STATS 402 – Interdisciplinary Data Analysis Depression Detection on Social Media with Natural Language Processing Approaches Final Report

Lin He, Jingcheng Wu
lh317@duke.edu, gw93@duke.edu

Abstract—**Depression is a significant factor that contributes to suicide globally every year. To screen out depression patients for early intervention and treatment, researchers have proposed to detect depression on social media. In this project, we explored how effective it is to detect depressed users based on their social account activities. We experimented with NLP methods, including BERT and LIWC, and supervised classification models, including kNN, SVM, LR, and Neural Network, to identify labeled depressed users from a Twitter database by their posts. To test if BERT captures the language patterns of depressed people and if the classification model can be applied to general social network data, we analyzed the data mislabeled by the models and searched for distinct patterns. We further applied more user profile features to the classification and checked which features helped increase the accuracy of the classification.**

I. Introduction

Depression is a common mental disorder that has serious impacts on human's lives. People with depressive disorder not only suffer from sad mood, but also experience other negative symptoms, which might include loss of sleep and appetite, restlessness and tiredness, loss of concentration, and disinterest in daily activity and job [1]. A huge number of people around the world live under the shadow of depression. According to WHO, the estimated population who suffer from depression takes up 5% globally [2]. Global Burden of Die discovered depressive disorder was the sixth major cause of disability to adjust to life among teenagers and middle-aged people [3]. Severe depression may lead to self-injury and even suicide. As stated in the risk factors for suicide by Hawton et al., depression is strongly associated with non-fatal self-injury and suicide [4]. Because of the huge damage depression can cause to patients and society, it is important to take depression seriously and provide patients with proper treatment.

However, many people who suffer from depression do not seek treatment. The barriers that stop people from seeking professional help might include social stigmatization and embarrassment, financial pressure, privacy concern and the belief that feeling depressed is a sign of weakness. According to a web-based survey by Yoshikawa et al., around 55% of people reported being unwilling to seek help for depression [5]. To better help these people who are reluctant or unable to actively seek mental treatment and reduce suicide cases, it is crucial to find a convenient and accurate way to classify people with the depressive disorder from the population and offer them professional help.

Since social media is becoming a more and more inevitable part of people's lives, it is also becoming a rich data source for AI techniques' application in depression diagnosis and detection [6]. People will pose content that shows subtle cues of their symptoms on social media, and their activities on social media (number of friends, number of interaction, when do they post on social media, and etc.) also reflects their inner world. Given the rich information embedded in social media and easy access to all the data, it is meaningful to explore more ways to interpret the data and leverage the data to help depressed people.

In this study, we have 3 objectives. First of all, test if we can use NLP method to classify depressed users and normal users solely based on the written content they post. In addition, check if the NLP method captures any language pattern of depressed patients in general. Last but not least, test if adding more social account activities features will improve the detection accuracy.

We use NLP models as our major methods because written content contains direct expressions of people's feelings, and it is a dominant part of social media. Since we want to do binary differentiation, we use supervised classification models to handle the encoded language data. We also do not want to waste additional information about the user profiles, so we will use feature extraction to find the important features.

With the objectives in mind, we found the Twitter database curated by Guangyao Shen et al. [7]. The database includes texts posted, the user's activity data, the user's network data, as well as clear "diagnosed with depression" labels and "normal people" labels. We have decided to proceed with Twitter as our target platform, for the three reasons listed below. First, Twitter is a personalized social media site that allows users to express themselves without limiting themselves to certain topics or contexts. This means that models trained for Twitter would be more likely to generalize well on other platforms. Two, the publicly available Twitter data can be scraped and collected without legal and ethical concerns, and it is relatively easy to do so. Three, there are trail records on detecting depression on Twitter by other researchers for our reference.

In the following part of the report, we will discuss the related work on depression detection on social media. Then we will introduce our methods of combining NLP methods Linguistic Inquiry and Word Count (LIWC) and BERT with classification models to do the classification and our methods for feature extraction. We will also show how we evaluate the performance of BERT model. Finally, we will analyze and discuss the result of our experiments.

