# DETECTING DEPRESSION IN SOCIAL MEDIA:
# AN EMOTIONAL ANALYSIS APPROACH

A Thesis

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Master of Science

by

Seunghyun Kim

May 2019

# ABSTRACT

Depression has been an ongoing mental health issue that has been affecting a wide range of humanity, particularly the young adults. To address and observe the more general public in a natural habitat, social media is examined for constructing a system to accurately detect depression. Despite the assiduous effort to construct a novel mechanism to detect depression from social media, behavioral approaches had underlying problems for users with a short activity span. To address this problem, emotion analysis was used as a tool to extract the emotion(s) of a user's post to identify those with depression. Via machine learning techniques to construct an emotion classifier which in turn creates emotion embeddings for a binary classifier, this study proposes a pipeline structure to identify reddit posts from the depression subreddit. The model yielded promising results, introducing emotional analysis as a novel methodology in assessing mental health within social media.

# BIOGRAPHICAL SKETCH

Seunghyun Kim was born in Seoul, Korea. After graduating from high school, Seunghyun enrolled into the Bachelor of Arts program in Computer Science at Cornell University. Seunghyun continued his education at Cornell through the Master of Engineering in Computer Science. Graduating in 2013, Seunghyun came back to Cornell in 2017 and enrolled in the Master of Science program in Computer Science. He will be joining the PhD program at Georgia Institute of Technology starting Fall of 2019.

# ACKNOWLEDGMENTS

I would like to start with my special thanks to my advisor and the chair of my committee, Claire Cardie. Her pieces of advice have helped me very greatly throughout my research and the writing of this paper. I cannot describe fully how much I am thankful to her.

I would also like to thank Lillian Lee. It was her class in natural language processing and social interactions that inspired me to pursue research in the field. The discussions throughout the course sparked my interest in natural language processing. I still vividly remember the lecture that introduced the research on social feedback in Reddit’s weight loss community; the course further intrigued me to look deeper into mental health and social media.

Walker White, Lillian Lee, Anne Bracy, and Daisy Fan helped me hone my skills as a teaching assistant and an instructor for introductory programming in Python. I was very fortunate to have them as my instructor and mentors. Jordan Dotzel and Jimmy Briggs were the best TAs for my Summer course.

My special thanks to Alexander Martin Ruch and Jennifer Ruch, who have been both great friends and project partners. I am lucky to have such talented and supportive friends throughout my Masters. Alex and Jen gave a great deal of help when doing the project on psychic modalities on Reddit. I can never forget Brian Hyun Choi as well; his experience in academic writing helped me further perfect my thesis.

I would like to thank the wonderful friends from various fields who made my two years special and memorable. I had a lot of fun spending time with Min Kyeong Lee, David Yu, Jeyon Kim, Seongha Park, Claudia Heeyeon Ro, and Sangwoo Park. I am very lucky to have made such great friends.

Last but not least, I would like to thank my family for the unconditional support through my journey.

TABLE OF CONTENTS

Biographical Sketch.......................................................................................... III

Acknowledgements ........................................................................................... IV

1. Introduction ....................................................................................................... 1

2. Related Work...................................................................................................... 4

2.1 Depression in Social Media.............................................................................. 5
2.2 Depression Detection in Social Media ............................................................. 6
2.3 Emotional Analysis ........................................................................................... 7
2.4 Sentence Classification..................................................................................... 9

3. Dataset ................................................................................................................ 9

3.1 EmotionLines ................................................................................................. 10
3.2 Reddit Dataset ................................................................................................ 11

4. Methodology...................................................................................................... 12

4.1 Models ............................................................................................................ 13
4.2 Emotion Embeddings ...................................................................................... 14
4.3 Word Embeddings ........................................................................................... 15
4.4 Statistical Methods .......................................................................................... 15

5. Experiments....................................................................................................... 16

5.1 Emotion Classifier............................................................................................ 16
5.2 Binary Classifier............................................................................................... 18

6. Results ............................................................................................................... 19

6.1 Emotion Classifier............................................................................................ 20
6.2 Binary Classification through Emotion Embeddings ........................................ 22

7. Discussion and Future Work ............................................................................ 25

7.1 Limitations....................................................................................................... 25
7.2 Conclusion and Future Work........................................................................... 27

# LIST OF FIGURES

Overview of the Pipeline Structure of the Model........................................................ 13

# LIST OF TABLES

Number of entries for each emotion in the EmotionLines dataset ............................... 10

Number of Reddit posts used from the four subreddits, r/depression, r/happy, r/loseit, and r/bodybuilding........................................................................................ 12

Accuracy on Emotion Classification on Friends Dialogues........................................ 20

Accuracy on Emotion Classification on Facebook Messenger Dialogues................... 21

Accuracy of SVM and RF on Binary Classification .................................................... 23

Coefficients for each Emotion over Classifiers and Datasets ...................................... 23

ANOVA Statistics on the seven emotions over the whole Dataset.............................. 24

ANOVA Statistics on the seven emotions over the Depression Posts ......................... 24

ANOVA Statistics on the seven emotions over the Control Posts ............................... 24

## 1. Introduction

Depression is a mental disorder from impaired emotion regulation. In Teasdale's study, people with depression differ from others not in the initial reactions to an antagonistic event, but during their recovery processes (Teasdale, 1988). The lack of properly regulating emotion(s) has been an indicator of people who are depressed as well as those who have recovered from depression; both groups have delineated to suffer from appropriately managing negative emotions (Ehring et al., 2008). Such findings raise the possibility that those who experienced depression might have a constant difficulty coping with negative events, ultimately leading to chronic or recurring depression.

Depression is a prevalent yet serious public health issue that influences more than 17 million American adults (National Institute of Mental Health, 2019). It is also believed to be a cause for more than 44,000 suicides annually (Xu et al. 2018; Luoma et al., 2002). Depending on the regions, fewer than half or even 10% of those affected with depression seek for/receive proper treatments. In the United States, depression is recognized as a major cause of disability for people of ages between 15 and 44 according to statistics from the National Institute of Mental Health in 2016 (Anxiety and Depression Association of America). Depression is polymorphic, which is categorized based on traits such as the length of the symptom or circumstance(s) impacting the disorder. The most common form, major depressive disorder, plagued approximately 16.1 million adults in the U.S. in 2015 (Anxiety and Depression Association of America). Those who have a depressed mood that lasts at least two years are categorized under persistent depressive disorder while many women experience this severe form of depression during or after pregnancy. Furthermore, depression is sometimes accompanied with delusions or hallucinations, which are labeled as psychotic depression (National Institute of Mental Health, 2018).

