A Time Series Analysis on Depression During the COVID-19 Period

Alejo Vallega#1, Devin Efendy#2, Kevin Kim#3, Victor Sharonov#4, Joseph Smith#5

#Department of Computer Science, University of Manitoba
66 Chancellors Cir, Canada

1vallegaa@myumanitoba.ca

2efendyd@myumanitoba.ca

3kimh3427@myumanitoba.ca

4sharonov@myumanitoba.ca

5smithj32@myumanitoba.ca

***Abstract*— The novel coronavirus disease (COVID-19) has been declared as a pandemic by the World Health Organization (WHO) on March 11, 2020, and significantly affected people's lifestyle. Depression is one of the common mental disorders found in people during the COVID-19 period.**

***Keywords*— COVID-19, Depression, Machine Learning Classifiers, Twitter, Time Series Analysis, Support Vector Model, Random Forest**

I. Introduction

The novel coronavirus disease (COVID-19) has been declared as a pandemic by the World Health Organization (WHO) on March 11, 2020. It is disheartening that, as of December 2020, around 67 million confirmed cases of COVID-19, including 1.5 million deaths, are reported to WHO [1]. This COVID-19 has continued to be prevalent even in the end of 2020, significantly affecting people's lifestyle and their physical and mental well-being [2].

Depression is one of the common mental disorders found in people during the COVID-19 period. It is a leading cause of disability worldwide and a major contributor to the overall global burden of disease [1]. Traditionally, depression is often diagnosed by a psychological doctor interviewing a patient face-to-face via referring to clinical depression criteria. However people are often ashamed to talk to a psychological doctor or unaware of depression, which leads to more than 70% of people in the early stages of depression not consulting a doctor, letting depression deteriorate their conditions [3]. Instead of seeking professional assistance, people are increasingly relying on social media platforms, such as Facebook and Twitter, to express their emotions in addition to sharing other information about their lives.

This study exploits data collected from . Several machine learning classifier techniques are utilized to identify the depression level, which include Support Vector Machines and Random Forest.

This paper aims to see if there exist new keywords or features from depression-related tweets during the COVID-19 period, and to compare them to existing keywords and features that indicate depression of users.

Our key contribution of this paper is a time series of tweets related to depression during the covid-19 period. Specifically, providing a visualization of the number of depression-related tweets on Twitter during the particular COVID-19 period (April to September). The tweets that indicate depression are classified by our Machine Learning model that we trained. In addition to that, after completing classification on tweets using our machine learning models, we are doing feature extraction on the tweets that are classified as depression-indicating during the COVID-19 period.

The rest of the paper is organized as follows: Section II provides background and related works on depression detection. Section III presents which datasets are used and how datasets are preprocessed prior to applying the machine learning algorithm. Section IV and V describe what features are extracted and which machine learning algorithms are used to detect depression from the datasets. Section VI provides time series analysis. Then, we conclude our study and provide a direction for future work in section VII.

II. Background and Related work

Efforts to analyze depression of an individual have existed since long before the appearance of the Internet. There are many widely-accepted scales and criteria that have been developed based on user study or questionnaire surveys. In the areas of medicine and psychology, several questionnaire-based measures for rating depression in individuals have been proposed [4, 5, 6, 7]. For instance, Center for Epidemiological Studies-Depression (CES-D), Beck's Depression Scale (BDI) and Zung's Self-rating Depression Scale (SDS) estimate the severity of depression in individuals from the self-reported answers to 20 questions [4, 6, 8]. The questions either have several options aligned with different scores or require participants to evaluate the severity of their circumstances. Then, the level of depression is determined according to the scale of the total score [9]. In general, however, obtaining data through a survey or questionnaire is often expensive and sometimes time-consuming.

Nowadays, it is commonly known that almost everyone has an active account on at least one social media platform, such as Facebook or Twitter, allowing a large amount of data to be generated in a short period of time. For example, Facebook has about 2.7 billion active users, and 350 million posts are uploaded each single day[1]. Twitter has around 340 million active users and 500 million tweets (user posts on Twitter) are generated each day, as of December 2020[2]. Through the use of social media's own Application Programming Interface (API), such as Graph API Explorer (for Facebook) and Twitter Search API, researchers can easily access and obtain large scale data of the users. As the large scale data is available publicly due to the presence of social media, approaches that use such data for depression analysis are seen to be compelling to researchers and thereby provide motivation to analyze the online behaviors of depressed users.