II. Related Work

Since both researches on depression and social media are heated topics, there are already many datasets curated and many experiments done for depression detection on social media. Most research focuses on the 3 types of data on the account: texts posted, the user's activity data such as how many times they post each day and when they usually post, and the user's network data, such as the number of followers [8]. According to Tadesse et al., the common method for depression detection on social media is to conduct feature extractions and feed the features into classification models [8]. The common feature extraction tools in depression detection include LIWC, N-grams and topic modeling [8], and the classification models include Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), Neural Network (NN), Convolutional Neural Network (CNN), etc. [9]. There is also an attempt to apply deep neural networks to distinguish depressed tweets and non-depressed tweets by Sijia Wen that yielded promising results [10].

Combining NLP models as encoders and classification models as decoders to do the depression is straightforward. The model LIWC model is pre-trained, and the N-grams model, SVM, RF and LR are all easy to train. However, we cannot tell from the result if the model learns any useful language pattern, or if it only learns how to match new data with the previous data it learned. Then we do not know if the models trained on one dataset can be applied to a more general situation. To try to answer this question and find if any language pattern exists among depressed people, we conducted an analysis of the texts based on the classification results.

Apart from the analysis of the NLP + classification models, another innovation of this study is using several posts of users instead of judging the emotion based on one single post. Since depression is a chronic disease and everyone has ups and downs in their lives, it is more reasonable to classify based on the overall content and tone of the user's posts. An extension of this innovation is to combine language features with account profile features, with the hope of capturing a more accurate user profile.

III. Proposed Methods

A. Data Set Preparation

We adopted a well-curated data set created by Shen et al. [7] that has been cited by other researchers. We wrote our Python notebook for data parsing, cleaning and pre-processing. We first parse all JSON files into CSV files, then concatenate CSV files for further cleaning and preliminary data analysis. According to our need at the current phase of our project, when parsing data we only preserved data columns of ”text”, ”user_name”, ”user_screen_name”, ”user_id”, ”user_follower_count”, ”created_at”, and ”user_lang”. ”User_name” and ”user_screen_name” are for distinguishing organization or media accounts, ”user_id” is the non-repetitive identification of the users. ”User_follower_count” is for filtering out influencers and organizations with high number of followers. We have also included user metrics including ”user_friends_count”, ”user_listed_count”, ”user_favourites_count”, ”user_statuses_count”, and ”user_profile_background_color”. We have also created a duplicate of the data set where all the tweets by a single user are concatenated into one long string so that the unit for training and testing is used instead of single tweet texts.

We have adopted the approach of filtering out those accounts with over 20,000 followers. Column ”created_at” contains the date and time when the tweets were posted. And ”user_lang” helped us drop those users whose languages are not English, who only take up a small fraction of the whole data set. Moreover, we have also removed duplicate lines using both ”user_id” and ”text”, and dropped empty rows. After all the cleaning and pre-processing, our data set now contains the tweets posted by 3946 undepressed users and 226 depressed users within a month. At this point, we have attained an imbalanced dataset from both the aspects of the number of users and tweets.

We have decided not to directly normalize our data set at this step, because we would like to test how our models perform on both balanced and imbalanced data sets.

In terms of labeling, we set the post's label the same as the user's label. To create a merged post for each user, we randomly selected the user's posts and combined the posts.

Label	No. of Users	No. of Tweets
Depressed	226	71210
Undepressed	3946	2489367

Table 1. Overview of data set after data cleaning and pre-processing

B. NLP + Classification model

1) Encoding methods: This study used two different encoding methods: DistilBERT-based pretrained uncase model and LIWC.

BERT was developed recently by Jacob Devlin et al. [11] and is a pre-trained model using a plain text corpus. BERT has shown promising results in detecting self-harm on social media [12]. We based our work on A Visual Notebook to Using BERT for the First Time by Jay Alammar [13]. The notebook is originally written to use BERT or DistilBERT to perform binary classification of positive and negative movie comments. The pre-trained models adopted are ”distilbert-based-uncased” and ”bert-base-uncased”. According to the authors of DistilBERT, the distilled version of BERT can reduce the size of BERT model while preserving most of its performance [14]. The input text goes through tokenization, padding and masking before being passed into BERT for learning.

LIWC is a popular method in the field of psychology designed to analyze the tone and psychological indications [6]. It is a pre-trained model that gives scores between 0 and 100 to over 100 psychological and language use categories.