Common measurements to treat depression are very responsive, with more than 80% returning positive results after treatment. Ranging from medication, psychotherapy, to electroconvulsive therapy, multiple methods are utilized to treat the disorder. (American Psychiatric Association, 2017). However, such treatments are based on the premise that patients/potential patients are actively asking for help from medical professionals; less than half of those who suffer from this highly treatable disease seek for help. The premise of the treatment stands as a barrier against people; the social stigma or the lack of professional health-care providers can often prevent someone from receiving treatment let alone realizing that he or she has depression (The World Health Organization, 2018).

The high percentage of those who are not treated for a highly recoverable illness raises the importance of building a sophisticated system that would identify those with depression. Surveys and clinical questionnaires, while providing a more thorough examination of a given subject, has the restriction that the subject has to be physically present. Thus, it is important to utilize a platform which the general public uses frequently and does not require a deviation from their natural behavior for the sake of identifying depression.

Online communities stand out as a suitable corpus as one of the distinct properties of social media is that people “broadcast” or “post” their opinions and other people can respond to the given post in the form of comments. The range of opinions that are posted are limitless; from political perspectives to hobbies and personal life stories, people have been treating social media platforms as a public, interactive diary.

Such unique characteristic enables the studying of the posts more carefully in detail, looking for patterns and indicators of the user’s mental state; more specifically, depression. The platform contains a large amount of data about each person’s thoughts and feelings, thus providing us with ample information on the current status of the person. Furthermore, these online communities provide a solution to the problem of social stigma associated with depression, which could often discourage individuals from seeking help from medical professionals (Corrigan, 2004).

There has been extensive research on analyzing online posts to assess the mental states of users. Many of the studies focus on the online behavior of the user, ranging from the normal time the user writes posts to the subcommunities that the user is associated with. Such methods, while effective, is based on the premise that the user has a consistent history on the particular social media website; a new user with a depression might go under the radar.

In this paper we address the challenge of addressing the mental state of users regardless of the lifespan of their activity in social media. To achieve this goal, we use machine learning techniques to first construct an emotion classifier through television dialogues and Facebook messages, and then create emotion annotations for Reddit posts to train a binary classifier to differentiate between depression and non-depression subreddits.

Via experimental studies, the influence of pre-trained word vectors on the task of emotional analysis was compared as well as the performance of different models.

Using the model with the highest accuracy, successful differentiation was achieved on the posts from the depression subreddit and those from the non-depression subreddits. Statistical tests corroborate the strong relationship between emotion embeddings and binary predictions.

The remainder of this study is organized as follows. Section 2 introduces the related work and background on depression, social media, emotion analysis, and machine learning techniques on text classification. Section 3 talks about the dataset used in this study while section 4 elaborates on the methodologies of the experiments. Section 5 further explains the setup of the experiments along with the associated hypotheses. Sections 6 and 7 discuss the results of the experiments and the future work that could be explored. Lastly, section 8 lists the related papers that were referenced for this study.

## 2. Related Work

The four keywords that are crucial in this study are as the following: depression in social media, detection of depression in social media, emotion analysis, and text classification.  Past research has demonstrated the suitability of social media as a platform to detect depression. Despite the extensive research on depression detection in social media, the majority focuses on the activity patterns, which may not be suitable for users with inconsistent activities. As an approach to potentially resolve this challenge, emotional analysis is used via sentence classification.

### *2.1 Depression in Social Media*

Due to the enhancement of creativity via sharing and expanding online connections, social media provides the grounds for many people to find and join communities in which possess common interests amongst the members. Such connections are arduous to develop and cannot be maintained without the help of such online communities; the rapidity of the internet further enables for a quick and responsive conversation among individuals (O'Keeffe et al., 2011). However, social media also entails a great deal of risk such as cyber-bullying and verbal sexual harassment. Moreover, those who spend a great deal of time on social media platforms are prone to developing classic symptoms of depression (Davila et al., 2009; Selfhout et al., 2009). Approximately 70% of the adults in the U.S. are using Facebook and 81% of those are between 18 and 29, the risk of depression through online communities is a grave issue that needs considerable attention (PEW Research Center, 2018).

Studying social media as a tool to assess mental and behavioral health has advantages in the sense that online activities are less prone to subjectivity from self-report methodology in behavioral surveys (De Choudhury et al., 2013). The variety in the age, gender, income, education, and etc. among the users provide us with ample data in analyzing methodologies to successfully detect depression. In addition, as more and more people use social media as the main medium of interactive communication with others, online communities such as Facebook or Reddit can serve as platforms to capture the naturalistic language behavior that is difficult to record offline.

De Choudhury comprehensively studied the potential of analyzing social media as a means to detect depression (De Choudhury et al., 2013). Using the Center for

Epidemiologic Studies Depression Scale questionnaire as the measurement, the study built a ground truth dataset using Amazon Mechanical Turkers. Then, a predictive model that determines whether a post is indicates any sign(s) of depression, was constructed using daily posts from Twitter users. This model estimates the degree of depression in large scale populations. Regardless of the main corpus of the study being Twitter, which is used by less than 10% of the 74% of young U.S. adults, it is clearly shown that social media can serve as a powerful domain for examining the mental health of users while addressing the challenge of detecting under-reported mental health issues.

### *2.2 Depression Detection in Social Media*

Various methodologies have been applied to resolve the fallacies in detecting depression via social media. In Wang et al.'s study, psychologists examined the behaviors of depressed users in online communities and identified distinctive features such as the increased use of first-person singular pronouns, decreased use of emoticons, and frequent original posts from the user between midnight and 6:00am (Wang et al., 2013). De Choudury et al. further analyzed the behavioral aspects of users in the online communities. When assessing depressive behavior, engagement (the number of posts, replies in addition to the time of the user activity), egocentric social graph (the undirected interactive graph between users), emotion (based on the psycholinguistic resource LIWC, http://www.liwc.net), linguistic style (based on LIWC), and depression language (based on the topical language of users with depression) were the key elements involved (De Choudhury et al., 2013). It is important to note that the criteria used in the emotion feature of the study were positive affect, negative affect, activation, and dominance. Although the categories have been shown to perform well when predicting future behaviors and moods in social networks (De Choudhury et al., 2013), the categories lack detail in the sense that only positivity and negativity are illustrated. Moreover, such prediction(s) in future behaviors and moods might not be suitable for those who have just recently joined the network; newcomers will lack the behavioral aspect which are the main features used for predictions, nullifying the accuracy of the approach who constantly change their social media legions or have multiple accounts, as the linking of numerous accounts to a single entity introduces a whole new challenge. To address this domain of users, this study utilizes emotion analysis on posts to determine the authors' mental states.