Moreno et al. propose that college students experiencing depressive mood show symptoms consistent with depression on Facebook and place greater investment in social media as a communication outlet because it could be viewed as a safe and indirect outlet for their emotions [10].

Park et al. explored the use of language in describing depressive moods by utilizing real time moods captured from Twitter users and analyzed the differences between Twitter users with and without depression by analyzing their activities [11]. In their later work, a similar analysis is done by analyzing data from Facebook [12].

De Choudhury et al. demonstrate the estimation accuracy that could be achieved by utilizing activities on Twitter to predict depression of the users [13]. They obtained training data for machine learning by crowdsourcing (the practice of engaging a crowd or group for a common goal). Then, models that could be used to predict risk of depression were identified from several features obtained from the records of user activity on Twitter by using Support Vector Machine. The result of their experiments show that depression can be recognized among Twitter users with an average accuracy of 70% [13]. These approaches are also applied to predict mothers' postpartum depression (the depression after giving birth) from Facebook and Twitter as well [13, 14].

Tsugawa et al. showed that word frequencies are useful for identifying depression and investigated how useful the various features extracted from Twitter user history are for recognizing depression, and the degree of accuracy with which the presence of active depression could be detected by using these features [7, 16].

Nadeem et al. employed a Bag of Words (BOW) approach which utilizes word occurrence frequencies to quantify the content of a tweet (i.e. putting all words within a bag and measuring how commonly each word appeared). Then they used

[1] https://www.omnicoreagency.com/facebook-statistics/#:~:text=350%20million%20photos%20are%20uploaded,300%2C000%20users%20helping%20in%20translation

[2] https://www.omnicoreagency.com/twitter-statistics/

four types of binary classifiers: linear SVM classifier, decision tree (DT), Naïve Bayes (NB) algorithm, and logistic regression. They found that NB algorithm produced an accuracy of 81% and precision of 0.86, achieving better performance than the other classifiers [15].

Beyond the technical considerations and challenges, there are also ethical considerations to be taken into account when proposing the use of social media data to judge levels of depression. The use of the large volume of data available from social media platforms has potential to aid in early detection of depression [18]. However, there is still doubt as to the ethically appropriate use of this data [17]. Mikal et. all found that many users of Twitter did not understand the permanence of their posts, as well as either a poor or total lack of understanding of the data tools that could be used to analyze their activity on the platform [17]. There are also concerns about privacy and consent, as often users are not informed that their data can and will be used in this way [19].

In this study we are considering purely the technical aspects of social media data analysis for the purpose of finding new keywords or features from depression-related tweets during the COVID-19 period. This paper should not be taken to be a statement of our position on the question of the ethics of putting ours or similar methods into practice.

III. Data collection

For data regarding the time series analysis of depressed tweets during the pandemic, we used datasets provided by IEEE Dataport where it contained multiple CSV files of tweets, specifically tweet IDs, from the timeline of March 20, 2020 and still ongoing. The dataset files had a CSV for each day during the time interval and ranged anywhere from 350,000 to 5,000,000 tweets for each file with a total of about 850,000,000 number tweets as of December 21, 2020. We decided to only extract CSV files of every Friday of the week for all of the months within the timeframe, considering how large the size would be downloading each CSV file. We chose every Friday as an arbitrary choice to reduce the number of CSV files to download and used random sampling to extract only 20 000 tweets per CSV file. We then used a script that transformed the IDs[3] to the proper formatting that included various columns such as the user's post, date of post, username and many more.

For training and test dataset for the different Machine Learning classifiers, we first focused on finding labeled depressed and non-depressed tweets through GitHub, but we found this resulted in some complications. First of all, we only found one dataset that had manually labeled depressed and non-depressed tweets using domain knowledge of depression however only had very few samples. There were many datasets with large samples of labeled depressed and non-depressed tweets but lacked domain knowledge of depression. The method of extraction of those tweets consisted only of finding tweets that mentioned the word "depression" which contained many false positives since that does not truly measure the factors that go into a person's depression. For instance, the phrase "I know someone who has depression" does not indicate depression and the datasets consisted of many cases of similar tweets.