2) Classification models: We used SVM, logistic regression, kNN and Neural Network as the classification tools. The Neural Networks that we used all contain 3 layers, one input layer with appropriate number of nodes corresponding to the number of input features, a dense layer with ReLU activation, and an output layer with 1 node and sigmoid activation.The epoch number is set as 50. With a relatively small data set, adding extra layers could easily lead to overfit, hence we have used this simple network structure.

C. BERT + LR Model Analysis

We first separated the merged posts of users that are classified by Bert + Logistic Regression into four categories: True Positive, True Negative, False Positive, False Negative. To test if BERT model actually captures any language pattern of depressed users, we came up with two methods. First, test BERT on single posts and see if the prediction differs from the merged posts. Check if there are any connections between the single posts prediction and merged prediction. Since our goal is to detect depressed users, we focused on the depressed users that are mislabeled as normal users by BERT+LR model. Among the False Negative group, we broke down the merged posts into single texts, ran Bert + LR model on the single texts again, and checked which single text got mislabeled.

We also checked if there were any differences between the four classification groups in terms of language use and emotions that could be picked up by LIWC. We ran the LIWC and analyzed the mean, standard deviation, higher quartile scores and lower quartile scores of the four groups across all the 99 categories of LIWC.

D. User Feature Extraction

Inspired by the work by Shen et al. [7], we have experimented with including some user features into our model. From the datatset, we first picked a series of user metrics, including ”user_friends_count”, ”user_listed_count”, ”user_favourites_count”, and ”user_statuses_count”.

We hypothesized that users' profile background colors could be an indicator of their emotional status. The dataset comes with a feature named ”user_profile_background_color”, which is the HEX format of the user's profile background color. In order to turn it into a feature compatible with the machine learning model, we converted the HEX format color into HSL format, and then extract the multiplication of Saturation and Luminance. We take multiplication. We then ended up with a feature called 'L_times_S', whose range is from 0 to 1, and the higher number represents lighter profile background color.

We have also created a feature called ”antidepressant_count” that records the mentions of antidepressant medicines. We referred to the Internet for a list of 10 popular antidepressants [15]. We then run through every user to count how many times all of their tweets mentioned these antidepressants.

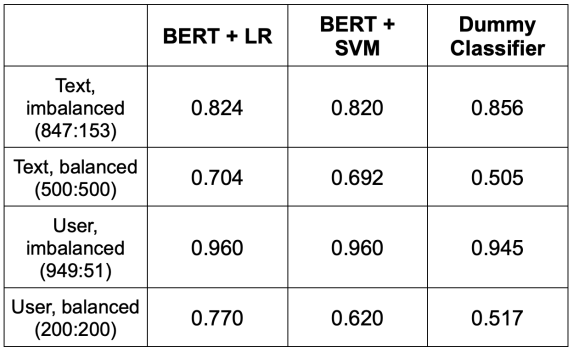
Finally, according to the work by Shen et al. [7], depressed users as a whole are more active than the undepressed users late at night, and undepressed users are more active in the morning than the depressed user. Therefore, we created a feature called ”time_feature” by the following steps. First, the default timestamp of the tweets is in UTC. We extracted another column called ”user_utc_offset” to convert the post time from UTC to the time relative to the users. Then, we divided the 24 hours of a day into four periods, morning (6AM-12PM), afternoon (12PM-6PM), evening (6PM-12AM), and night (12AM-6AM). Then for each user, we calculate the portions of how many tweets were posted during these four periods of a day. Finally, we assigned 1 point to the morning period, and 2 for afternoon, 3 for night and 4 for midnight, and calculated the score for each of the users by multiplying the number of points of the periods by the corresponding portions of the periods. Then, the ”time_feature” ranges from [1.0, 4.0].

