as it only relies on patterns in the text. Therefore, tweets that use the same words are more similar than tweets that use different words with the same meaning (see Table 6.1).

Even if the dataset is large and the function considers semantic coherence, there are nontechnical reasons that may cause the same results. The first concerns the assumption that depressed users express their ruminations. It is not necessary that depressed users put their thoughts on social media. In fact, it is possible that healthy controls are generally more expressive than depressed users. This would mean that these users may convey their thoughts on a topic repeatedly, while

phrases	coherence score
'im so sad right now', 'depressing day', 'it sucks to have depression'	0
'im so worthless', 'am i even worth it', 'worthless sigh'	0.3664
'hey hows everyone', 'i hung out with everyone today', 'everyone listen to this'	0.1802

Table 6.1.: Cosine Similarity Efficacy: the function gives very low scores for semantically coherent phrases that use different words

depressed users may be hesitant to express and so only communicate sparingly. In addition to this, depressed users may express their thoughts more on other social media sites, not only on Twitter. In such cases, perhaps rumination can be detected only across different social media sites.

Another explanation would be the possible nature with which depressed users express their thoughts. As mentioned in the DSM-5, people with MDD tend to have impaired social functioning [8]. More specifically, those afflicted with MDD can have difficulty in practicing proper social behavior like withdrawing in the face of emotional expressions [34]. This may render depressed users to be less effective in expression than healthy users. In light of this, it is possible that depressed users communicate their negative thoughts more vaguely compared to standard users. That is, perhaps a healthy user would directly express their sadness through a statement "I am sad today" while a depressed user would quote a song or a poem that indirectly communicates their sadness.

6.2. Accuracies Without Coherence Feature

The accuracy produced by the linear (76.65%) and RBF(75.10%) kernels are decent (see Table 5.1), considering that the results of De Choudhury et al. was 72.38% [3] and that the accuracy from our human check experiment was 72.5% (see Table 3.5). However, this does not show that our classification method is better than either. For De Choudhury et al., their dataset for depressed users is likely more representative of the depressed population on social media in light of their collection method. This makes the depressed users in their data to have more diverse behaviors, making them more difficult to classify. For the human check experiment, the average score from the testers are inconclusive since it is possible that having more testers may increase or decrease the testers' mean accuracy.

6.3. Feature Importance

We used the linear kernel to extract the top features used by the algorithm. The important features or keywords generally change across different time window sizes–this reflects that time window size has some impact on what main features the classifier uses to estimate depression. From these important features in time, we produce some insights on the behavior of depressed users and healthy controls on social media. We were also able to draw signs on the potential of the coherence feature for certain time window sizes.

From the results on feature importance across time window sizes, depressed users seem to be interested on mental health topics. The important features to classify depressed users include words like "depression", "anxiety", and "mental" (see Table 6.2) This reflects that depressed users in the dataset talk about mental health often enough that these keywords are detected even in short periods. However, this also shows how our dataset only contains a subset of depressed users on Twitter. The method we used to collect depressed users would require the user to be vocal about their diagnosis. Therefore, it is not surprising that they would express their thoughts on mental health.

On the other hand, healthy controls seem to talk about topics on different

	DP Important Keywords
1 week	**mhtchat**, getglue, ada, tapi, **depression**, sticker, **mental**, major, human, ako, xd, bc, dog, **anxiety**, pun, ka, stream, pain, stand, president
1 month	**mental**, stand, human, major, sticker, **depression**, cat, cover, **anxiety**, took, bc, fact, xd, president, bet, clean, thoughts, hands, rt, using
3 months	**depression**, human, cause, id, become, name, love, said, major, **mental**, youre, feel, friend, bc, face, ball, took, seen, early,
6 months	like, go, think, one, amp, **depression**, line, **mental**, major, love, **anxiety**, seen, making, hey, im, good, already, waiting, night, god
1 year	would, one, major, **mental**, **depression**, called, early, cats, thanks, **anxiety**, big, fat, huge, awesome, piece, realize, stand, hell, get, cry

Table 6.2.: Important Keyword for Detecting DP: words in bold highlight the inclusion of mental health related terms throughout all time slices.

