Depression Detection in Tweets

Abstract

Basing off the sentence, "you are what you post", there has been continuous study conducted to identify depression by leveraging social media data. Depression detection traditionally is done through a face-to-face meeting with a clinical professional. However, only 52% of patients come in seeking help (Boerema AM, 2016) early enough for effective intervention. Nowadays, social media is a place in which people post their feelings more often then talking to a clinician about their emotions. Therefore, many research has been conducted to detect depressive symptoms in social media data. In this paper, we are doing three different task with some novel approaches. First, applying statistical analysis to find different word usages between depressed and non-deppressed users. Second, training a classification model for detecting depressed users by utilizing a different library. In addition, we use topic modeling to yield insight into how how depressed and non-depressed individuals use language differently on social media.

1 Introduction

Depression is an illness that affects more than 300 million people worldwide. Major depression can result in impairments that interfere with and limit an individual's daily activities such as family and personal relationships, work, school life. It is believed to be responsible for close to 800,000 suicides every year, as suicide is the second leading cause of death in 15-59 years old people. (WHO, 2018). Early detection and intervention in depression can dramatically speed up the rehabilitation process and minimize the emotional and financial burden of the patient and the people around him/her (Halfin, 2007).

However, current methods to diagnose and treat mental illnesses such as depression are sometimes considered insufficient as they are based on behavior questionnaires and surveys reported by the patients themselves alongside a simple mental health status examination (Nadeem, 2016). These questionnaires tend to be costly, time-consuming and have the potential to be easily manipulated to mask one's depression due to social stigma and denial. (Halfin, 2007). However, among those who were treated for depression, 80% of patients showed improvement in their symptoms within just four to six weeks after their diagnosis (Mann et al., 2005). Therefore, it can be inferred that when depression is correctly identified in the early stages, it can lead to significant changes in the lives of the patients and of the people around them.

In this day and age of the internet, a lot of people tend to share their thoughts and emotions on social media through tweets, posts, pictures, messages, etc. Now, social media is an integral part of many people's daily lives and like any other aspect of life, it can answer questions that can make the world a better place to live. The question at hand is not to judge the benefits of social media but to utilize the hundreds and millions of people generated data. One such field that can benefit from this data is clinical psychology. As people's postings on platforms such as Twitter or Facebook can represent a day-to-day social interaction in a natural setting, it has the benefits of being a real-time, no memory loss, truthful data collection of the individuals of interest. Thus, we believe that social media postings can be used to construct a model that identifies and predicts depression of the user.

The main research question of this work is to classify the different ways that people express their depression through social media postings. It will be based on the previous work done on detecting depression or other mental illness based on Tweets of individuals. Also this paper analyzes the ways in which the topics of tweets differ among

users who are positively and negatively labeled.

2 Previous Work

As interactions on social media become increasingly prominent in the lives of people, automated analysis of social media has shown the potential to provide early detection of depression. Naturally, many researchers conducted studies that examined the correlation between mental illness with social media postings. One of the early works in this field was using publicly-available text data such as the tweeter API to analyze the thoughts, emotions of the users. Work done by Saif Mohammad et al. is notable for work on detecting emotion using hashtags on tweets. Using emotional hashtags such as #sad as a guideline, the study builds on a model that detects emotion from tweets. First, the hashtag predictor was trained on a self-labeled data-set (including the hashtags) which performed than a random classifier, which means that the text of tweets has some characteristics related to hashtags. The significance of this study lies in the fact that the data-set created and validated by this classifier can also be applied to other domains, which means that their data-set can be leveraged for emotion classification tasks (Mohammad and Kiritchenko, 2014).

Based on depressions detection using hash-tags, more researches were conducted to produce better labeled data set and more predictive classification models that analyzed the textual and linguistic information that the tweets provided rather than just using hashtags. There were difficulties in using just emotion labeled data or not-labeled tweets to identify depressive symptoms. Thus, Munmun De Choudhury et al. created a labeled tweet data of individuals suffering from clinical depression using the crow-sourcing technique. Using Amazon's Mechanical Turk interface, the study combined the response of the crowd workers on the depression survey and the self-report to yield depression scores of people based on the CES-D (Center for Epidemiologic Studies Depression Scale) questionnaire. The crowd workers had the choice to opt in to share their Twitter usernames of their public profile. Out of 1,583 participants, 637 participants (40%) agreed to provide access to their twitter feeds. These tweets were fed into Besides the tweets, they made use of other features such as LIWC emotion detection, linguistic style, timestamp of tweets and user-related features such as the volume of replies and tweets were fed into their statistical model. This model analyzed a real-time "behavioral fingerprint" of an individual and predicted whether or not this user will be depressed. This research was significant in that it achieved high accuracy by incorporating linguistic and user-related features that could hint to behavioral environments the user was surrounded by.(Munmun De Choudhury, 2013)