IV. Performance evaluation

A. *Performance of BERT and Classifiers*

We first passed in the tweet texts directly with labels of ”depressed” or ”undepressed”. Every single text posted by a depressed user is marked as ”depressed”, and the same for undepressed. It is expected that the model would not perform quite well, because not every single text posted by a depressed user has to be associated with depression. Depressed users might have good days when they post happy tweets and undepressed users might have a bad time when they post negative tweets. This problem makes the labels of texts unclear and perhaps confusing for the machine. As expected, training the model based on these unclear labels did not yield very impressive results. We encoded the texts using BERT, and we tried using logistic regression and support vector machine as the discriminators. On a balanced dataset, the LR yielded an accuracy of 0.704 and SVM got 69.2%, while the baseline dummy classifier has an accuracy of 50.5%. On an imbalanced dataset, which is more close to the real-life situation, LR yielded an accuracy of 82.4%, and SVM 82.0%, both even failing to exceed the dummy classifier’s accuracy of 85.6%. We, therefore, conclude that giving texts individual labels and training BERT model based on these labels is not effective enough.

We then merged the texts posted by the same user and treat each user as the smallest unit in our classification. However, we ran into the problem that when encoding merged texts, BERT model can only take up to 512 tokens at a time, which resulted in a truncation of the texts. This results in the unavoidable loss of useful information in the merged texts and we have no idea how the truncated part will affect the final decision. Not surprisingly, due to the loss of information, the model’s performance does not have a big improvement. On a balanced dataset, the logistic Regression has 88% accuracy and the SVM has 62% accuracy, whereas the dummy classifier has an accuracy of 51.7%. On an imbalanced dataset, both LR and SVM yield an accuracy of 96%, but it is not much improvement from the baseline of 94.5% To overcome this truncation problem, our next step plan is to encode the texts individually and then merge the encoded texts, before applying SVM and LR to perform the binary classification on users.



	BERT + LR	BERT + SVM	Dummy Classifier
Text, imbalanced (847:153)	0.824	0.820	0.856
Text, balanced (500:500)	0.704	0.692	0.505
User, imbalanced (949:51)	0.960	0.960	0.945
User, balanced (200:200)	0.770	0.620	0.517

Table 2. The results of DistilBERT-based classifier models compared to dummy classifier

B. *Analysis of BERT+ LR model*

We randomly selected 200 depressed users and 200 normal users for the test. We randomly selected 10 posts for each user so that the word count did not exceed the BERT’s limit. In this sample, 83 depressed users and 83 normal users were mislabeled. This might indicate that the model has no preference for either group. In addition, we did not find any apparent connection between the mislabeled separate text and the merged text. Among the 83 mislabeled users, there are 36 of them who have 0-3 posts mislabeled, 23 of them who have 4-6 posts mislabeled and 24 of them who have 7-10 posts mislabeled. The mislabeled case occurs in a situation where either all the single posts are labeled positive or none are labeled positive.

To see if the overall language use of users is reflected in the classification, we analyzed the mean score (Fig 1), standard deviation, higher quartile, and lower quartile (Fig 2) of the four groups across 99 LIWC features.

There are several interesting points in the mean score analysis graph (Fig 1). First of all, among the 99 features, there are only 8 features in which the four classification groups have evidently different mean values. These 8 features are personal pronoun use, verb use, cognition, cognition process, conversation, and internet slang use, which are 0, 15, 22, 24, 94, 95 according. It is worth noticing that in these stand-out features, the FN score and TN score are close, the FP score and TP score are close, and more importantly, the scores of these four groups line up in the order of FP, TP, FN, and TN. The fact that True Positive and False Negative is close while True Negative and False Positive is far from each other might suggest that diagnosed depressed users have much more in common regarding the content and language used in their posts. However, the fact that the mean scores are similar in most of the features suggests that the BERT-logistic model does not capture many psychological nuances in language, at least not in a way that the popular psychology language analysis tool LIWC can analyze.