	HC Important Keywords
1 week	danger, **morocco**, haha, **tokyo**, etc, **fitbit**, china, nak, lfc, gt, surprised, tak, london, **dance**, **business**, **japan**, experience, **steps**, album, brother
1 month	**business**, danger, hahaha, **steps**, side, star, surprised, test, haha, **morocco**, final, train, note, album, happening, songs, **china**, **book**, experience, head
3 months	japanese, **business**, hi, course, soon, short, learning, 100, case, free, chinese, men, well, something, definitely, week, **steps**, test, date, long
6 months	vs, that, short, chinese, fitbit, start, steps, week, **traveled**, cannot, say, **business**, soon, ask, **dance**, home, free, article, great, trying
1 year	**steps**, **fitbit**, week, start, **fitstatsensus**, twitter, great, **traveled**, find, spend, ask, ng, chinese, tomorrow, back, new, question, cannot, two, actually

Table 6.3.: Important Keyword for Detecting HC: words in bold highlight the inclusion of terms related to activities throughout all time slices.

	HC Important Keywords (coh)
1 week	cossim_mean (3)
1 month	cossim_max (14)
3 months	cossim_max (15)
6 months	cossim_max (5)
1 year	cossim_max (9)

Table 6.4.: Important Coherence Features in Detecting HC: some coherence features generated using cosine similarity was included in the top 10 important features for detecting HC. Note that the number in the parenthesis indicates the rank of the feature.

activities. Compared to important keywords for depressed users, there are no common keywords across time window sizes. One possibility for these varying keywords would be because of a lack of data. Since healthy controls are expected to have varied interests, more data would be needed to be able to detect what is generally common among healthy controls. Nonetheless, we observed there are keywords in every time window size that seem to reflect activities or hobbies (see Table 6.3). For instance, for time window sizes of week and month, the important keywords include "travel", country names, and "business." For the rest of the time proximities, words related to fitness like "steps", "fitbit", "fitstats" are in the top features. This may be indicative of how healthy controls tend to take part in more activities compared to depressed users.

The last observation on feature importance concerns the coherence feature. In the first section of this chapter, we discussed how the coherence feature only provided insignificant improvement to the algorithm. Despite this, the feature importance analysis is indicative of the potential of the coherence feature (see Table 6.4). The important feature for time window sizes of 1 month, 3 months, and 6 months include the max cosine similarity score of tweets within the time slice. Moreover, top keywords for time window size of a week includes the coherence score (mean cosine similarity score) of the time slice.

7. Conclusion

Investigating a measure for rumination as a symptom of depression, this study served as a first step in exploring the potential of including coherence over periods of time as a feature in estimating risk of depression using social media data. For our material, we executed and evaluated a method of collecting depressed users and healthy controls from Twitter data. With this data, periods of time was accounted for by slicing the collected user data according to time window slices. We then calculate coherence by applying cosine similarity to the resulting time slices. The time slices were then used as input to SVM classifiers that delineates these slices as coming from depressed users or healthy controls. To evaluate the potential of coherence in estimate depression, the results from the classification using the coherence feature was then compared to the results of classifying without coherence as a feature.

From our experiment, we found that our coherence feature using cosine similarity only produced a negligible effect in accuracy across all time window sizes. We speculate that this is because of the naive metric, as well as numerous factors that were not considered like the possible lack of expression and linguistic style of depressed users. Various insights were also extracted from using different time window sizes by analyzing the important keywords used in the linear SVM kernel. Across time window sizes, the important keywords to identify depressed users from our dataset involved mental health topics while words relating to different activities were important in recognizing healthy controls. In our investigation, we found that coherence was part of the top important features, which may indicate the potential of coherence.

This study functions as a first step towards our goal of exploring the possible use of coherence in estimating depression. In light of this, we can improve on many aspects of our research, particularly our material and method. Our data

collection process can be refined by increasing the number of users considered in the study and by employing clinical depression surveys through crowd sourcing. Building on our efforts of using time window size, grouping tweets according relevant periods of time can also be improved using various clustering techniques. With regard to measuring coherence, we opted to use cosine similarity since we were working on a small dataset. When we have a larger corpus, we can explore other means to calculating coherence such as using Gensim's Doc2Vec [35].

There are also a number of angles we should consider for future work. For example, parts of speech (POS) can be examined. We started to look into this by investigating the ratio of verb and nouns for both the depressed users and healthy controls (Table A.1). The ratio concerning verbs seems to be higher for depressed users, while the ratio for nouns are higher for healthy control. The significance and implications of these results may be explored further in the future.

