Understanding Communication Differences on Mental Health and Daily Life Related Subreddits

Anonymous EMNLP submission

Introduction & Related Work

With the rise of internet technology, smart phones, and social media platforms, support groups now rest in our pockets and are only a few clicks away. Online platforms for peer-to-peer support have grown in popularity: a survey report from the Pew Research center estimates that 23% of internet users with chronic health concerns seek peer-to-peer help on the internet (Fox, n.d.). Peer-to-peer support on the internet has become popular, providing unprecedented accessibility and open means for communication without borders (White & Dorman, 2001). Online communities also offer the benefit of anonymity, promoting open and intimate discourse, and people with sensitive issues have been reported to benefit from the anonymity in online communities as compared to face-to-face support groups (Hwang et al., 2010; White & Dorman, 2001).

In addition to unprecedented accessibility for peer-to-peer support, online communities provide an unprecedented amount of public data for computer scientists, computational linguists, and clinical researchers (Chancellor & De Choudhury, 2020). The availability of this data has given birth to a growing number of studies using machine learning to predict the presence of mental disorders using people's text on social media (Chancellor & De Choudhury, 2020). While some of these studies have been focused on making predictions to classify mental health statuses based on people's written language, other studies have focused on more descriptive analyses of the content of these online support communities (Chancellor & De Choudhury, 2020; Park, Conway, & Chen, 2018; Shen & Rudzicz, 2017). Understanding the communication patterns of these online communities will shed light on ways in which the communities can be made more effective (De Choudhury & De, 2014; Park et al., 2018). In the present work, we use a prediction task via interpretable machine learning models as a means of understanding the communication patterns in various types of mental health support communities on reddit. In particular, the present work looks at reddit data from three categories of online support communities: depression related communities, anxiety related communities, and daily life related communities. We collected posts from subreddits in these three categories, and used interpretable machine learning methods to try to classify which type of community a post comes from to investigate whether or not machine learning models can understand communication differences across these three categories and use the knowledge for classification.

The present work will build off of previous work of machine learning models for classifying the type of reddit community that a post comes from. In Shen and Rudzicz (2017), the authors collected text from posts on anxiety related and daily life related subreddits. Then, they trained a logistic regression model on LIWC features derived from the text of posts and classified the subreddit type from posts in the testing data. While they found strong performance, their training data was not strictly before testing data, therefore their systems are not as useful for real-world applications. This work was in 2017, and it remains unknown whether or not the logistic regression model trained on posts from before the COVID outbreak will be effective in classifying posts from after the COVID outbreak.

In Fatima et al. (2019), the authors collected text from posts on depression related and daily life related subreddits. The authors found high accuracy in classifying depression versus controls using a lasso logistic regression model. The robustness of their models is equivocal since they only collected data from 3,000 posts, and similar to Shen and Rudzicz (2017), the testing data was not strictly after the training data in time.

These two previous studies both found strong performance in a binary classification problem. In the present work, we investigated whether or not logistic regression could learn the patterns of communication in a multi-class setting with posts in depression related, anxiety related and daily life related subreddits. Therefore, our work provides an extension of these two previous studies.

In Park et al. (2018), the authors used topical modeling and k-means clustering to compare and contrast the communication differences on anxiety related and depression related subreddits. While they found that depression and anxiety related subreddits often shared discussions about stresses in school, work, issues with sleep, and gratitude, anxiety related subreddits tended to discuss symptoms of anxiety disorders and depression related subreddits tended to talk about the feelings of depression. The present work therefore builds on the work of Park et al. (2018), by investigating whether a logistic regression model can learn relevant differences in communication and then use those differences to make predictions about which type of subreddit a post originates from. Furthermore, unlike the present work, Park et al. (2018) was a 2018 study which did not investigate the differences in communication during COVID era.

In order to determine if machine learning models can understand communication differences amongst subreddits, we conducted three experiments in order to understand the robustness of the multi-class lasso logistic regression for classifying the reddit community type from the post's content. In experiment 1, we aimed to understand how well the multi-class logistic regression model performs when testing data is not strictly after training data in time, an extension of the work of Shen and Rudzicz (2017) and Fatima et al. (2019) to a multi-class setting. Experiment 2 investigates whether training data from the past can predict the subreddit type of future posts, when all data is before the COVID outbreak. Experiment 3 provides a manipulation of experiment 2, in which the models are trained on data from before the COVID pandemic but tested on posts created during the COVID pandemic.