From these ground works that developed models to tag tweets as depressive or not, various researches were conducted with improved classification models that utilized different scoring methods. In 2014, Schwartz et al. developed a model predicting the degree of depression (DDep) based on their language usage on Facebook. Although the accuracy scores of this model were not significant (r=0.386), this study is worth noting as it was the first work to detect continuous-valued depression scores from social media posts (Schwartz et al., 2014). This scoring method can be used to create a more effective and accurate label for depressive tweets as words indicating a person's state of mind can be ambiguous in many ways. Many studies were also conducted to compare performances of models targeting depression. Nadeem, in 2016, examined multiple algorithms to identify Major Depressive Disorder via social media. This study was different from other studies in that its focus on whether the tweets were depressive in nature on a document-level basis. Most of the analysis involved comparing different classifiers and with the priority in accuracy over F-1 score, a Multinomial Naïve Bayes classifier was the best performing model (Nadeem, 2016).

Along with the technological advancements in the field of Machine Learning and NLP over the past few years, Husseini Orabi et.al applied deep learning to find people at risk of depression based on their tweets using the CLPSych 2015 shared task data-set which consists of tweets from people classified into 3 categories; depression, PTSD, and controlled. This classifier primarily used word embedding with multiple loss functions and was tested on several different network architectures including three different CNN setups and on LSTM setup to find the best performing network. A new word embedding optimization technique was introduced in this study and it showed to have improved the performances of the classifiers. Also, between different architec-

tures, the MultiChannelCNN achieved 83.117% accuracy and CNNWithMax achieved 87.957 % accuracy. (Husseini Orabi et al., 2018).

Studies relevant to the field of depression detection in social media expanded from classifying depressed vs controlled to targeting different, more specific types of mental illness such as PTSD. For instance, as part of the CLPsych 2015 Shared Task(evaluation task to detect users with PTSD or depression), Preoţiuc-Pietro et. al developed a model that distinguished between (a) control users and users with depression (b) control users and users with PTSD and (c) users with depression and PTSD using a corpus of users who consented to dispose their mental illness diagnoses on Twitter. This research utilized a wide variety of features such as automatic clustering methods to infer topic and binary unigram vectors which made a note of whether or not a user ever tweet a word. The final model was an ensemble model which was a linear combination of different classifiers and down weighting less informative features. This study is significant as its model tested various ways to group words and to identify which grouping best-indicated depression signal (Preoţiuc-Pietro et al., 2015).

Another study was conducted by Padugupati et.al to specifically analyze heavily suicidal behavior of tweets. A martingale-based framework was built to detect emotion change in tweets using user-related features and linguistic features like the research conducted by De Choudhury et. al. This study classified tweets to 3 classes, No distress, Minimal distress, moderate distress, and severe distress (Padugupati et al., 2019).

Branching out from these foundations paper aims to focus primarily on three things. We have found some limitations to the previous researches in their lack of detailed explanation or discussion of the actual syntactical differences in language that the depressive and non depressive users used on social media. Also, previous works relied significantly on the gensim and NLTK library for their language models only. Therefore, this paper aims to do three things.

(1) Conduct a statistical analysis to discuss the statistical difference between users labeled depressed and non-depressed.

(2) Using a new library called Fasttext, aim to construct an improved classification model.

(3) Apply topic modeling to understand the differences in topics that that depressive and non-depressive users discuss to test if there are any difference in the way in which they utilize their social media.

3 Method

3.1 Pre-processing

As tweets tend to be shorter in length than conventional text data that machine learning models use, to get a more accurate representation of how each *person* uses the languages the tweets were aggregated by users. For transforming the tweets which are from human language to machine-readable format, some text normalization are needed to be done before further processing. The first step of normalization is to lowercase all the letters to have consistent words. Since the focus is on words of the tweets and their semantics, numbers, punctuation, and white spaces are removed. To reduce errors in classifier and topic modeling, non-English vocabulary was removed using NLTK libraries (Bird et al., 2009). Also, the English stop-words are removed using the NLTK stop-words list. Finally, the tweets are tokenized and ready for processing.