### *2.3 Emotional Analysis*

For the past two decades, numerous studies have been conducted on emotional analysis. For instance, Koelstra et al. used electroencephalogram (EEG) and peripheral physiological signals of the participants to solve binary classification problems of arousal, valence and liking (Koelstra et al., 2012). In Quan et al.'s study, weblogs were used to construct an emotional expression model (Quan et al., 2009). Mishne also studied the problem of mood classification in Livejournal, another type of online community where people write in the form of blog posts (Mishne, 2005). Mihaclea and Liu studied LiveJournal as well and examined the problem of identifying happiness within a corpus of annotated happy/sad posts, analyzing different aspects such as the time of the post or semantic dimensions (Mihalcea et al., 2006).

Despite such extensive research on identifying emotions online as forementioned, there has been a discrepancy in the number of categories for the emotions used in the study. For example, Quan et al. used eight emotion classes (expect, joy, love, surprise, anxiety, sorry, angry, and hate) to classify the Chinese emotional expression categories whereas Jung et al. utilized only the four mood categories (happy, sad, angry, and fear) based on the posts extracted from LiveJournal.(Jung et al., 2006) In a previous study on the history of defining emotions, Gendron explores other previous work to define the term "emotion" in a scientific setting. Through a review of Izard's survey, Gendron was unable to identify the scientific criteria behind differentiating one emotion from another. (Gendron, 2010; Izard, 2010). Reaching a consensus on the exact number of emotions to be considered in a scientific setting is an underlying future work for this study; however, it is substantial to emphasize that the study is focused on the effectiveness of emotional analysis on detecting depression and not on the construction of a universally accepted category of emotions.

Emotional analysis on the online corpus share a common task of text classification; the task of emotion classification constructs a system that enables emotion extraction of posts, leading to a binary classifier that will detect elements that signal the presence of depression.

### *2.4 Sentence Classification*

Within the popular challenge of text classification, there have been relatively fewer studies that specifically scrutinized the methodology on classifying sentences. In Yoon Kim's study, the use of convolutional neural network (CNN) to resolve this issue was proposed (Kim, 2014). Kim's proposed model performed very competitively against numerous deep learning models such as the Dynamic Convolutional Neural Network with k-max pooling (Kalchbrenner et al., 2014) and Matrix-Vector Recursive Neural Network with parse trees (Socher et al., 2012). The series of studies in Kim's work delineated that CNN performed greatly with pre-trained word embeddings, illustrating how pre-trained vectors can be utilized in various tasks and datasets.

Sentence classification provides a more refined analysis as it can be used to study both the entirety of the document and the individual sentences that serve as building blocks. This further allows the establishment of emotion embeddings in each sentence in a given post, leading to a more detailed analysis on the correlation between the emotions at the start, middle, end of the post and the subreddit that it is associated with.

## **3. Dataset**

The datasets for this study originate from two sources: EmotionLines dataset from the EmotionX Challenge of SocialNLP 2018 (SocialNLP EmotionX Challenge, 2018) and a subset of a Google BigQuery dataset on Reddit. Public datasets that have emotion annotation on social media posts are not available due to the protection of privacy. Such a challenge has led to select a dataset that is constructed in an online conversation with emotional annotations. In addition, the intrinsic structure of Reddit allows the construction of an experimental group and a control group based on topical similarity to depression. The section further elaborates on how the datasets were constructed as well as the distribution(s) of labels within each dataset.

### *3.1 EmotionLines*

EmotionLines dataset was used for the EmotionX Challenge in SocialNLP 2018, in which the participants analyzed the corpus from Friends TV scripts and Facebook messenger dialogues. The dataset was annotated according to the six emotions mentioned in Ekman’s study (Ekman et al., 1987) and a seventh emotion *neutral* and eighth emotion *non-neutral* (to the entries labeled with more than one emotion) using Amazon Mechanical Turk. The Facebook messenger dialogues were pre-processed to remove all critical personal information that might expose the users. Table 1 represents the label distribution of the dataset.

**Table 1.** Number of entries for each emotion in the EmotionLines dataset - the Friends dialogue dataset is divided into Friends Train and Friends Test and the Facebook dialogue is divided into Facebook Train and Facebook Test

| Emotion | Friends Train | Friends Test | Facebook Train | Facebook Test |
|---|---|---|---|---|
| Joy | 1283 | 304 | 1482 | 458 |
| Anger | 513 | 161 | 94 | 37 |
| Sadness | 351 | 85 | 389 | 87 |
| Fear | 185 | 32 | 36 | 2 |
| Surprise | 1220 | 286 | 435 | 93 |
| Disgust | 240 | 68 | 85 | 15 |
| Neutral | 4752 | 1287 | 7148 | 1882 |
| Non-Neutral | 2017 | 541 | 1064 | 233 |

### *3.2 Reddit Dataset (RD)*

The dataset used for binary classification was collected through Google BigQuery on Reddit. In this work, posts from December 2015 to December 2018 were used, which were subsequently divided into two subsets: depression dataset and negative control group. The depression dataset is consisted of posts from the subreddit *r/depression*. For the control group, the same number of posts were selected from three different subsets *r/happy*, *r/loseit*, and *r/bodybuilding*, identical to the ones from Park et al.'s study on Reddit and the written-communication challenges those with mental disorders face on the online communities (Park et al., 2018). The *r/happy* subreddit is aimed to share positive stories; *r/loseit* is a place where people share concerns about their weight and discuss healthy ways to lose weight; *r/bodybuilding* is a subreddit devoted to those who are interested in bodybuilding ranging from nutrition to training methods and preparing for various contests. Park's study further elaborates the selection criteria for the control group subreddits: *r/happy* was selected for its high activity as well its focus on positive emotion; *r/loseit* was selected for the relatively low possibility of medical or technical terminology while exhibiting an abundance of emotional support (Cunha et al., 2016); *r/bodybuilding* was chosen for the notable amount of emotionally supportive posts (Ploderer et al., 2008). All subreddits used in the study were selected due to their rich emotional content, enabling them as a suitable experimental dataset.