Another angle to investigate for future work concerns how negative affect should be considered in trying to capture rumination. This study looks into coherence in light of the repetitive nature of rumination. However, rumination for the depressed often involves repetitively thinking about negative thoughts. A healthy user may express repetitive thoughts for various reasons such as their interests or current obsessions. Although we think instances of these are unlikely, such cases should still be considered to produce better results.

Acknowledgements

I would like to thank my supervisors and panelists Professor Yuji Matsumoto, Professor Satoshi Nakamura, Assistant Professor Hiroki Tanaka, and Associate Professor Eiji Aramaki for their guidance and patience as I wrote this thesis. I would also like to thank Shoko Wakamiya and Kaoru Ito for dedicating their time and energy in reading and commenting on my first drafts. I thank my labmates for their ideas and kinship. I especially thank Associate Professor Eiji Aramaki for providing a light yet conducive environment as he heads the social computing laboratory with his brilliance and positive energy. I would also like to thank him and Chihiro Honda for being supportive, kind, and understanding as I struggled with many obstacles in finishing my degree.

I thank the Filipino community in NAIST for making me feel more at home here in Japan. I especially thank my friends Carlo, Nicko, and Jayzon for their unwavering support; their company and the bond we formed certainly contributed in bringing joy and meaning in my stay here in Japan. I thank Victor for all that he has done for me–I will forever be grateful of his effort and sacrifices to bring care, comfort, and meaning to my daily life. I would also like to thank Yya for her continued support–somehow she is able to keep me company even from miles away.

Finally, I thank my dad who has always been integral in my growth as a person and as an academic. He is my inspiration and my raison d'être. He kept me in track by reminding me of what is truly important amidst all the challenges and successes I faced in life. I would not be here without him.

To everyone who has been part of it all, thank you.

References

[1] World Health Organization. Depression, 2018. `http://www.who.int/news-room/fact-sheets/detail/depression`, Last accessed on 2018-06-29.

[2] RH Belmaker and Galila Agam. Major depressive disorder. *New England Journal of Medicine*, 358(1):55–68, 2008.

[3] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. *ICWSM*, 13:1–10, 2013.

[4] 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.

[5] 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.

[6] 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.

[7] Wendy Treynor, Richard Gonzalez, and Susan Nolen-Hoeksema. Rumination reconsidered: A psychometric analysis. *Cognitive therapy and research*, 27(3):247–259, 2003.

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

[9] Costas Papageorgiou and Adrian Wells. An empirical test of a clinical metacognitive model of rumination and depression. *Cognitive therapy and research*, 27(3):261–273, 2003.

[10] Greg J Siegle, Patricia M Moore, and Michael E Thase. Rumination: One construct, many features in healthy individuals, depressed individuals, and individuals with lupus. *Cognitive Therapy and Research*, 28(5):645–668, 2004.

[11] Rui Wang, Weichen Wang, Alex daSilva, Jeremy F Huckins, William M Kelley, Todd F Heatherton, and Andrew T Campbell. Tracking depression dynamics in college students using mobile phone and wearable sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(1):43, 2018.

[12] Asma Ahmad Farhan, Chaoqun Yue, Reynaldo Morillo, Shweta Ware, Jin Lu, Jinbo Bi, Jayesh Kamath, Alexander Russell, Athanasios Bamis, and Bing Wang. Behavior vs. introspection: refining prediction of clinical depression via smartphone sensing data. In *Wireless Health*, pages 30–37, 2016.

[13] Sharifa Alghowinem, Roland Goecke, Michael Wagner, Julien Epps, Matthew Hyett, Gordon Parker, and Michael Breakspear. Multimodal depression detection: fusion analysis of paralinguistic, head pose and eye gaze behaviors. *IEEE Transactions on Affective Computing*, 2016.

[14] Aron Culotta. Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the first workshop on social media analytics*, pages 115–122. ACM, 2010.

[15] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. Twitter catches the flu: detecting influenza epidemics using twitter. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1568–1576. Association for Computational Linguistics, 2011.

[16] David A Broniatowski, Michael J Paul, and Mark Dredze. National and local influenza surveillance through twitter: an analysis of the 2012-2013 influenza epidemic. *PloS one*, 8(12):e83672, 2013.