3.2 Statistical Analysis

In order to verify our hypothesis, we run some statistical analysis to see if the presence of the words in depressed (positive) tweets are significantly different from non-depressed (negative) ones. We counted the word frequency of the pre-processed tweets included in both positive and negative data sets and created a table similar to the left table in figure 1 . After that, we run the chi-square on each word. Equation 1 shows the chi-square formula and we created the right table in figure 1 to do our calculation.

$$\chi^2 = \sum \frac{(Observed - expected\,\#\,of\ individual)^2}{Expected\,\#\,of\ individual} = \sum \frac{(O-E)^2}{E} \tag{1}$$

3.3 Classification Model

Recognizing depressed users based on their tweets would be really helpful, for example it can help to diagnose and treat them it in first stages. The next task we tried to address was to classify the users to

	Word count	Other words count	total
Positive	A	B	A+B
Negative	C	D	C+D
Total	A+C	B+D	T

	Word count	Other words count
Positive	(A+C)(A+B)/T	(B+D)(A+B)/T
Negative	(A+C)(C+D)/T	(B+D)(C+D)/T
Total	A+C	B+D

Figure 1: The top table is the observed table for each word. The button table is the expected table calculated from observed table.

depressed and not-depressed based on their tweets. For doing so, we used Word2Vec (Mikolov et al., 2013) for embedding the tweets to vectors in an dense space. In summary, the Word2Vec uses the co-occurrence of the words to generate their embeddings. Eventually, words that have similar meanings would be near each other in this dense space. Also, there is some notion of meaning in the resulting embedding. One of the most popular examples in Word2Vec applications is given in figure **??**. After training the Word2Vec on a huge corpus, we can see that if we add the difference between ”woman” and ”man” vectors to the ”king” vector, we get a vector which is pretty near to the ”queen” vector.

After embedding our words, we add all the vectors of each word and n-gram in tweets of a user and use the sum as the input feature for the classifier. For embedding and classifying the users, we used the fastText (Joulin et al., 2016) package provided by Facebook. This library is for text analyzing and it is optimized to be fast even when you run it on CPU.

3.4 Topic Modeling

Used Latent Dirichlet Allocation to model topics with the given text data. LDA, short for Latent Dirichlet Allocation, is a generative probabilistic model that considers each document as it is a combination of different topics. LDA returns mainly the topics and the probabilities of representative

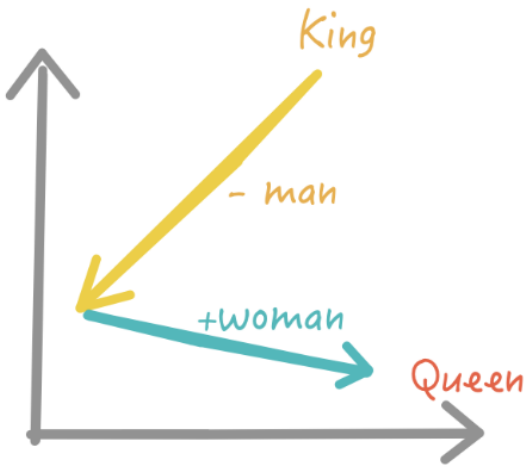

Figure 2: An example of word2vec embeddings

words in the topics. The topics can be viewed as the number of clusters and the probabilities as the proportion of the clusters generated (David M Blei and Jordan, 2003) . However, LDA is most effective for longer documents so the data was aggregated by each user. This increases the amount of relevant word co-occurrences within the documents and made the data more suitable for LDA modeling.

First, the tweets were loaded into a dictionary user:tweets. Then, an LDA model variable was created for each labels using the LDA Model object in the gensim library. Each negative and positive labeled LDA model was trained on the corresponding data-set and produced five topics with proportions and representative words. In the form shown in Figure 3. The negative label LDA took significantly longer because the data set was larger.

```
lda.show_topics(5,10)

[(0,
  '0.019*"like" + 0.017*"im" + 0.016*"na" + 0.011*"dont" + 0.010*"cant" + 0.009*"really" + 0.008*"want" + 0.008*"gon" + 0.007*"go" + 0.007*"wan"'),
 (1,
  '0.028*"rt" + 0.013*"’" + 0.013*"new" + 0.012*"’" + 0.010*"video" + 0.010*"lol" + 0.009*"happy" + 0.008*"yes" + 0.008*"ur" + 0.008*"youtube"'),
 (2,
  '0.089*"rt" + 0.009*"depression" + 0.008*"follow" + 0.008*"amp" + 0.008*"’" + 0.006*…" + 0.005*"tweet" + 0.005*"diagnosed" + 0.005*"help" + 0.005*"peopl
e"'),
 (3,
  '0.039*"rt" + 0.029*"im" + 0.015*"dont" + 0.015*"like" + 0.013*"get" + 0.012*"know" + 0.009*"people" + 0.009*"one" + 0.008*"need" + 0.007*"never"'),
 (4,
  '0.041*"rt" + 0.029*"love" + 0.011*"thank" + 0.010*"much" + 0.009*"foto" + 0.009*"hate" + 0.007*"thanks" + 0.007*"hope" + 0.007*"x" + 0.006*"sorry"')]
```