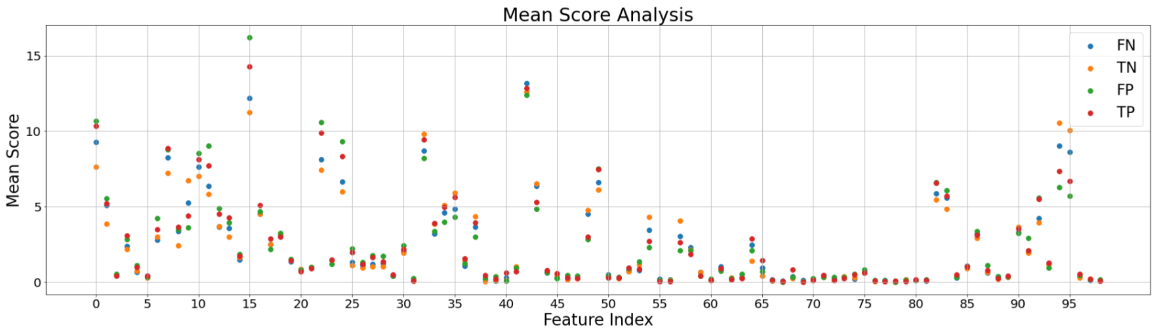


Figure 1. The mean LIWC score for merge texts

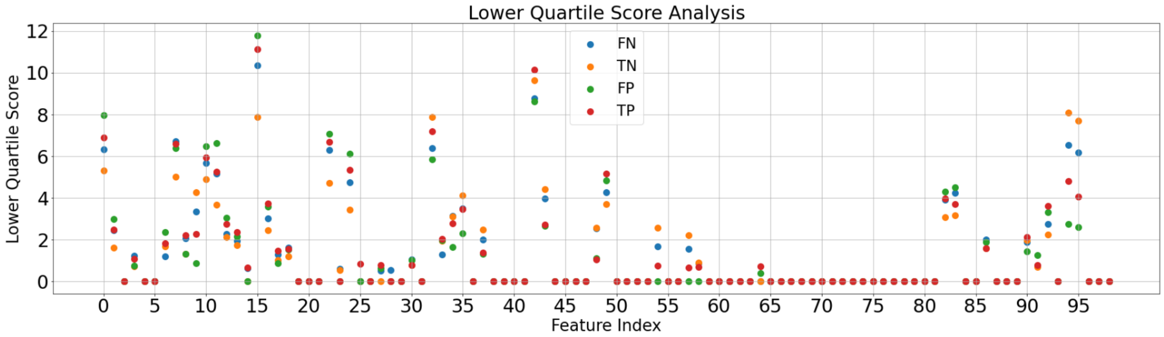


Figure 2. Lower quartile LIWC scores of merged texts

Another interesting discovery is that many feature scores are 0 within the True Positive group. It might be another piece of evidence that the BERT-LR model does not make classification decisions based on the psychological nuances in language, at least not in a way that we usually think about people's tones. It is also possible that such general language patterns among depressed people that we are looking for do not exist.

C. Performance of LIWC and Classifiers

We manually chosed 33 depression-related categories in LIWC to be analyzed. We balanced the data set and trained the data using K-Nearest Neighbors and Logistic Regression. We found that when the number of neighbors is around 11, the model achieves the best performance, which is 71% (Fig 3). Since there are only around 200 users in each label, having large k values increase the variance and compromise the performance. Logistic Regression The accuracy of Logistic Regression is around 65%.

D. Neural Network Experiment Results

1) LIWC Features Only, 33 Features: We first manually chose 33 depression related features to be analyzed. We then run the data into a neural network with an input layer of 33 nodes, a dense layer of 33 nodes with ReLU activation, and an output layer with 1 node and sigmoid activation. We set the epoch number to be 50. On an imbalanced data set where a dummy classifier can score an accuracy of 0.9517, the neural network test accuracy reaches 0.9415 and test loss 0.2857.

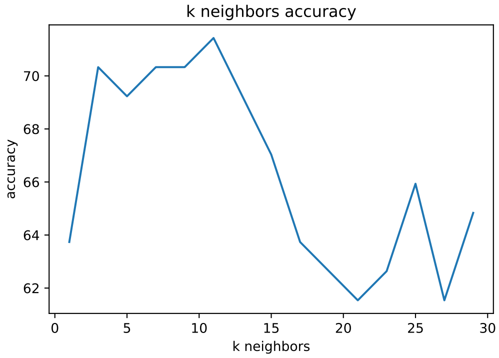


Figure 3. The accuracy of KNN model using a different number of neighbors

2) LIWC Features Only, 6 Features: We then used Scikit-leran Feature Selection to pick out the 6 most important features. Then we pass data into neural network with an input layer of 6 nodes, a dense layer of 6 nodes and an output layer. Test accuracy is 0.9497 and test loss is 0.1706.

