INTRODUCTION

Depression is estimated to affect 350 million people worldwide (WHO, 2015). Characterized by feelings of intense sadness or apathy, depression hinders basic aspects of everyday life and can also lead to suicide. The leading cause of disability (WHO, 2015), it is estimated that depression cost approximately $50 billion in lost earnings in the United States in 2010 alone (Greenberg et al., 2015).

Despite its ubiquity, depression is currently underdiagnosed and undertreated. The vast majority of those diagnosed with depression improve after proper treatment, but fewer than half of those struggling with depression receive such treatment (WHO, 2015). At Princeton University, Counseling and Psychological Services (CPS) is the main service that provides treatment and support for those who are struggling with depression. Often students will struggle for several months to a year before visiting CPS, and some students may never seek help at all. Barriers to seeking timely treatment include the social stigma associated with depression, self-denial of illness, and inaccurate assessment and diagnosis. The question then arises, is there any way we can identify those silently struggling with depression and encourage them to seek help sooner rather than later?

With the recent rise of social media, platforms such as Twitter and Facebook have become popular mediums through which people often share personal and candid thoughts, providing snapshots of an individual’s day-to-day life. Thus, it might be possible that the content on these platforms might contain subtle clues on the well-being of their users.
Previous studies have shown that there do exist differences in the language usage between depressed and non-depressed individuals (Rude et al., 2004), and in the last four years, scholars have begun exploring the possibility of using various data analysis techniques on Twitter data to detect signs of depression in its users (Coppersmith et al., 2014; De Choudhury et al, 2013; Park et al., 2014). However, given the relative novelty of such studies, there is still much debate regarding the true capabilities of such a method.

In my thesis, I continue exploring the possibility of using machine learning methods to identify depressed users on Twitter. The results from this study corroborate the idea that we may indeed be able to detect signs of depression by observing aspects of a user’s behavior on Twitter.

**METHODOLOGY**

The procedure to create the classifier, the tool I developed, can be broken down into the three steps. First, I gathered a set of depressed and non-depressed users and their tweets. Afterwards, I identified a set of features, characteristics of a user and his or her tweets, that could be used to differentiate between depressed and non-depressed users. Finally, I used machine learning methods to create a classifier to differentiate between depressed and non-depressed users based on these features.

To gather a group of depressed users, I searched for Twitter users who discussed their depression online within a 24-hour window. I then used five raters to check that these users were genuine in their expression of depression. Any users that were rated to be disingenuous were removed. To find a group of non-depressed users, I randomly

---

1 “Non-depressed users” is a little bit of a misnomer. Because ~5% of the population has depression, by randomly sampling users on Twitter, it is likely that some of my “non-depressed users” do have depression. For the sake of brevity, I will continue to refer to them as non-depressed users. A more nuanced handling of this issue is elaborated upon in my full thesis (Tang, 2016).
sampled Twitter for a set of users within that same 24-hour window. After gathering the depressed and non-depressed users, I obtained the last 3,200 public tweets of each user. I did this procedure over two 24-hour windows, once on February 2-3, 2016 and once on April 4-5, 2016.

After gathering the tweets of the depressed and non-depressed users, I observed many of their behavioral (how a user interacts with Twitter) and linguistic (characteristics of a user’s language on Twitter) features. Behavioral features included how often a user posted each day and how often a user mentioned other users in his or her posts. Linguistic features included how often a user uses first-person pronouns and the average number of terms related to sadness in each tweet. The full set of features observed is included in the full text of my thesis (Tang, 2016).

To put it all together, I used various machine learning methods to create a classifier to differentiate between depressed and non-depressed users based on the features I identified. In my thesis, I specifically used a group of machine learning methods called supervised machine learning. These methods work by first providing a computer with many examples of depressed and non-depressed users. The computer analyzes these examples, and then based on what it observes, “learns” a set of rules through which it can classify future users. For example, a computer might decide that a user who references other users in his or her posts is not likely to be depressed, but if that user also uses many terms related to sadness, then the user is very likely to be depressed. A classifier is then created based on this set of rules. I used the February 2016 group of users as the examples, and then tested the resulting classifier by having it classify the April 2016 group of users.
MAJOR FINDINGS

Classifier Performance

Overall, I was able to create a classifier that could differentiate between depressed and non-depressed users with an approximately 80% rate of accuracy. The relatively good performance of this classifier provides a few key insights. First, it tells us that we may indeed be able to identify users with depression based on their behavior and language on Twitter. This finding is very promising and opens many doors regarding new ways to identify individuals struggling with depression. Secondly, however, this indicates that a lot of work still must be done. Studies have found that incorrectly identifying an individual to have depression can have adverse effects (Sheehan et al., 2013). On the flip side, if a user without depression is not identified, then the user may continue struggling indefinitely without getting the help he or she needs. An 80% performance is far from a sure guarantee that someone who is struggling with depression can be identified, and given the sensitive nature of this topic, further work is necessary before such a tool is implemented.