Experiment 1 provides a replication of previous work, and demonstrates that the logistic regression model can also be extended to discriminating between posts from depression and anxiety related reddit communities. In experiment 2, we find the model provides strong performance even when the data splits are controlled for time. In experiment 3, the performance of the model decreased substantially when testing on COVID data, suggesting that the patterns of communication changed across these three communities during the COVID era. As a result, the model did not generalize as well to the posts created during the COVID pandemic.

Data

Collecting Reddit Posts

We used Pushift's API for collecting Reddit data. In each subreddit, all posts every day from December 2012 to April 2021 was collected in the subreddits we used. From each post we collected the following information: date and time, day of week, subreddit, text in post, author name, title of post, group type (depression related, anxiety related or daily life related controls), and the season of the year.

Group	Subreddit
Depression Related Subreddits	r/depression, r/depression_help, r/depressed
Anxiety Related Subreddits	r/Anxiety, r/PanicAttack, r/socialanxiety, r/Anxietyhelp, r/HealthAnxiety
Daily Life Controls Related Subreddits	r/books, r/business, r/movies, r/technology, r/YouShouldKnow, r/Music, r/graphic_design

Figure 1. Subreddits used in each of the three classes.

Pre-Processing Posts

We removed all posts with a deleted author or text. We eliminated special characters because the language detection tool could not handle them. Language detection is essential to ensure all text is in English, so that our feature extraction methods using affect dictionaries work properly. Then, we removed stop words from posts and replaced all web links with "(link here)" in order to make the posts understandable to the language detection tool. Finally, we removed all rows that contain less than 50 words because we only want to include sentences that have a substantial amount of content for us to analyze, and short posts would introduce undesired variability into our observations.

Feature Extraction

Features were extracted via two affect dictionaries: National Research Council Canada Affect Lexicon (NRCLex) and Linguistic Inquiry and Word Count (LIWC) (Tausczik & Pennebaker, 2010). After inputting the text from one post into the NRClex function, a dictionary will be returned that contains specific emotions (such as fear, surprise, joy) as keys and the number of words matching the NRClex dictionary for a selected emotion in the post

as values. We then divided the word counts in affect categories by the total number of words in the post so that longer posts do not unduly have high values for features solely as a consequence of the length. LIWC contains categories that look at parts of speech, emotion, punctuation and numerous psychological categories (Tausczik & Pennebaker, 2010). Given a post, the LIWC application will return multiple new columns representing the number of words in a post that match each of the LIWC categories divided by the total number of words in the post.

Model

All three experiments used lasso regularized multi-class logistic regression, using the l_1 norm penalty, to perform multi-class prediction (Tibshirani, 1996). Logistic regression is extended to the multi-class setting by using the control group as the reference class in two separate sub-model fits. In the first model fit, we model the log-odds of the probability that a post is from a depression related subreddit and not the control group as a linear combination of the features. In the second model fit, we model the log-odds of the probability that a post is from an anxiety subreddit and not a daily life subreddit also as linear combination of the features. Then, in order to perform this multi-class prediction, for a new feature vector we can then derive an estimate for the probability of each class. See Mohamad, Ali, Noor, and Baharum (2016) for other work that put this multi-class logistic regression method to practice. The python statsmodels package was used to perform this multi-class lasso logistic regression. The lasso penalty parameter was tuned on the validation set in each experiment.

Experiments

Experiment 1

Experiment 1 is similar to the work of Shen and Rudzicz (2017) and Fatima et al. (2019), insofar as we are predicting which subreddit type a post comes from, when the training and testing data are all from the same time period. However, the present investigation differs by an extension to the multi-class prediction setting, and more rigorously controlled data set splits than these previous studies, which are described subsequently. In order to expand on the work of Shen and Rudzicz (2017), we investigated the influence of removing the "anxiety" LIWC feature on model performance, which is a feature specifically engineered to capture psychologically relevant words to an anxiety affect.

We address the following questions in experiment 1:

(1) How does multi-class logistic regression perform when posts in the training data are not strictly before testing data in time?

(2) How does the performance change with and without LIWC's "anxiety" feature?

Training & Testing Split

In experiment 1 our training, validation, and testing data are all between 2-8 years ago. Our training, validation and testing sets splits had the following constraints: (1) an approximately equal balance of posts amongst the three classes, (2) the balance across days of the week and seasons is maintained in each of the three sets as it is in the raw data, and (3) a single author cannot appear in more than one of the training, testing and validation set. The first constraint is imposed so that our baseline accuracy is close to chance. The second constraint is imposed to ensure that our data represent the general trends of the respective categories and not just very specific patterns that appear in particular seasons or days of the week. The third constraint is imposed so that we can demonstrate that our systems generalize to new authors under the three classes, rather than just learning patterns of particular authors.