We decided to broaden our search by looking at datasets from other social media sites and found a source of Reddit mental health and non-mental health datasets with features extracted of different subreddits (a web forum of a particular topic) during a specific timeframe [29]. For instance, these datasets included subreddits such as depression, anxiety, suicide watch, fitness, meditation and others. These datasets had specific timeframes which were divided as follows: January 2018 to April 2018 or January 2019 to April 2019, which was used as a control for posts during the mid-pandemic; December 2018 to December 2019, to indicate a full year of posts from before the pandemic; lastly, January 2020 to April 2020, indicating mid-pandemic posts. For training data, we used subreddits from depression, anxiety and suicide watch as our labeled depressive posts since, by observation of the posts and the features

[3] Documenting the Now. (2020). Hydrator [Computer Software]. Retrieved from https://github.com/docnow/hydrator

extracted, they showed good results of signs of depression. For our labeled non depressive posts, we chose subreddits from fitness and meditation as from observation, the posts were not indicative of depression. We only chose timeframes from before the mid-pandemic since the goal of our project was to find predictions on mid-pandemic posts.

In order to extract features from our extracted tweets during the COVID-19 period, we first cleaned all tweets using several steps. Before mentioning the steps, Twitter has a lot of different features that we cleaned from the tweets that are worth mentioning first. Retweet is a repost of another tweet which starts with "RT" from the beginning of the tweet. Hashtag is a word preceded by the hash sign "#". Mentioning is mentioning a particular user in a tweet represented by a "@mention". Twitter also allows emojis in tweets, which are a visual representation of an emotion. The first step of preprocessing included removing any instances of URLs, hashtags, the "RT", and emojis as these caused unnecessary noise. Secondly, we translated any acronyms such as "LOL" or "BRB" to its non-abbreviated form for readability purposes. Thirdly, we converted each word to lowercase and also removed any leading and trailing whitespaces for overall consistency of each tweet. We then removed any stop words, which are common words such as articles, prepositions, and pronouns, as we only wanted to focus on important words that are indicative of depression. We did keep first-person pronouns as depressed people tend to use them more often. This is mentioned by Wang et al., where they showed that "the online depressed writers use more first person singular pronouns" [30]. Lastly, we used lemmatization to reduce inflected and variant forms of a word to its common base form. This is important for feature extraction since we will be using a language corpus, and reducing words to a common base form helps to reduce the search in the language corpus.

IV. Feature Extraction

For feature extraction, we decided to use the following feature set [21]: Textacy readability index [23]; Textacy text statistic and word count [23]; manually built lexicons which include: suicidality, economic stress, isolation, substance use, domestic stress, guns; sentiment intensity analysis from nltk library; and, tf-idf vector. The differences between our features and the referenced paper is that we don't use the LIWC features set (explained in the limitations section).

For extracting words using tf-idf we selected the top 256 words which occur in at least 2 documents (in our case these are tweets) and at maximum of 80% of the corpus (the whole set of tweets). Among all the tweets there are a lot of common words such as "I", "you", "be", "this". The result was that on our first attempt to use tf-idf to extract the features we did not find a lot of words that are interesting because those common words occurred more than the other words. Hence, when calculating tf-idf, we used English stopwords from nltk.corpus to make sure that we extracted meaningful words.

For our training data, the features set that is mentioned before is already included. Therefore, we transformed our test data to match the feature set of our training data.

The same tf-idf calculation (occurrence frequency and number of extracted words) is used for both our training data set and our test data set. To achieve this, we first fit our training data set with TfidfVectorizer from sklearn and save it for later use. Then, we load the same vectorizer that we fit on the training data set to transform the test data set.

V. Machine Learning Algorithm

In this study, two different classifiers are used: Support Vector Model (SVM) and Random Forest.

SVM is a supervised learning model that draws a hyperplane in a high-dimensional space to classify two different classes [27]. In this study, we used the linear-kernel SVM.