Figure 3: Sample LDA results

4 Data Used

Data-set with depression labeled users was created by a Shen et al in 2017 (Shen et al., 2017), while conducting a research to identify at-risk users in social media. After the study the data-set was released for public usage. The data set is based on the tweets between 2019 and 2016 and users

were labeled as depressive if their anchor tweets (tweets that they uploaded to their tweeter stream) that included a strict pattern "(I'm/ I was/ I am/ I've been) diagnosed depression". These tweets were labeled as positive. The non depressive users were labeled as positive if they have never posted any tweet containing the word "depress". Figure 4 shows the number of users and tweets been used in the sampled data set for doing this study. The used sample data-set can be found here.

	D1 - positive	D2 - negative
Users	**2626**	**5373**
Tweets	**483419**	**1890167**

Figure 4: Number of users and tweets been used in the sample data set for this study.

We are aware that this is not the most accurate way of categorizing depressive and non depressive users. However, due to limitations in getting patient consent, it is extremely challenging to find a data-set with data from users that were clinically diagnosed with depression. We understand this limitation, but as depression is often self-diagnose and as 1 in 5 people are receiving treatment consistent with current practice clinical guidelines and social media is an outlet in which people expose their feelings, we considered tweets that self-diagnosed themselves with depression a reliable source to label them as depressive for the purpose of this research.

5 Results

5.1 Statistical Analysis

For every word included in both positive and negative data set, the statistical analysis is run and chi-square is calculated. The calculated chi-square are compared to two different critical chi-square values with p-values of 0.01 and 0.05. For each p-value, there is two different word cloud, showing the significant word related to positive tweets and negative tweets. Looking at the positive word cloud in figure 5, the most significantly different words are some words like "depression", "diagnosed", "anxiety", "SOS", and "PTSD" which are related to depression or mental issues.

However, looking at the negative word cloud in figure 6, words like Trump are there. It can be interpreted that not depressed users are more interested in politics rather than their health situation. Also, by looking at other words, their tweets are related to the everyday concerns in compare to the tweets from depressed users.

Figure 5: Word cloud of depressed users tweets

Figure 6: Word cloud of not-depressed users tweets

5.2 Classification Model

As mentioned before, we used fastText library to do our text classification analysis. This tool requires the input to have an special format. Specifically, it needs a file where each line is one data sample where the first word should be the label. Labels should be prefixed with a "__label__" phrase. Moreover, we need to pass hyper-parameters such as the learning rate, number of epochs to the fastText trainer. We also can specify the word count of n-grams used in this model.

In the figure 7, the classification results can be seen. For each model, the f1 score for both positive and negative classes and overall accuracy of the model has been calculated. We have used a bunch of different parameters to find the best performing model. The model with learning rate equal to 0.5, n-gram equal to 8 and trained for 2 epochs gave the best results.

5.3 Topic Modeling

With the word clouds we analyzed the words in each topic carefully. First, from word cloud results from depressive users as shown in Figure 8, we

	Lr	epoch	Ngram	P F1	N F1	precision
1	0.1	1	2	0.57	0.70	0.65
2	0.5	1	3	0.60	0.70	0.66
3	0.5	2	4	0.63	0.70	0.67
4	0.5	2	8	**0.63**	**0.71**	**0.68**