In order to curate our training, validation and testing data split to meet these new constraints, we devised the following algorithm. We created a dictionary mapping every author to a list of their posts. Then, we classified each author into low activity authors with less than 15 posts total, medium activity authors with between 15 and 30 total posts, and high activity authors with more than 30 posts. Next, we sampled authors from each category of low, medium and high to go into training, validation and testing, such that one of training, validation and testing gets all of that authors' posts. Ultimately, this procedure led to a balance across the three subreddit classes, and maintained a balance across days of the week and seasons. We used 68,860 posts in the training data, 24,074 posts in the validation data and 58,342 posts in the testing data.

Results & Discussion

The logistic regression model, in conjunction with LIWC and NRCLex features achieved outstanding performance on the testing data. With the LIWC feature engineered to capture anxiety related words

in text, the multi-class logistic regression model correctly classified the subreddit type of 88.2% of the testing data posts. When this “anxiety" feature was removed, the misclassification error rate only dropped to 85.7%.

Investigating model performance via confusion matrices, we found a large discrepancy between the performance in classifying each of the three classes (see Figures 2 and 3). The true positive number of the controls was the highest by far, with a true positive number around 0.95 both with and without the “anxiety" feature. The true positive numbers for classifying posts from anxiety and depression related subreddits was substantially lower. The confusion matrices indicate a disproportionate number of depression misclassifications when the true class is anxiety and vice versa, suggesting that the model has a harder time discriminating between posts in the anxiety and depression related classes compared to discriminating a post from the controls. The misclassifcation rate of anxiety posts when depression is the true class increases substantially when the “anxiety" LIWC feature is removed, suggesting that the “anxiety" feature is helpful in discriminating between the depression and anxiety related posts. These results suggest that the communication patterns in the anxiety and depression related subreddits have similarities which make it harder for the model to discriminate between the classes. Nonetheless, the true positives of both anxiety and depression subreddit types are high, near 0.80 regardless of whether or not the “anxiety" feature is included, demonstrating that the logistic regression model achieves high performance even in this multi-class setting.

Experiment 1 replicates the strong performance of the logistic regression model in previous work. Experiment 1 demonstrates that this model can be extended to a multi-class setting, and the performance is robust to the removal of the “anxiety" feature, and to ensuring that the same author does not appear in more than one of the training, validation and testing set.

	Prediction		
Truth	Controls	Anxiety	Depression
Controls	0.953	0.023	0.024
Anxiety	0.033	0.820	0.147
Depression	0.028	0.102	0.870

Figure 2. Experiment 1 Normalized Confusion Matrix with “Anxiety" Feature

	Prediction		
Truth	Controls	Anxiety	Depression
Controls	0.940	0.025	0.034
Anxiety	0.036	0.837	0.128
Depression	0.025	0.181	0.794

Figure 3. Experiment 1 Normalized Confusion Matrix without “Anxiety" Feature

Experiment 2

We investigate the following questions in experiment 2:

(1) With the training data strictly before validation and testing data, how does performance on the testing data differ as compared to experiment 1?

(2) Which features are particularly useful for classifying which subreddit type a post comes from?

Training, Validation, & Testing Split

The training data for experiment 2 included posts only from before 2017, while the validation and testing data included posts from 2017 to 2019, intentionally leaving out posts that were created during the COVID outbreak. The splits ensured a balance across three classes (depression, anxiety, controls), days of the week and seasons, but did not control for each author only appearing in one data set. There were 84,084, 33,270, and 33,487 posts included in the training, validation, and testing data sets respectively.

Results & Discussion

In experiment 2, the accuracy of the model’s classification on the testing data only decreased to 86.6% accuracy when being compared to the accuracy of experiment 1. Experiment 2 demonstrates that the model’s performance is robust even when the training data are strictly before the testing and validation data in time, when all data is before the COVID outbreak. Similar to experiment 1, the true positive classifications for the daily life controls group are

better than those for the anxiety and depression related subreddit classes. The accuracy only decreases by 1% when the LIWC “anxiety" feature is removed. While z-score analysis and statistical significance tests reveals that the “anxiety" feature is important in the classification problem, the small 1% accuracy decrease suggests the feature is not imperative to good performance. Subsequent studies may compare the performance with the removal of features correlated with the anxiety feature to further understand what gives rise to the strong performance of the model.