Random forest is a machine learning algorithm that solves regression and classification problems using the ensemble of decision trees. The algorithm is a modification of a bagged decision tree that builds a large collection of decorrelated trees to further improve predictive performance[31]. Random forest has a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest [28]. In our paper, we use a classification version of random forest and a random forest classifier from scikit learn for the prediction of possible depressed and nondepressed tweets.

```
Elapsed time to train the model (in seconds): 64.470012
Confusion matrix:
[[ 8323  2127]
 [ 1091 28381]]
Classification report:
              precision    recall  f1-score   support

           0       0.88      0.80      0.84     10450
           1       0.93      0.96      0.95     29472

    accuracy                           0.92     39922
   macro avg       0.91      0.88      0.89     39922
weighted avg       0.92      0.92      0.92     39922
Prediction score: 0.919393
```

Fig 1. Classification results of Random Forest

```
Elapsed time to train the model (in seconds): 7026.242794
Confusion matrix:
[[10344   169]
 [ 3944 25465]]
Classification report:
              precision    recall  f1-score   support

           0       0.72      0.98      0.83     10513
           1       0.99      0.87      0.93     29409

    accuracy                           0.90     39922
   macro avg       0.86      0.92      0.88     39922
weighted avg       0.92      0.90      0.90     39922
Prediction score: 0.896974
```

Fig 2. Classification results of SVM

We compared the SVM and Random Forest performance. At first, we trained SVM and Random forest with a large dataset of approximately 200000 already classified as depressed and non-depressed Reddit posts. SVM took 7026.24 seconds for training with a prediction score of 89.7%. Random forest was much faster and was trained in 64.47 seconds with a prediction score of 91.94%. Complete performance reports can be found on Figure 1 and 2.

Number of tweets	SVM Time (sec)	Random Forest Time (sec)
5150	28.72	0.07
10452	58.35	0.14
16408	91.58	0.23
24442	136.45	0.34
31813	177.72	0.46
40317	225.11	0.59
50643	282.81	0.74
61363	342.66	0.89
71836	401.05	1.04
82550	460.88	1.23

Table 1. Classification times for datasets of different sizes

After the training process was done, we saved the trained models and compared the classification speed of the SVM and Random Forest algorithms on the datasets that we will be using in the future section for depression analysis. From Table 1, we can see a linear relationship between the number of tweets and time algorithms needed to classify tweets. As the number of tweets increases, the time to complete the classification process also increases. Also, we can see that the random forest outperforms SVM in classification speed.

VI. Result, Time Series Analysis, and Word Cloud

In this section, we will talk about the result of our trained Machine Learning models. We did the time series analysis from March 20, 2020 to October 09, 2020. For each week during that period, we pick one day to be classified using our trained model (e.g., March 20, March 27, April 10, etc.). We have a total of 31 dates. For each date, the dataset for that date is about ranging from 1 million to 3 million tweets. Since our computing power was limited, we took 20 000 samples from each date. The 20 000 tweets sample might contain re-tweets of the same thing, so we made sure to remove these duplicates prior to processing. So, our sample may be less than 20 000 tweets. Lastly, we run our trained model on those particular dates, which contain sampled tweets, for classification.

After classifying our test data, we found that our result was not as we expected. The tweets that are classified are not necessarily related to depression. There are some tweets that are completely unrelated to depression. But, the majority of tweets seem to have a negative tone. For example (note: these are *original* tweets that are not preprocessed):

1. When you think you,Äôre starting to get sick and can,Äôt tell if it,Äôs the corona virus,your depression acting up, or you,Äôre just simply sick. üò≥üò≥üò≥
2. RT @cloakzy: I just want this corona shit to pass my anxiety is brutal 24/7
3. This corona virus is bad for my anxiety.