Figure 7: Classification results using different parameter sets

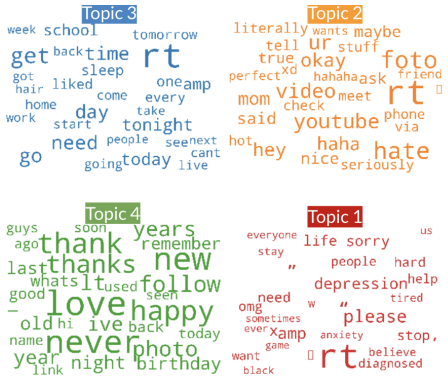

Figure 8: WordCloud result from topic modeling Depressive Users

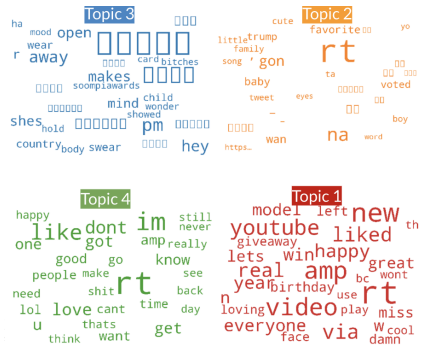

Figure 9: WordCloud result from topic modeling Non-Depressive Users

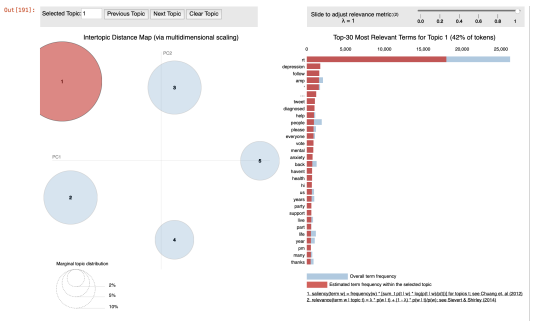

Figure 10: Topic bubble for topic 1

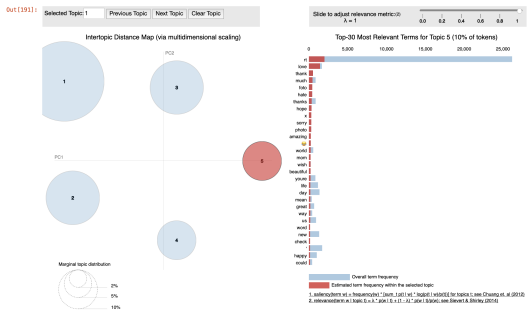

Figure 11: Topic bubble for topic5

were able to detect that topics did not vary much between them except for topic 1(red) indicating words that indicated depression or at least some type of struggle. We were unable to identity significant patterns from word cloud from topic modeling positive users as the data was too vast. However, a thing to note is that the words for each topics aren't coherent and inclusive to that topic only. There are two possible explanation for this. The LDA is producing low coherence scores, but this is unlikely because the topics were coherent for "depressive related" topics for Figure 11. The more likely explanation we found is that the topics don't converge to just 5 topics because of the size of the data-set and that users with non-depressive symptoms talk about a more diverse fields on their social media, compared to those who self-diagnosed themselves with depression.

After analyzing the topics from the word cloud, we decided to further analyze the distance between topics for the topics created for the positive labels using a visualization tool pyLDAvis. The results are shown in Figures 10 and 11. Figure 10 shows the topic bubble for topic 1 and Figure 11 shows the topic bubble for topic 4. The topic bubble shows that the distance between the topics for positive labels are further apart and the topic bubble for topic 1 containing words such as "Depression", "please" ,"help, "need" etc is the biggest bubble and the rest of the topics have similar sizes in terms of proportion.

6 Conclusion

In this work we conducted a study over the depression phenomena and its signs in humans behaviors in social networks, specifically Twitter. We performed several analysis and found out that we can see and detect signs of depression in people's tweets.

At first, our statistical analysis showed that there the depressed users have a statistically significant difference word usage than not-depressed people. They more tend to talk about stuff like anxiety and that could be used as signs of depression. On the other hand, as expected, not-depressed people are talking over lots of different topics including politics in Twitter and there is not a really specific topic that is the center of discussions.

By applying text classification on the tweets authored by different users, we were able to predict if the author is depressed or not to some extent. We looked at the tweets generated in the past 30 days by the user and tried to predict user's depression based on that and we got an accuracy of about 68 percent. This means that we are able to detect depression based on users tweets to some extent which is pretty valuable. We can use such classifier to help detect people who are prone to depression and try to help them, maybe even before they know themselves.