The title of the post was not included in the dataset; any posts that were either removed or deleted, which replaces the body of the post with an empty string or the token "removed", were also excluded from the dataset. Furthermore, any posts that were shorter than 4 sentences were excluded from the dataset to effectively compare among the first quarter, middle half, last quarter, and the total post for the binary classification task. Table 2 represents the total number of posts from *r/depression* and the control group before and after the exclusion described above.

**Table 2.** Number of Reddit posts used from the four subreddits, r/depression, r/happy, r/loseit, and r/bodybuilding

| | Total pool of posts | Posts with length > 4 |
|---|---|---|
| r/depression | 445534 | 219180 |
| r/happy, r/loseit, r/bodybuilding | 238553 | 102388 |

## 4. Methodology

In this study a two-part pipeline structure was constructed. The former established an emotion classifier which created emotion embeddings that were used in the latter, which created a model to classify Reddit posts related to depression. Figure 1 illustrates the proposed pipeline model. The system was trained on Friends dialogues and Facebook messages mentioned above with pre-trained word vectors and produced emotion embeddings for sentences. The classifier generated emotion embeddings for each sentence of a reddit post, and the binary classifier that detected reddit posts from *r/depression* was trained via emotion embeddings. Permutation tests and the analysis of variance tests (ANOVA) were performed to further explore the relationship between each emotion and its performance towards binary classification.

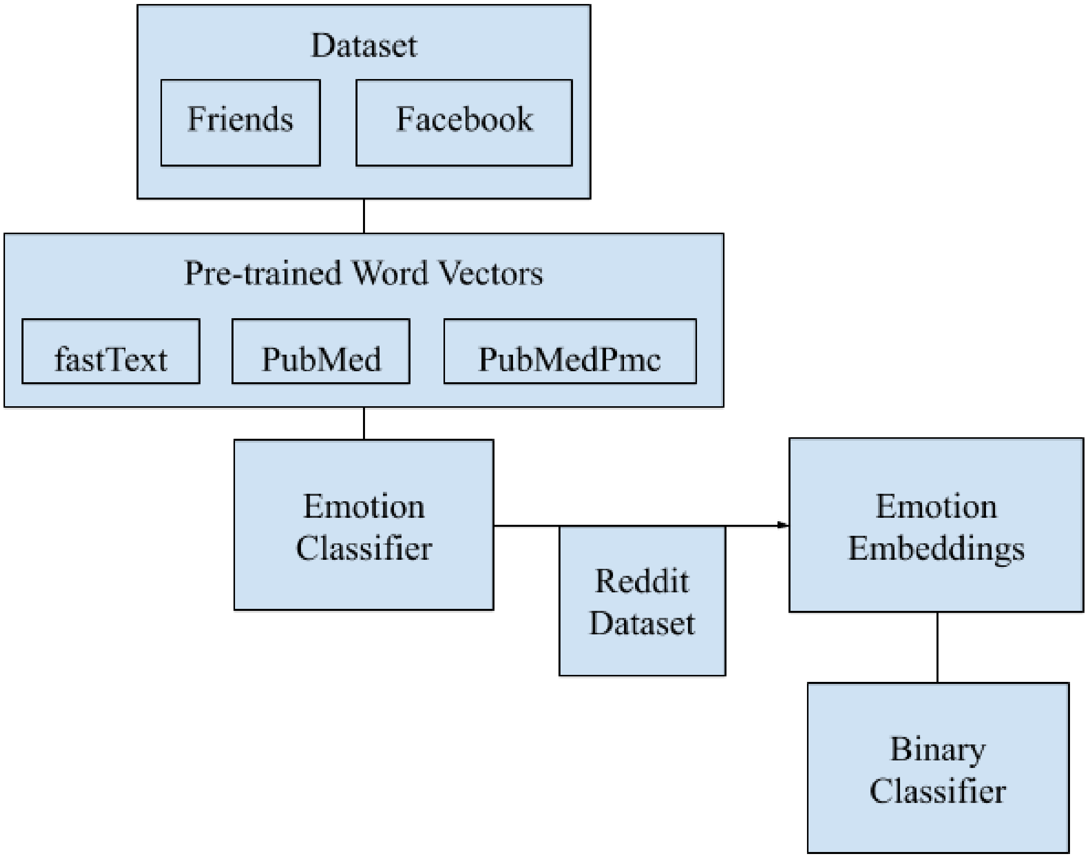

**Figure 1.** Overview of the Pipeline Structure of the Model

*4.1 Models*

The main model for this study was based on Kim's CNN model for sentence classification. For baseline models, support vector machines (Joachims, 1998) and random forest classifiers (Liaw and Wiener, 2002) were used. The study compared the performances of the baseline models and the CNN with pre-trained word vectors. The model and the pre-trained vector that yielded the highest accuracy were chosen to create emotion embeddings for the binary classification task. For the binary classification task, the two base models for the first part of the experiment were used.

The coefficients of the binary classifier were examined to further explore the influence of each emotion on the outcome of the classification process.

### *4.2 Emotion Embeddings*

The first part of the study established an emotion classifier that created emotion embeddings of a sentence. The performances of the models were compared across four different emotion label settings: all eight emotions, without neutral and non-neutral, without neutral, and without non-neutral. The comparisons assessed the influence of neutral and non-neutral emotion annotations on constructing an accurate emotional description of online texts. For the binary classification task, positional analysis of the emotion embeddings was utilized. For a given Reddit post, an emotion embedding for each sentence in the post was first constructed, which then was averaged across the first quarter, the middle, the last quarter, and the whole post. This allowed for a detailed analysis on which part of the post contains the most information that would help detect depression; while a post might have a “general” emotion that dominates the post overall, there may be parts of the post where the user shows different emotions. For example, it could be the case that many of the users who write on the depression subreddit start their posts with an introduction, which leads to their body of the post where they elaborate in detail about why they feel so depressed and how they are feeling. Similarly, posts might have a conclusion stage where the author wraps up the story.