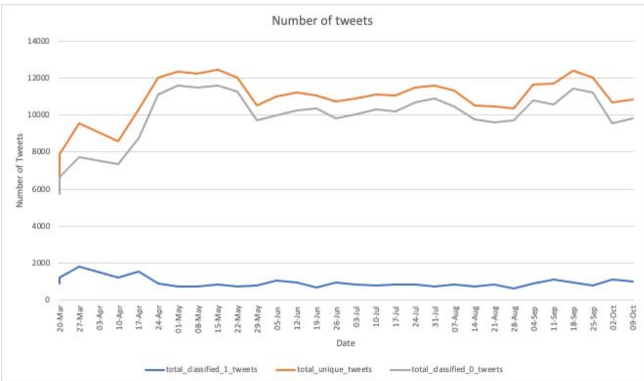

Fig. 3. Visualization of tweets classified as depressed and non-depressed

Fig. 3 shows that the number of tweets that are classified as depressed (blue line) is way less than the number of tweets that are not classified as depressed (gray line). note: orange line is the total tweets that the model tries to classify.

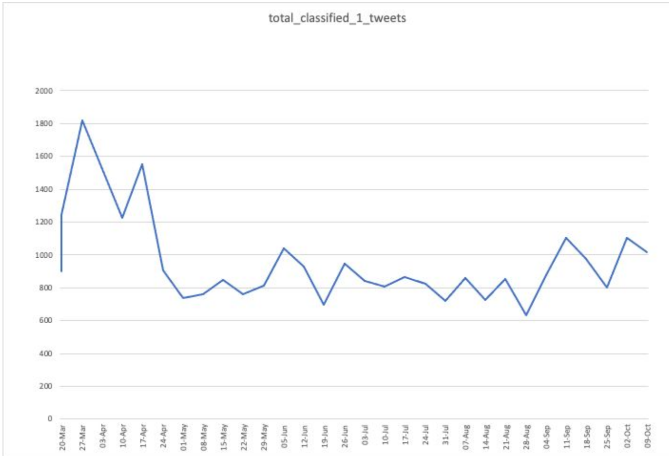

Fig. 4. Visualization of number of depressed tweets during the COVID-19 period

Fig. 4 shows the trend of the number of classified depressed tweets during the period. From the graph, we can see that there is a spike between March 20 and April 17. When we take a look at the classified tweets during that period, we saw some negative toned tweets related to Covid-19. For example: *"This corona virus is bad for my anxiety."*.

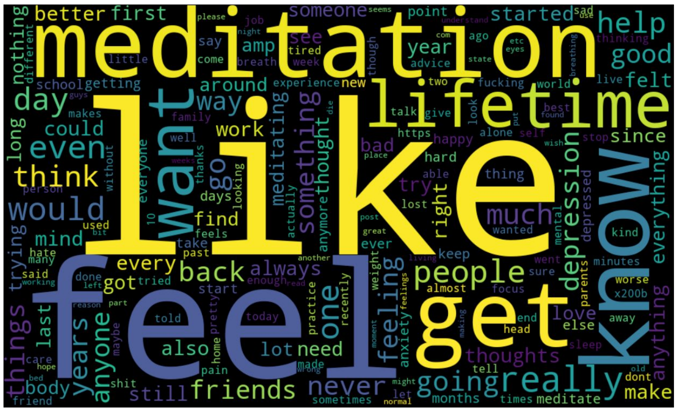

Fig. 5. Top 256 most frequent words

Fig. 5 shows the word cloud of the most frequent 256 words that appears on the classified tweets.

VII. Evaluation

There are various limitations throughout the process of our technical implementation of this project. Firstly, the training dataset includes a LIWC features set. But, our test dataset does not have LIWC features set since we do not have the license for LIWC API. Thus, we needed to remove the LIWC features set and work with the remaining features. From numerous research papers that we read, many of them use LIWC to help with the text classification. Therefore, we think that the LIWC features set is an integral part of text analysis.

Our method of preprocessing could be improved. For example, removing special characters; many of the tweets contain slang and acronyms. In this case, we need to transform those words into their proper form of words; using correct spelling. We did try to correct the spelling of mis-spelled words however, with our current machine or computing power, it is not feasible to correct the spelling for all of the documents, given our time restriction.

Since we did not have sufficient computing power, this also led us to other difficulties such as: we are only able to process a limited amount of tweets. For instance, we only use 20,000 tweets for each date meanwhile there are about a millions tweets per day. Furthermore, by only limiting ourselves to 20,000 tweets we might not get a significant amount of samples that are related to depression. Additionally, given our large dataset, training the Machine Learning models (particularly SVM) took a significant amount of time to finish. Hence, it is not feasible to tune our parameters multiple times to get the best model for the task.