3) Expanded Features, Imbalanced Data Set: We then included the previously mentioned user features that we have extracted into the data set. With the additional 7 features, the data set now contains 41

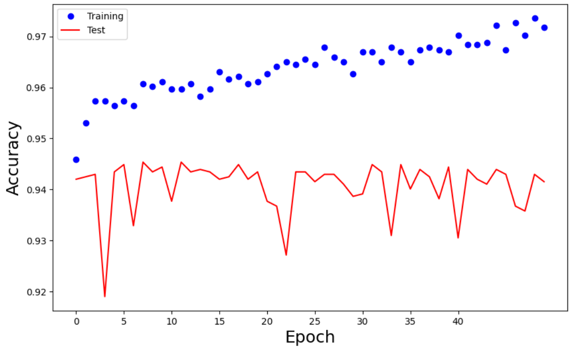


Figure 4. Neural network accuracy over epochs, 33 features, imbalanced data set

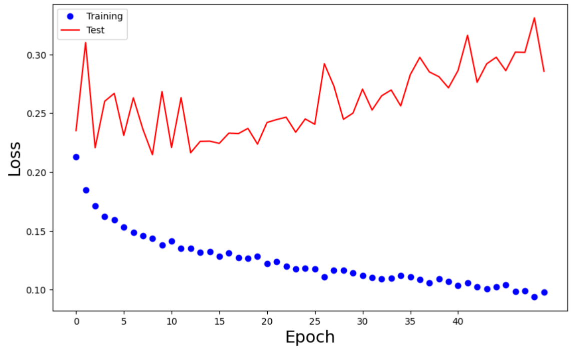


Figure 5. Neural network loss over epochs, 33 features, imbalanced data set

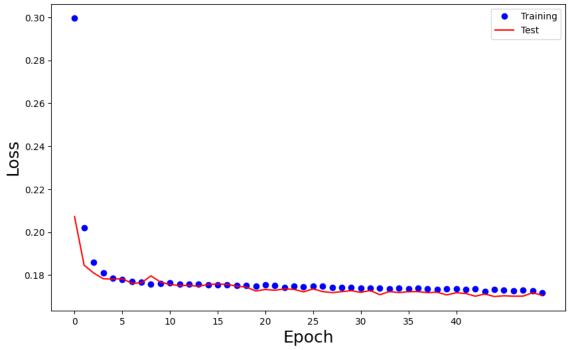


Figure 7. Neural network loss over epochs, 6 features, imbalanced data set

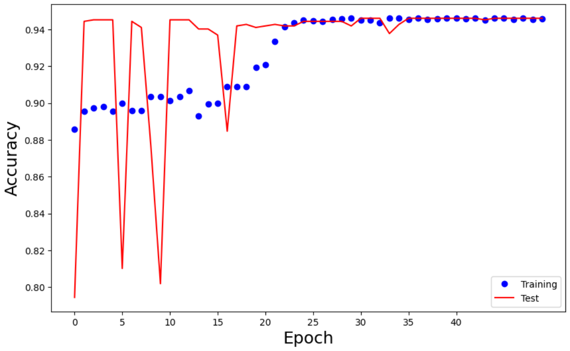


Figure 8. Neural network accuracy over epochs, 41 features, imbalanced data set

features. We passed in the data to the neural network (41-node input layer + 16-node dense layer + output layer). Test accuracy is 0.9461 and test loss is 0.6315.

4) Expanded Features, Balanced Data Set: The distribution histogram exposes that there is an inherent issue with the profile background color feature. The value 0.467320 has 1468 records, and 0.323529 has 489, and 0.005882 has 186, and finally, 0.000000 has 168. This is because Twitter comes with a series of default profile background color, and a large portion (over 90%) of the users are using these default colors, which makes this feature a lot less indicative.