### *4.3 Word Embeddings*

Three pre-trained word vectors were mainly used for this study: fastText word vectors (Mikolov et al., 2017), PubMed vectors, and PubMedPmc vectors (Moen et al., 2013). The fastText vectors were trained on Wikipedia, UMBC WebBase corpus, and statmt.org news while the PubMed, PubMedPmc vectors were trained on PubMed and PubMed Central texts using the word2vec tool, respectively. While fastText word vectors were based on the words used in the average daily settings, the PubMed vectors were more accustomed to the medical terminology. There was a notable difference in the corpus used to train these vectors: fastText utilized the web encyclopedia dataset along with various news articles whereas PubMed vectors trained on the scholarly articles related to medical research. While fastText word vectors would intuitively be more suitable for the task of analyzing social media compared to PubMed vectors, the task of detecting a mental health issue raised the importance of studying the influence of word embeddings on this task.

### *4.4 Statistical Methods*

To further explore the differences on the emotion embeddings among datasets, this study utilized a permutation test on each emotion over all datasets to reject the null hypothesis which states that permuting the given emotion has no significant effect on the performance of the binary classifier (Fisher, 1937). If the original observed accuracy was lower than the permuted accuracy for more than 5% of the permutations, it would suggest that the certain emotion has negligible effect on the output of the binary classifier. Additionally, the analysis of variance (ANOVA) test (Girden, 1992) was performed to study the differences in each emotion within the three datasets; first quarter, last quarter, and middle half to examine the dynamics of emotions throughout a given post.

## 5. Experiments

The work is mainly divided into two sections: first, constructing an emotion classifier and using the emotion embeddings from the classifier to train a binary classifier to differentiate between depression and non-depression subreddits. While the former experiment compared the performance of the classifier across CNN, linear SVM, and random forest model and three different pre-trained word embeddings - fastText, PubMed, and PubMedPmc - on dialogues from Friends and Facebook messages, the latter experiment used the best emotion classifier from the first experiment to create emotion embeddings for each sentence in a given post and train binary classifiers. The two base models were trained across four different proportions of the post to compare the positional influence of the emotion; the first quarter, middle half, the last quarter, and the total post. The section elaborates on each experiment along with the hypotheses tested.

### *5.1 Emotion Classifier*

This experiment was mainly aimed to compare the performance of different models across the two datasets while also examining the influence of word vectors and the data construction on the performance of the classifier. For this experiment, the following hypotheses were tested:

h1. The performance of the classifiers on the Facebook Messenger dialogues will be higher than that of the classifiers on the Friends dialogues.

h2. Models that exclude the neutral and non-neutral utterances will perform better than those that include all emotions or exclude only one of neutral and non-neutral.

h3. Excluding only the non-neutral utterances will have a positive effect on the performance.

h4. Excluding only the neutral utterance will have a positive effect on the performance.

h5. The PubMed vectors will have the best performance among the three different word embeddings.

Friends, which was one of the main sources for the EmotionLines dataset, was a television sitcom that starred numerous actors and actresses. Each member of the sitcom used numerous ways to convey messages and emotions, such as voice tones, body language, and facial expressions. The dialogues, which is one channel out of several for the sitcom, would be a fragment of the data, which would be inadequate for training the model in detecting emotions.

In contrast, the dialogues from the Facebook messenger have an intrinsic limitation in the number of ways to communicate. Other than using emoticons or emojis, text is the main and only channel of conversation between two people. Therefore, it was hypothesized that the lack of different ways to express emotions would motivate people to express their emotions more through text.

Another prediction was that excluding the neutral labels and non-neutral labels, which represent the absence of emotion and combination of emotions, respectively, would have negative effects on the classifier. Training the classifiers without such entries would result in a higher accuracy.

### *5.2 Binary Classifier*

After training different models with different word vectors for the emotion classifier, the model with the highest accuracy, random forest, was used to establish emotion embeddings for each post. A post was divided into sentences, which was given to the emotion classifier to generate emotion embeddings. The sentence emotion embeddings of the post were then divided into four different sets: first quarter, middle half, last quarter, and total. The first and last quarter represented the first and last 25% of the post while the middle half represented the remaining 50%. The embeddings were averaged to get a single vector that represents the first quarter, middle half, last quarter, and total of the post. The four datasets were utilized to train the random forest and linear SVM classifiers, which performed binary classification with an 80:20 train-test ratio. For this experiment, the following hypotheses were tested:

h1. There will be a significant difference in the neutral emotion between the three datasets: first quarter, last quarter, and middle half.

h2. The model trained on the middle half will perform better than those trained on the first and last quarter.

h3. The model trained on whole posts will perform better than those trained on the other three datasets.

h4. Joy, anger, sadness, and disgust will be the most contributing emotions in the binary classification.

A person's overall emotion might not be completely consistent for a given post; the user could start the post with an "opening" which leads to the body of the post that would represent the dominant emotion property of the whole post. The last quarter of the post could depict the user's closing comments, which could also suggest that the user would "tone down" and incorporate more sentences with the neutral emotions. There would be a significant difference in the amount of neutral emotion one shows between the first quarter, middle and the last quarter. Examining the difference in the absence of emotion throughout a given post could further help understand the dynamic of emotion expression in social media.

## 6. Results

The experiment to establish an emotion classifier showed that model trained on the word vectors trained on a corpus not related to medical articles had a higher accuracy. Training on Facebook messages yielded a higher accuracy over the Friends dialogues, supporting the possibility of a correlation between the limitation of communication medium and the amount of emotion conveyed through text. Furthermore, excluding the "combined" emotion labels while keeping the neutral emotion labels returned the highest performance. The experiment on binary classification returned promising results with miniscule difference between the performance between the linear SVM and random forest classifier. Between the four different parts of the post analyzing the total post proved to be the most effective, followed by the middle half and the first, last quarters.

### *6.1 Emotion Classifier*

Table 3 depicts the performance of each model on all eight emotions. It is worth noting that for the linear SVM, excluding the neutral and non-neutral emotions had a negative effect on its accuracy. For both baseline models, exclusion of non-neutral improved the overall performance.