Lastly, due to Twitter's 280 character limit per tweet, the average tweet length that we have is fairly short. Thus, extracting features are difficult given the lack of words. This is because it is difficult to determine the meaning and sentiment of very short sentences.

There is certainly significant room for future development of our methods outlined in this paper to improve accuracy and specificity. As Seabrook et. all explained, while specific keywords and topics in social media posts can be used to indicate depression with a high degree of accuracy, this approach also leads to a high degree of false positives being marked as indicative of depression [24]. They further explain that analyzing the change in emotion across different posts for patterns that indicate depression could provide improvement in specificity of the results [24] to help screen out false positives.

A future version of our algorithm could add additional processing to analyze the changing depression indicators between different tweets from the same person to prune the data of those users whose emotional patterns do not correspond to those considered indicative of depression.

It would also be useful to investigate the posting patterns of users who are depressed as compared to those who are not, such as when they tend to post, how often they post, and what location they generally make posts with different indicators from. If this data has a statistically significant difference between depressed and non-depressed people, it could be used to further prune and refine the results.

Additionally, it would be useful in the future to test our methods on tweets from known individuals, who could be classified into known depressed and control groups. This would allow us to test the accuracy of our method using posts made by people whose depression status is known, and compare those results to those for which we used indicators suggested by previous works.

VIII. Conclusion

In this paper, a time-series analysis was performed on the datasets from IEEE Dataport submitted by Rabindra, Lamsal are presented [20]. Our aim was to investigate if there exist new keywords or features from depression-related tweets during the COVID-19 period and compare them to existing keywords and features that indicate depression of users in social media. We employed two different classifiers to identify the depressed tweets among the datasets and conducted a time-series analysis. We found that there is a spike of depression related tweets during the end of March to mid April 2020. After that the number of tweets related to depression went down and fluctuated until October 2020. Of course, as mentioned before, our classification might not be accurate and the result is not representative of what actually happened. In modern life, use of social media to express an individual's emotion and feeling has become ubiquitous. Hopefully, in the future, we expect this study to provide more insights and perspectives for researchers in depression analysis and psychology.

References

[1] W. H. Organization. (2017). Depression and Other Common Mental Disorders: Global Health Estimates. Geneva: World Health Organization. [Online]. Available: http://www.who.int/en/newsroom/fact-sheets/detail/depression

[2] Bueno-Notivol, Juan et al. “Prevalence of depression during the COVID-19 outbreak: A meta-analysis of community-based studies.” International journal of clinical and health psychology : IJCHP, 100196. 31 Aug. 2020, doi:10.1016/j.ijchp.2020.07.007

[3] Shen, G., Jia, J., Nie, L., Feng, F., Zhang, C., Hu, T., Chua, T.-S., & Zhu, W. Depression detection via harvesting social media: A multimodal dictionary learning solution in IJCAI 3838-3844 (2017).

[4] Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., and Erbaugh, J. An inventory for measuring depression. Archives of General Psychiatry 4, 6 (June 1961), 561–571.

[5] Hamilton, M. Development of a rating scale for primary depressive illness. British Journal of Social & Clinical Psychology 6, 4 (Dec. 1967), 278–296.

[6] Radloff, L. S. The CES-D scale a self-report depression scale for research in the general population. Applied Psychological Measurement 1, 3 (1977), 385–401.

[7] Tsugawa, Kikuchi. “Recognizing Depression from Twitter Activity.” Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, ACM, 2015, pp. 3187–96, doi:10.1145/2702123.2702280.

[8] Zhung, W. W. K. A self-rating depression scale. Archives of General Psychiatry 12, 1 (1965), 63–70.

[9] Shen, Guangyao, et al. "Depression Detection via Harvesting Social Media: A Multimodal Dictionary Learning Solution." IJCAI. 2017.

[10] Moreno, Jelenchick. “Feeling Bad on Facebook: Depression Disclosures by College Students on a Social Networking Site.” Depression and Anxiety, vol. 28, no. 6, Wiley, June 2011, pp. 447–55, doi:10.1002/da.20805.