Some Significant Features

To help gain a better understanding of how my classifier makes its decisions, here I briefly highlight some features that helped the classifier differentiate between depressed and non-depressed individuals:

First-person Pronouns: I found that depressed individuals tend to use more first-person pronouns than non-depressed individuals. One theory in depression literature is that depressed individuals tend to be more self-focused (Pyszczynski et al., 1987). This finding corroborates this theory and seems to indicate that this self-focus also manifests itself in the language a person uses.

Negative Emotion: Another characteristic of the language of depressed users is that depressed individuals tend to use more words that carry negative emotion, such as
“hurt” and “useless,” than non-depressed individuals. This finding corresponds well with the theory that depressed individuals tend to have a negative cognitive bias—they generally have more negative thoughts than positive thoughts (Beck, 1967).

Anxiety: Depressed individuals also tend to use more words related to feelings of anxiety, such as “afraid” and “confused,” than non-depressed individuals. Anxiety is one of the mental disorders that occur most often with depression.

**KEY RECOMMENDATIONS**

While my classifier is not yet accurate enough for implementation, with further research and testing, I believe we can develop a reliable (>95% rate of accuracy) classifier. If such a classifier were to be created, I have three recommendations regarding how CPS could leverage such a tool.

*A Pre-Screening “Quiz”*

People often enjoy using apps such as *Who Will I Marry?* that claim to determine who you are likely to marry or who your best friends are based on your social media posts. CPS may be able to leverage this trend by offering the classifier as a pre-screening device for people to use on their own social media accounts to assess their risk for depression. If a user receives a positive result, then the user could be encouraged to visit CPS or otherwise seek a mental health professional for a formal diagnosis. This classifier could also be used in conjunction with University efforts to promote mental health awareness such as with Mental Health Week. The hope is that by having students use this classifier as a pre-screening tool, they may be able to identify signs of depression early on and seek help sooner rather than later.
An Extra Diagnostic Tool

Diagnosing depression is mainly performed through patient interviews in which mental health professionals interview patients and use the information they provide on possible symptoms and other factors for a diagnosis. My classifier could provide additional insight during the diagnosis process by acting as an additional quantitative metric on whether a user has depression. Furthermore, my classifier can offer information that cannot be obtained in a clinical setting. Posts on social media are done in a natural setting—patients are just talking with friends or their online followers. By incorporating this additional information, this tool could provide another dimension of analysis for clinicians when determining whether a user has depression.

Automatic Monitoring of Symptoms

Beyond just diagnosing depression, this tool could also help with monitoring the status of a patient after he or she has been diagnosed with depression. Current methods of monitoring rely on weekly hour-long meetings. However, it can be hard to gauge how well someone is doing based on an hour of conversation. Such a tool could provide a daily analysis of a patient’s well-being and symptoms based on the posts collected each day. This could provide clinicians and psychotherapists real-time updates on the state of their patients and even alert them if something is awry.

CONCLUSION

In my thesis, I created a classifier that could differentiate between depressed and non-depressed users on Twitter with 80% accuracy. While my classifier is not yet ready for real-world application, my thesis corroborates the idea that it may indeed be possible to differentiate between depressed and non-depressed Twitter users based on subtle linguistic and behavioral cues.
CPS and other mental health clinics can greatly benefit from such a tool through its capabilities as a pre-screening tool, an extra diagnostic tool, and for automatic monitoring of symptoms post-diagnosis. Social media websites themselves could also leverage such a tool. For example, Twitter could incorporate the tool into their platform by setting up a system to immediately contact a trusted family member, friend, or primary care provider if a user were found to be at high risk for depression. My research could also benefit those who are struggling with other mental illnesses. The techniques I have applied in my thesis may also be applied to identify signs of other mental illnesses on various social media platforms. With further research and refinement, such classifiers could be used to encourage those struggling with mental illnesses to seek help and combat one of the largest hidden problems facing our society today.

WORKS CITED


and college counseling center directors annual survey,” 2012.