**Table 3.** Accuracy on Emotion Classification on Friends Dialogues - three word vectors (fastText, PubMed, PubMedPmc) were tested for four different scenarios

| | All eight emotions | | | excluding neutral, non-neutral | | | excluding neutral | | | excluding non-neutral | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| WV | fast Text | Pub Med | Pub Med Pmc | fast Text | Pub Med | Pub Med Pmc | fast Text | Pub Med | Pub Med Pmc | fast Text | Pub Med | Pub Med Pmc |
| SVM | 47.21 | 48.95 | 49.57 | 35.58 | 41.67 | 43.27 | 37.64 | 40.42 | 41.30 | 57.94 | 59.51 | 59.56 |
| RF | 50.90 | 49.06 | 48.91 | 46.15 | 47.65 | 46.37 | 40.89 | 39.74 | 39.68 | 63.88 | 62.71 | 62.98 |
| CNN | 53.05 | 50.45 | 50.57 | 58.35 | 54.00 | 52.15 | 43.79 | 43.11 | 40.22 | 66.97 | 63.25 | 62.88 |

**Table 4.** Accuracy on Emotion Classification on Facebook Messenger Dialogues - three word vectors (fastText, PubMed, PubMedPmc) were tested for four different scenarios

| | All eight emotions | | | excluding neutral, non-neutral | | | excluding neutral | | | excluding non-neutral | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| WV | fast Text | Pub Med | Pub Med Pmc | fast Text | Pub Med | Pub Med Pmc | fast Text | Pub Med | Pub Med Pmc | fast Text | Pub Med | Pub Med Pmc |
| SVM | 67.05 | 67.05 | 67.05 | 66.18 | 66.18 | 66.18 | 51.78 | 52.22 | 51.24 | **73.12** | 73.12 | 73.12 |
| RF | 70.75 | 70.04 | 69.86 | 69.65 | 69.51 | 67.63 | 55.14 | 51.78 | 52.97 | **78.01** | 75.95 | 77.23 |
| CNN | 49.95 | 50.78 | 50.53 | 58.52 | 55.54 | 53.90 | 44.32 | 43.71 | 39.96 | 67.08 | 64.70 | 64.76 |

Table 4 displays the performance of the model on Facebook dialogues. Most of the models performed similarly or better than those with the Friends dataset, which suggest limitations within communication methods may lead to displaying various emotions through texts. The initial hypothesis of a performance gain by excluding neutral and non-neutral emotion labels did not seem to be uniform across the models. Excluding neutral labels led to a drop in the accuracies; in contrast, excluding the non-neutral labels performed better across the models. This indicates that the existence of neutral labels for the task of emotion classification is crucial; the absence of one's emotion provides ample information when training for the existence of a certain emotion. In addition, the neutral labels could further aid in understanding the relationships between different emotions by providing the "zero point." Establishing word embeddings from medical articles did not provide more gain to the accuracy;

fastText word vectors had more power, further showing how the vectors from general articles are more appropriate for social media.

### *6.2 Binary Classification through Emotion Embeddings*

Table 5 delineates the scores of the binary classifiers on four different datasets. It is noteworthy to observe that the accuracies for the total posts and the middle 50% of the post were higher than those for the first and last quarters. These results support the hypothesis that the body of a post is more effective in emotion embeddings than other counterparts of a post. For the coefficients from the SVM, the squares of the coefficients were ranked based on Guyon et al.’s study, which serve as a substantial criterion when ranking individual features (Guyon et al., 2002). The reasoning was further corroborated by examining Optimum Brain Damage (OBD) algorithm which estimated the cost function of the linear SVM through the expansion in Taylor series to the second order (LeCun et al., 1990). Furthermore, Guyon explained how Lecun et al.’s study suggested using the square of the coefficients over the absolute value when ranking features. As shown in Table 6, joy, anger, disgust, and neutral were the most influential emotions when training the classifier. It was remarkable to find that neutral, which is recognized as an as the absence of a specific emotion, achieved comparable training of the classifiers as much as the “strong” emotions such as anger and disgust. Also, different emotions dictated the differentiation two groups of posts for the models examined; while anger and joy were the dominant emotions for linear SVMs, random forest classifiers focused on joy, sadness, and disgust.

Each emotion went through 100 permutations and statistic comparison determined the accuracy of the model for subject set(s). All seven emotions had no case where the permuted statistic was larger than the original accuracy. Therefore, the null hypothesis was rejected for all emotions.

**Table 5.** Accuracy of SVM and RF on Binary Classification

| | First 25% | Middle 50% | Last 25% | Total |
|---|---|---|---|---|
| Linear SVM | 0.78661 | **0.86393** | 0.77573 | **0.92020** |
| Random Forest | 0.79505 | **0.87595** | 0.78621 | **0.92496** |

**Table 6.** Coefficients for each Emotion over Classifiers and Datasets

| | Joy | Anger | Sadness | Fear | Surprise | Disgust | Neutral |
|---|---|---|---|---|---|---|---|
| SVM-First | **13.20522** | **41.32698** | 4.05970 | 2.40859 | 0.93813 | 4.40013 | 7.39003 |
| SVM-Middle | **36.30318** | **70.42947** | 6.35496 | 3.72187 | 3.99750 | **23.46421** | **17.78865** |
| SVM-Last | **11.35683** | **53.83883** | 3.41504 | 3.07201 | 0.09861 | 0.86653 | 6.91133 |
| SVM-Total | **66.19978** | **88.66724** | 6.95809 | 3.37623 | 14.44245 | **46.82331** | **28.34717** |
| RF-First | **0.50930** | 0.00619 | **0.12685** | 0.00027 | 0.02448 | **0.21200** | 0.12091 |
| RF-Middle | **0.46582** | 0.01181 | **0.16241** | 0.00022 | 0.02764 | **0.24918** | 0.08293 |
| RF-Last | **0.43970** | 0.00921 | **0.19847** | 0.00049 | 0.05680 | **0.16939** | 0.12593 |
| RF-Total | **0.38451** | 0.01307 | **0.19362** | 0.00045 | 0.05844 | **0.24969** | 0.10022 |

**Table 7.** ANOVA Statistics on the seven emotions over the whole Dataset

| | Joy | Anger | Sadness | Fear | Surprise | Disgust | Neutral |
|---|---|---|---|---|---|---|---|
| Statistic | 168.60888 | 4.07045 | 55.99323 | 1.13655 | 499.27991 | 150.47397 | 112.50172 |
| P-Value | 6.16717 e-74 | **0.01707** | 4.83294 e-25 | **0.32093** | 2.02138 e-217 | 4.59962 e-66 | 1.40684 e-49 |