E. User Feature Evaluation

We exported a feature correlation heatmap for the expanded 41 featuers. We discovered that the feature ”time_periods” has a correlation of 0.4 with ”label”. To verify the validity, we made quick visualization of the histograms, which revealed that the overall distribution of the depressed users’ ”time_feature” is slightly different from the distribution of the undepressed ones, further indicating that user post time

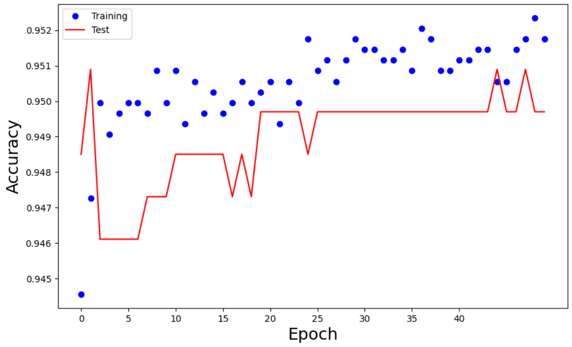


Figure 6. Neural network accuracy over epochs, 6 features, imbalanced data set

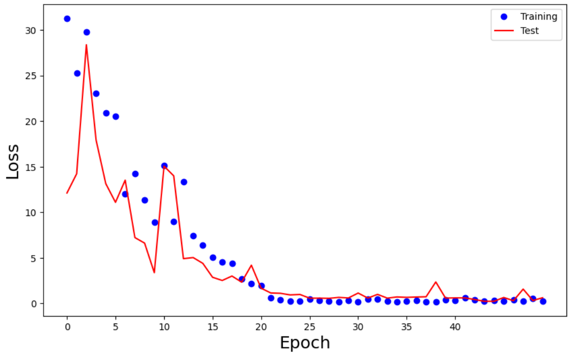


Figure 9. Neural network loss over epochs, 41 features, imbalanced data set

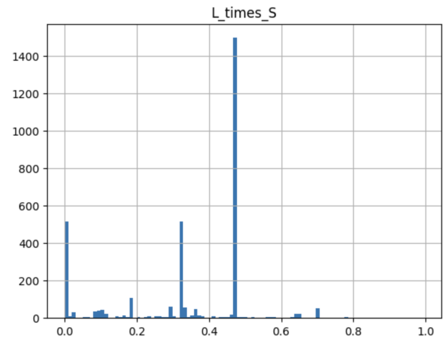


Figure 10. Distribution of feature ’L_times_S’

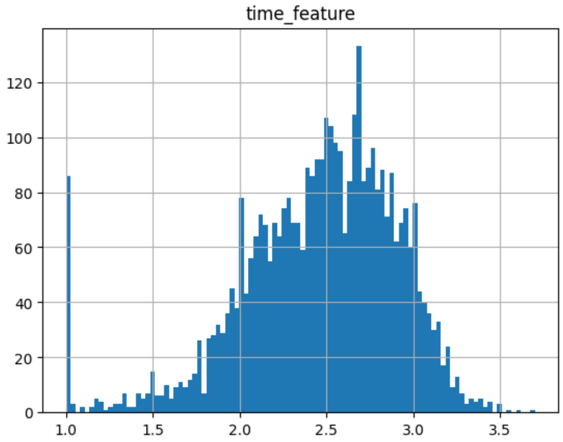


Figure 11. Distribution of feature ’L_times_S’

could be engineered into a feature that contributes to depression detection on social media.

V. Conclusion and Future Work

The NLP+classification models can be helpful in detecting depressed users, but the models only show their success in matching the new users to the previous users they learned. Our results showed that the BERT+LR model is not capturing the language feature of the texts. One possible reason is that the data is highly diluted: only 5% of all users are depressed, and not all depressed users’ tweets contain depression-related expressions. Feeding BERT with such a diluted data set may have hindered the model from picking out certain features. Another possible reason is that BERT cannot capture many features that LIWC can pick up. The features BERT capture might be too abstract to be simply detected by word count based model. And simply adding up single texts might not be able to recreate the features BERT extracted from merged texts. It is also possible that depressed users do not have a common, distinct language pattern from normal users, and we might only be able to match users with users.