[11] Park, M., Cha, C., and Cha, M. Depressive moods of users portrayed in twitter. In Proceedings of the ACM SIGKDD Workshop on Healthcare Informatics (HI-KDD’12) (Aug. 2012), 1–8.

[12] Park, S., Lee, S. W., Kwak, J., Cha, M., and Jeong, B. Activities on Facebook reveal the depressive state of users. Journal of Medical Internet Research 15, 10 (Oct. 2013), e217.

[13] De Choudhury, M., Gamon, M., Counts, S., and Horvitz, E. Predicting depression via social media. In Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM’13) (July 2013), 128–137.

[14] De Choudhury, M., Counts, S., and Horvitz, E. Predicting postpartum changes in emotion and behavior via social media. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI’13) (Apr. 2013), 3267–3276.

[15] Nadeem, Moin. Identifying Depression on Twitter. July 2016.

[16] Tsugawa, S., Mogi, Y., Kikuchi, Y., Kishino, F., Fujita, K., Itoh, Y., and Ohsaki, H. On estimating depressive tendency of twitter users from their tweet data. In Proceedings of the 2nd International Workshop on Ambient Information Technologies (AMBIT’12) (Mar. 2013), 29–32.

[17] Mikal, H. (2016) Ethical issues in using Twitter for population-level depression monitoring: a qualitative study. BMC medical ethics. [Online] 17 (1), 22–22.

[18] Reece, R. (2017) Forecasting the onset and course of mental illness with Twitter data. Scientific reports. [Online] 7 (1), 13006–13011.

[19] Nicholas, O. (2020) Ethics and Privacy in Social Media Research for Mental Health. Current psychiatry reports. [Online] 22 (12), 84–84.

[20] Rabindra Lamsal, March 13, 2020, "Coronavirus (COVID-19) Tweets Dataset", IEEE Dataport, doi: https://dx.doi.org/10.21227/781w-ef42.

[21] Low, D. M., Rumker, L., Talkar, T., Torous, J., Cecchi, G., & Ghosh, S. S. (2020, July 13). Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on Reddit during COVID-19: an observational study. https://doi.org/10.31234/osf.io/xvwcy

[22] Webb, Geoffrey I., Eamonn Keogh, and Risto Miikkulainen. "Naïve Bayes." Encyclopedia of machine learning 15 (2010): 713-714.

[23] DeWilde B. textacy Documentation. 2020. Available from: https://buildmedia.readthedocs.org/media/pdf/textacy/latest/textacy.pdf

[24] Seabrook, K. (2018) Predicting Depression From Language-Based Emotion Dynamics: Longitudinal Analysis of Facebook and Twitter Status Updates. Journal of medical Internet research. [Online] 20 (5), e168–e168.

[25] Rish, Irina. "An empirical study of the naive Bayes classifier." IJCAI 2001 workshop on empirical methods in artificial intelligence. Vol. 3. No. 22. 2001.

[26] Zhang, H.. “The Optimality of Naive Bayes.” FLAIRS Conference (2004).

[27] Noble, William. “What Is a Support Vector Machine?” Nature Biotechnology, vol. 24, no. 12, Springer Science and Business Media LLC, Dec. 2006, pp. 1565–67, doi:10.1038/nbt1206-1565.

[28] Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

[29] Low, Daniel M., Rumker, Laurie, Talker, Tanya, Torous, John, Cecchi, Guillermo, & Ghosh, Satrajit S. (2020). Reddit Mental Health Dataset (Version 01) [Data set]. http://doi.org/10.17605/OSF.IO/7PEYQ

[30] Wang X., Zhang C., Ji Y., Sun L., Wu L., Bao Z. (2013) A Depression Detection Model Based on Sentiment Analysis in Micro-blog Social Network. In: Li J. et al. (eds) Trends and Applications in Knowledge Discovery and Data Mining. PAKDD 2013. Lecture Notes in Computer Science, vol 7867. Springer, Berlin, Heidelberg. https://doi-org.uml.idm.oclc.org/10.1007/978-3-642-40319-4_18

[31] Boehmke, Brad & Greenwell, Brandon. (2019). Hands-On Machine Learning with R. 10.1201/9780367816377.