**Table 8.** ANOVA Statistics on the seven emotions over the Depression Posts

| | Joy | Anger | Sadness | Fear | Surprise | Disgust | Neutral |
|---|---|---|---|---|---|---|---|
| Statistic | 529.95899 | 1.06973 | 64.79508 | Nan | 14.68470 | 50.62119 | 404.22793 |
| P-Value | 1.06424 e-230 | **0.34310** | 7.28833 e-29 | Nan | 4.19430 e-07 | 1.04038 e-22 | 3.57993 e-176 |

**Table 9.** ANOVA Statistics on the seven emotions over the Control Posts

| | Joy | Anger | Sadness | Fear | Surprise | Disgust | Neutral |
|---|---|---|---|---|---|---|---|
| Statistic | 644.44922 | 4.91084 | 42.19918 | 1.82411 | 660.22520 | 268.74635 | 410.44947 |
| P-Value | 5.06798 e-280 | 0.00737 | 4.73855 e-19 | **0.16136** | 7.62767 e-287 | 2.43748 e-117 | 9.59031 e-179 |

It is notable to see that most of the emotions seem to reject the null hypothesis that states a significant difference in the average number of emotions in the first quarter, middle half, and the last quarter of a given post. However, there seemed to be an equal average amount of anger throughout the depression posts. This was not the case for the posts in the control group, which could suggest how anger consistently had a high coefficient for the linear SVM model. In combination with the accuracy performance on the four different portions of the posts, the ANOVA statistics help further support that people exhibit different emotions throughout the post while maintaining a main "theme" emotion.

## 7. Discussion and Future Work

This study was based on a single interpretation of the possible human emotions, and there are yet more studies to be conducted to successfully create a universally accepted set of emotion categories for scientific research. The datasets represent a subset of the online communities, thus leaving room for further research on other popular social media platforms. Moreover, there were other features that could have been considered but were rendered uniformly across users in the study. Nonetheless, the results demonstrated that emotional analysis is a potent tool in solving the challenge of accurately detecting depression in social media.

### *7.1 Limitations*

Emotions used for this study were defined upon the six emotions from Ekman et al.'s study along with the addition of the neutral emotion and the non-neutral emotion, which represents *any* combination of emotions. A non-neutral label could be a combination of sadness and anger, joy and surprise, disgust and anger and sadness, and so on. While the emotion labels used in this study were comprehensive, the selection of the subjects was not to claim these are the universal definition of human emotions. There is yet to finalize on exactly how many and what categories of emotions a human being can express. As previously mentioned, the goal of this study is aimed to scrutinize the potential of emotion analysis based on one interpretation of the human emotion rather than defining set(s) of emotions.

The main corpus used for training the emotion classifier was mainly from dialogues in a real-world setting and a television series. The utterances within the corpus possess different traits from those commonly observed in posts written in online communities; one was focused on a stream of communication that goes back and forth in a many-to-many fashion while the other was similar to a broadcast in the sense that one writes in a one-to-many format.

Binary classification task was based on Reddit, a platform which has subreddits each with a specific purpose and rules made by associated moderators. This unique system of reddit is very different from other popular social media platforms such as Facebook or Twitter. Reddit users always write posts that are part of a subreddit(s), suggesting that the post would most likely abide the rules made for the certain subreddit. Facebook or Twitter, on the other hand, provide users each with his/her own page where one can post without any affiliation to a community within the platform. Users, when writing within a subgroup that has a set topic, may go under a "learning phase," in which they start to align their linguistic styles with those of the respective communities; depending on the life cycle of their activities within the group, users will go through a "conservative phase" halting them from conforming to the norm, which widens the discrepancy between their and the group's language (Danescu-Niculescu-Mizil et al., 2013). This study assumed that all users are in the same "phase" when writing in the subreddit.

### *7.2 Conclusion and Future Work*

The objective of this study is to identify and compare different models for establishing an effective emotion classifier. The pre-trained word embeddings on Wikipedia were more useful than medical articles when training for emotions within a given sentence. Compared to the dialogues from Friends, the conversations in Facebook messengers led to a better performance of the examined models. This highlighted the potential of social media as a platform to detect human emotions and analyze the mental state of individuals. The lack of vocal tones and facial expressions when communicating through online communities resulted in a more emotion-rich text, which resulted in capturing the natural mental state of the user. The baseline models demonstrated comparable or better performance than the CNN.

Emotion embeddings from the random forest classifier trained on pre-trained fastText vectors without non-neutral labels showed to be an excellent representation to identify reddit posts about depression.  The proposed model was able to differentiate the posts from *r/depression* and those from *r/happy*, *r/loseit*, and *r/bodybuilding* with a high accuracy. This further suggests the potential of emotion embeddings in analyzing a user's mental state in online communities. Compared to other methodologies of examining the behavior of the user such as the posting activity, social embeddings, and so on, emotion embeddings proved to be a successful approach that would also address users without data on previous activities.

A possible future work expanding on this study could explore the use of emotion embeddings in differentiating posts from "similar" subreddits such as *r/depression*, *r/SuicideWatch*, and *r/anxiety*. The study would further examine the strength of emotion analysis on mental health, which would look into whether emotion analysis could help identify those with immediate suicidal intent from those who are experiencing a mild form of temporary depression. Another possible study is applying emotion analysis across different social media platforms such as Facebook or Twitter. Through the different platforms the study would expand on the relationship between the properties of each platform and performance of emotion analysis. It would help determine the establishment of efficient systems for detecting mental illnesses in these communities, either by forming a general cross-platform model or multiple models tweaked for maximum efficiency on each website.

# REFERENCES

1. American Psychiatric Association. https://www.psychiatry.org/patients-families/depression/what-is-depression

2. Anxiety and Depression Association of America. https://adaa.org/understanding-anxiety/depression

3. Anxiety and Depression Association of America. https://adaa.org/about-adaa/press-room/facts-statistics

4. Corrigan, Patrick. "How stigma interferes with mental health care." *American psychologist* 59.7 (2004): 614.

5. Cunha, Tiago Oliveira, et al. "The effect of social feedback in a reddit weight loss community." *Proceedings of the 6th International Conference on Digital Health Conference*. ACM, 2016.