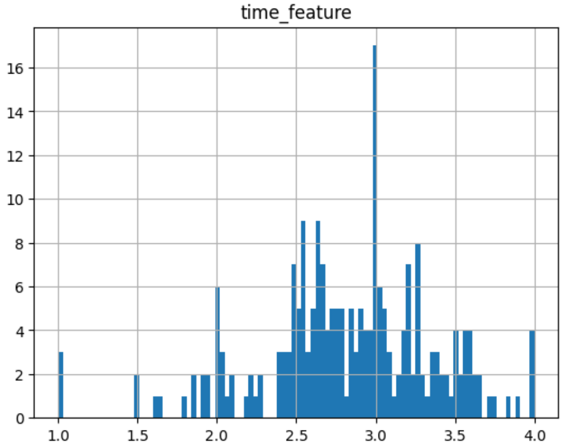


Figure 12. Distribution of feature ’L_times_S’

As for user feature extraction, we discovered that user post time could be engineered into a contributing feature toward depression detection. Further experiments with other different ways to encode this feature could perhaps yield higher performance.

Since word count based LIWC gives more easy-to-interpret results and is less affected by the total number of words included in the text, we propose a direction of future work is to design a mechanism that uses LIWC and user features to screen out suspected text content, then feed them into BERT models for training and testing. In this way, the data gets saturated.

Another future research direction is to investigate if the texts that are labeled positive by BERT+LR model show any sign of symptoms described in the depression self-assessment quiz. Since the self-assessment quiz is a common method used in depression diagnosis, it describes some general problems that most depressed people face. Including the symptoms in the training of depression detection models might increase their generalizability.

References

[1] e. a. Kanter, Jonathan W., “The nature of clinical depression: Symptoms, syndromes, and behavior analysis,” *The Behavior Analyst*, vol. 31, 4 2008. [Online]. Available: www.ncbi.nlm.nih.gov/pmc/articles/PMC2395346/,10.1007/bf03392158

[2] WHO, *Depression and Other Common Mental Disorders: Global Health Estimates.* World Health Organization, 2017.

[3] T. Vos, S. S. Lim, and C. Abbafati, “Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the global burden of disease study 2019,” *The Lancet*, vol. 396, p. 1204–1222, 2020.

[4] K. Hawton, “Risk factors for suicide in individuals with depression: A systematic review,” *Journal of Affective Disorders*, vol. 147, 5 2013. [Online]. Available: 10.1016/j.jad.2013.01.004

[5] E. Yoshikawa, “Factors associated with unwillingness to seek professional help for depression: A web-based survey,” *BMC Research Notes*, vol. 10, 12 2017. [Online]. Available: 10.1186/s13104-017-3010-1

[6] Y. R. Tausczik and J. W. Pennebaker, “The psychological meaning of words: Liwc and computerized text analysis methods,” *Journal of Language and Social Psychology*, vol. 29, pp. 24–54, 12 2009.

[7] G. Shen, J. Jia, L. Nie, F. Feng, C. Zhang, T. Hu, T.-S. Chua, and W. Zhu, “Depression detection via harvesting social media: A multimodal dictionary learning solution,” *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 08 2017.

[8] S. C. Guntuku, D. B. Yaden, M. L. Kern, L. H. Ungar, and J. C. Eichstaedt, “Detecting depression and mental illness on social media: an integrative review,” *Current Opinion in Behavioral Sciences*, vol. 18, pp. 43–49, 12 2017.

[9] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, “Detection of depression-related posts in reddit social media forum,” *IEEE Access*, vol. 7, pp. 44 883–44 893, 2019.

[10] S. Wen, “Detecting depression from tweets with neural language processing,” *Journal of Physics: Conference Series*, vol. 1792, p. 012058, 02 2021.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” arXiv.org, 10 2018. [Online]. Available: https://arxiv.org/abs/1810.04805

[12] R. Martínez-Castaño, A. Htait, L. Azzopardi, and Y. Moshfeghi, “Early risk detection of self-harm and depression severity using bert-based transformers : ilab at clef erisk 2020,” *CEUR Workshop Proceedings*, vol. 2696, 09 2020. [Online]. Available: https://strathprints.strath.ac.uk/72995/

[13] J. Alammar, “A visual guide to using bert for the first time,” 11. [Online]. Available: https://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/

[14] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter,” arXiv.org, 2019. [Online]. Available: https://arxiv.org/abs/1910.01108

[15] J. Bennington-Castro, “Which medication is best for treatment of depression? — everyday health,” EverydayHealth.com, 02 2018. [Online]. Available: https://www.everydayhealth.com/depression/guide/medications/