From topic modeling we were able to infer two things. First, depressive users tend to post topics that are irrelevant to one another. For non depressive users, the topic modeling performance was poor in that it was difficult to distinguish the clear difference between the topics. However, by analyzing the word cloud and the distance mapping of topic bubbles for depressive users, we were able to detect that the topics among the depressive user varied. From this, we can infer that depressive users, or users who self-diagnose them as "depressed" on tweeter, tend to use twitter to post their feelings and personal area of interest rather than posting something that will be retweeted often or responding to a tweet. If they were responding more to each other's or other user's tweets, the topic bubbles would have been closer to each other because of the semantic overlap of words used in each tweets. However, there are other researches indicating that passive social media use that include behaviors such as scrolling down news feed is associated with depressive symptoms. (George Aalbers, 2018) This paper can be further improved by utilizing the network perspective of social media that signals depression that is mentioned in this paper published in 2018 analyzing networks.

We are fully aware that data set that we currently used may be limited by the non-specific use of the term "depression", in other words "I am depressed" might not be a clinically valid indicator for depressive users. While earlier research identified depression predictors in twitter tweets for depression, we used a different libraries and statistical approach to tackle different ways of classifying depression. Also, we added topic modeling to get new insights for twitter usage of depressed groups of people.

7 Future Work

Topic modeling could have been improved by aggregating the user by user tweets by weekly or monthly posts, as people's interest and the things they talk about can differ significantly by the time period. Also, for a more model that targets analyzing depressive tweets, the research could have implemented supervised LDA model like the pre-trained LDA model (Philip Resnik and Resnik, 2013) that were trained on Pennebaker stream-of consciousness essays to predict Neuroticism on College students in 2013. Another way to improve topic modeling could have been using bigrams and trigrams, which could not be conducted due to hardware constrains as the word matrices would have grown to too large sizes.

Overall, the performance of models could have been improved by use of a better data-set. A data-set that includes tweets from patients that were diagnosed of depression with evidence from clinical and psychiatry professionals that are collected strictly under consent, rather than self-diagnosed patients.

The significance of this field of research is that it can aid depression management services that utilizes Machine Learning. Some of the services include Virtual Counseling, Tweet Recommendation, Patient Monitoring and Precision Therapy.

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Beekman AT van Zoonen K Dijkshoorn H Cuijpers P Boerema AM, Kleiboer A. 2016. Determinants of help-seeking behavior in depression: a cross-sectional study. *BMC Psychiatry*, pages 16–78.

Andrew Y Ng David M Blei and Michael I Jordan. 2003. Latent dirichelt allocation. *The Journal of Machine Learning Research*, pages 993–1022.

Alexandre Heeren Sanne de Wit Eiko I. Fried George Aalbers, Richard J. McNally. 2018. Social media and depression symptoms: A network perspective. *J Exp Psychol Gen.*

Aron Halfin. 2007. Depression: The benefits of early and appropriate treatment. *The American journal of managed care*, 13:S92–7.

Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. 2018. Deep learning for depression detection of twitter users. pages 88–97.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

J. Mann, Alan Apter, José Bertolote, Annette Beautrais, Dianne Currier, Ann Haas, Ulrich Hegerl, Jouko Lonnqvist, Kevin Malone, Andrej Marusic, Lars Mehlum, George Patton, Michael Phillips, Wolfgang Rutz, Zoltán Rihmer, Armin Schmidtke, David Shaffer, Morton Silverman, Yoshitomo Takahashi, and Herbert Hendin. 2005. Suicide prevention strategies: A systematic review. 294:2064–74.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Saif Mohammad and Svetlana Kiritchenko. 2014. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31.

Scott Counts Eric Horvitz Munmun De Choudhury, Michael Gamon. 2013. Predicting depression via social media.

Moin Nadeem. 2016. Identifying depression on twitter.

Penchalaiah Padugupati, K Nikhitha, Ch Devi, and Ch Ramya. 2019. Detection of suicide related posts in twitter data streams. 09:81 –86.

Anderson Garron Philip Resnik and Rebecca Resnik. 2013.

Daniel Preoţiuc-Pietro, Maarten Sap, H. Schwartz, and Lyle Ungar. 2015. Mental illness detection at the world well-being project for the clpsych 2015 shared task. pages 40–45.

H. Schwartz, Johannes Eichstaedt, Margaret Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. 2014. Towards assessing changes in degree of depression through facebook.

Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, Tat-Seng Chua, and Wenwu Zhu. 2017. Depression detection via harvesting social media: A multimodal dictionary learning solution. In *IJCAI*, pages 3838–3844.

WHO. Depression [online]. 2018.