6. Danescu-Niculescu-Mizil, Cristian, et al. "No country for old members: User lifecycle and linguistic change in online communities." *Proceedings of the 22nd international conference on World Wide Web*. ACM, 2013.

7. Davila, Joanne, et al. "Romantic and sexual activities, parent–adolescent stress, and depressive symptoms among early adolescent girls." *Journal of adolescence* 32.4 (2009): 909-924.

8. De Choudhury, Munmun, Scott Counts, and Eric Horvitz. "Social media as a measurement tool of depression in populations." *Proceedings of the 5th Annual ACM Web Science Conference*. ACM, 2013.

9. De Choudhury, Munmun, et al. "Predicting depression via social media." *Seventh international AAAI conference on weblogs and social media*. 2013.

10. De Choudhury, Munmun, Scott Counts, and Eric Horvitz. "Predicting postpartum changes in emotion and behavior via social media." *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 2013.

11. Ehring, Thomas, et al. "Characteristics of emotion regulation in recovered depressed versus never depressed individuals." *Personality and Individual Differences* 44.7 (2008): 1574-1584.

12. Ekman, Paul, et al. "Universals and cultural differences in the judgments of facial expressions of emotion." *Journal of personality and social psychology* 53.4 (1987): 712.

13. Fisher, Ronald Aylmer. *The design of experiments*. Oliver And Boyd; Edinburgh; London, 1937.

14. Gendron, Maria. "Defining emotion: A brief history." *Emotion Review* 2.4 (2010): 371-372.

15. Girden, Ellen R. *ANOVA: Repeated measures*. No. 84. Sage, 1992.

16. Guyon, Isabelle, et al. "Gene selection for cancer classification using support vector machines." *Machine learning* 46.1-3 (2002): 389-422.

17. Izard, Carroll E. "The many meanings/aspects of emotion: Definitions, functions, activation, and regulation." *Emotion Review* 2.4 (2010): 363-370.

18. Joachims, Thorsten. "Text categorization with support vector machines: Learning with many relevant features." *European conference on machine learning*. Springer, Berlin, Heidelberg, 1998.

19. Jung, Yuchul, Hogun Park, and Sung Hyon Myaeng. "A hybrid mood classification approach for blog text." *Pacific Rim International Conference on Artificial Intelligence*. Springer, Berlin, Heidelberg, 2006.

20. Kalchbrenner, Nal, Edward Grefenstette, and Phil Blunsom. "A convolutional neural network for modelling sentences." *arXiv preprint arXiv:1404.2188* (2014).

21. Kim, Yoon. "Convolutional neural networks for sentence classification." *arXiv preprint arXiv:1408.5882* (2014).

22. Koelstra, Sander, et al. "Deap: A database for emotion analysis; using physiological signals." *IEEE transactions on affective computing* 3.1 (2012): 18-31.

23. LeCun, Yann, John S. Denker, and Sara A. Solla. "Optimal brain damage." *Advances in neural information processing systems*. 1990.

24. Liaw, Andy, and Matthew Wiener. "Classification and regression by randomForest." *R news* 2.3 (2002): 18-22.

25. Luoma, Jason B., Catherine E. Martin, and Jane L. Pearson. "Contact with mental health and primary care providers before suicide: a review of the evidence." *American Journal of Psychiatry* 159.6 (2002): 909-916.

26. Mihalcea, Rada, and Hugo Liu. "A Corpus-based Approach to Finding Happiness." *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*. 2006.

27. Mikolov, Tomas, et al. "Advances in pre-training distributed word representations." *arXiv preprint arXiv:1712.09405*(2017).

28. Mishne, Gilad. "Experiments with mood classification in blog posts." *Proceedings of ACM SIGIR 2005 workshop on stylistic analysis of text for information access*. Vol. 19. 2005.

29. Moen, S. P. F. G. H., and Tapio Salakoski2 Sophia Ananiadou. "Distributional semantics resources for biomedical text processing." *Proceedings of LBM* (2013): 39-44.

30. National Institute of Mental Health. https://www.nimh.nih.gov/health/topics/depression/index.shtml

31. National Institute of Mental Health. https://www.nimh.nih.gov/health/statistics/major-depression.shtml

32. O'Keeffe, Gwenn Schurgin, and Kathleen Clarke-Pearson. "The impact of social media on children, adolescents, and families." *Pediatrics* 127.4 (2011): 800-804

33. Park, Albert, and Mike Conway. "Harnessing Reddit to Understand the Written-Communication Challenges Experienced by Individuals With Mental Health Disorders: Analysis of Texts From Mental Health Communities." *Journal of medical Internet research* 20.4 (2018).

34. PEW Research Center. http://www.pewinternet.org/fact-sheet/social-media/

35. Ploderer, Bernd, Steve Howard, and Peter Thomas. "Being online, living offline: the influence of social ties over the appropriation of social network sites." *Proceedings of the 2008 ACM conference on Computer supported cooperative work*. ACM, 2008.

36. Quan, Changqin, and Fuji Ren. "Construction of a blog emotion corpus for Chinese emotional expression analysis." *Proceedings of the 2009 Conference on*

*Empirical Methods in Natural Language Processing: Volume 3-Volume 3*. Association for Computational Linguistics, 2009.

37. Selfhout, Maarten HW, et al. "Different types of Internet use, depression, and social anxiety: The role of perceived friendship quality." *Journal of adolescence* 32.4 (2009): 819-833.

38. SocialNLP 2018 EmotionX Challenge. http://doraemon.iis.sinica.edu.tw/emotionlines/index.html

39. Socher, Richard, et al. "Semantic compositionality through recursive matrix-vector spaces." *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*. Association for Computational Linguistics, 2012.

40. Teasdale, John D. "Cognitive vulnerability to persistent depression." *Cognition & Emotion* 2.3 (1988): 247-274.

41. The World Health Organization. https://www.who.int/news-room/fact-sheets/detail/depression

42. Wang, Xinyu, et al. "A depression detection model based on sentiment analysis in micro-blog social network." *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, Berlin, Heidelberg, 2013.

43. Xu, Jiaquan, et al. "Deaths: Final data for 2016." (2018).