[17] Janaína Gomide, Adriano Veloso, Wagner Meira Jr, Virgílio Almeida, Fabrício Benevenuto, Fernanda Ferraz, and Mauro Teixeira. Dengue surveillance based on a computational model of spatio-temporal locality of twitter. In *Proceedings of the 3rd international web science conference*, page 3. ACM, 2011.

[18] M Odlum. How twitter can support early warning systems in ebola outbreak surveillance. In *2015 APHA Annual Meeting & Expo (Oct. 31-Nov. 4, 2015). APHA*, 2015.

[19] Nicola J Reavley and Pamela D Pilkington. Use of twitter to monitor attitudes toward depression and schizophrenia: an exploratory study. *PeerJ*, 2:e647, 2014.

[20] Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. From adhd to sad: Analyzing the language of mental health on twitter through self-reported diagnoses. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 1–10, 2015.

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

[22] Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. Who is tweeting on twitter: human, bot, or cyborg? In *Proceedings of the 26th annual computer security applications conference*, pages 21–30. ACM, 2010.

[23] Susan Nolen-Hoeksema. The role of rumination in depressive disorders and mixed anxiety/depressive symptoms. *Journal of abnormal psychology*, 109(3):504, 2000.

[24] Susan Nolen-Hoeksema, Blair E Wisco, and Sonja Lyubomirsky. Rethinking rumination. *Perspectives on psychological science*, 3(5):400–424, 2008.

[25] Bunmi O Olatunji, Kristin Naragon-Gainey, and Kate B Wolitzky-Taylor. Specificity of rumination in anxiety and depression: a multimodal meta-analysis. *Clinical Psychology: Science and Practice*, 20(3):225–257, 2013.

[26] Paul O Wilkinson, Tim J Croudace, and Ian M Goodyer. Rumination, anxiety, depressive symptoms and subsequent depression in adolescents at risk for psychopathology: a longitudinal cohort study. *BMC psychiatry*, 13(1):250, 2013.

[27] Gillinder Bedi, Facundo Carrillo, Guillermo A Cecchi, Diego Fernández Slezak, Mariano Sigman, Natália B Mota, Sidarta Ribeiro, Daniel C Javitt, Mauro Copelli, and Cheryl M Corcoran. Automated analysis of free speech predicts psychosis onset in high-risk youths. *npj Schizophrenia*, 1:15030, 2015.

[28] Susan T Dumais. Latent semantic analysis. *Annual review of information science and technology*, 38(1):188–230, 2004.

[29] Twitter. Api reference index, 2018. `https://developer.twitter.com/en/docs/api-reference-index.html`, Last accessed on 2018-07-08.

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

[31] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[32] Observatory on Social Media, Indiana University Network Science Institute and the Center for Complex Networks and Systems Research. Botometer, 2018. `https://botometer.iuni.iu.edu/#!/`, Last accessed on 2018-07-08.

[33] Steven Bird and Edward Loper. Natural language toolkit, 2017. `https://www.nltk.org/`, Last accessed on 2018-07-08.

[34] Aleksandra Kupferberg, Lucy Bicks, and Gregor Hasler. Social functioning in major depressive disorder. *Neuroscience & Biobehavioral Reviews*, 69:313–332, 2016.

[35] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. `http://is.muni.cz/publication/884893/en`.

Publication List

[1] Camille Ruiz, Kaoru Ito, Shoko Wakamiya, and Eiji Aramaki. Loneliness in a connected world: analyzing online activity and expressions on real life relationships of lonely users. In *The AAAI 2017 Spring Symposium on Wellbeing AI: From Machine Learning to Subjectivity Oriented Computing Technical Report SS-17-08*, pages 726-733, 2017.

Appendices

A. Part of Speech Analysis

Verb		
Tag	**DP**	**HC**
VBP	0.048	0.043
VBD	0.027	0.026
VBN	0.013	0.013
VBG	0.023	0.021
VB	0.049	0.046
VBZ	0.027	0.027
V*	0.187	0.176

Noun		
Tag	**DP**	**HC**
NNP	0.001	0.001
NNS	0.049	0.050
NN	0.301	0.314
NNPS	0.000	0.000
Nouns (N*)	0.351	0.365
PRP	0.042	0.038
PRP$	0.023	0.019
Pronouns (PR*)	0.065	0.057
N* and PR*	0.416	0.422

Table A.1.: Verb and Noun Ratios: all the words in the tweets were tagged and the ratio between number of verbs/noun and total number of words was computed.