In the experiment 2 model fit including the “anxiety" feature, we analyzed the relevant features by computing z-scores and p-values with the help of the statsmodels package. For the sub-model performing the binary classification on anxiety versus daily life related subreddits, the following features were most important in the classification: anticipation (NRCLex), anxiety (LIWC), health (LIWC), and leisure (LIWC). The model discovered that increased anticipation, anxiety and health was associated with the anxiety related subreddit, Consistent with clinical psychology research which suggests that people with anxiety think about future negative events more than controls, the logistic regression model discovers that a greater frequency of anticipation related words is in posts from anxiety related subreddits (MacLeod & Byrne, 1996). The model’s discovery that health related words were associated with anxiety related subreddits raised the concern that our results were confounded by the fact that we included a health anxiety focused subreddit. In order to investigate this confound, we re-ran the model with all features, including the “anxiety" feature but with the health feature removed. We found that the model still performed strong, with 88.0% accuracy on the testing set, suggesting that the “health" feature was not imperative to the strong performance. The model discovered that increases in leisure associated words were associated with the daily life controls, which makes sense given that our subreddits in the daily life controls pulled from leisure categories.

For the sub-model performing the binary classification on depression versus daily life related subreddits, z-score and p-value analyses suggest the following features were particularly meaningful: the LIWC leisure feature is strongly associated with daily life controls, the LIWC sadness feature is associated with depression related subreddits, and exclamation points were found to be more associated with daily life controls. The high true positives for the controls may be due to the leisure feature’s ability to capture the relevant text patterns in the daily life controls group, and subsequent studies may investigate the model performance with the leisure feature removed to provide insights into the most essential features that give rise to the strong performance.

		Prediction	
	Controls	Anxiety	Depression
Truth			
Controls	0.921	0.028	0.051
Anxiety	0.033	0.849	0.118
Depression	0.025	0.150	0.824

Figure 4. Experiment 2 Normalized Confusion Matrix with “Anxiety" Feature

		Prediction	
	Controls	Anxiety	Depression
Truth			
Controls	0.933	0.030	0.037
Anxiety	0.043	0.803	0.154
Depression	0.024	0.148	0.829

Figure 5. Experiment 2 Normalized Confusion Matrix without “Anxiety" Feature

Experiment 3

We investigate the following question in experiment 3: How does the performance differ when the validation and testing data are all posts from the first year of the COVID pandemic, and the training data are all posts from before the COVID outbreak?

Training & Testing Split

Experiment 3 used the same training data set as experiment 1, which included posts from before 2019. Validation and testing data included posts that were created during the first year of the COVID outbreak. The splits ensured balance across three classes (depression, anxiety, controls), days of the week and seasons, but did not control for each author appearing only in one of the three data sets. There were 68,860, 26,480, 28,762 posts included in the training, validation, and testing data sets respectively.

Results & Discussion

In experiment 3, the accuracy of the model’s classification on the testing data decreased to 77.6%

from 86.6% in experiment 2. This indicates that the model's performance was greatly affected when testing on COVID-era data. Analyzing the confusion matrix for experiment 3, the true positives for daily life controls was higher than the the true positives for the anxiety and depression groups, similar to experiments 1 and 2. Contrasting the confusion matrices of experiment 3 with those of experiments 1 and 2, the true positive number for the controls decreased drastically, by almost 15%, thereby suggesting that the different patterns of communication changed in the COVID-era data.

When the "anxiety" feature was removed, the accuracy of the model had a small decrease of 3.8%, similar to the small decrease in performance found in experiments 1 and 2. Without the "anxiety" feature, the true positives number (TPN) of depression decreased significantly, from 0.71 to 0.596 while the TPN for anxiety increased. This result suggests that the model required the "anxiety" feature to discriminate between the depression and anxiety related subreddits in experiment 3.

		Prediction	
Truth	Controls	Anxiety	Depression
Controls	0.802	0.079	0.119
Anxiety	0.045	0.815	0.140
Depression	0.056	0.234	0.710

Figure 6. Experiment 3 Normalized Confusion Matrix with "Anxiety" Feature

		Prediction	
Truth	Controls	Anxiety	Depression
Controls	0.794	0.084	0.122
Anxiety	0.051	0.821	0.128
Depression	0.048	0.355	0.596

Figure 7. Experiment 3 Normalized Confusion Matrix without "Anxiety" Feature

Ethical Considerations

De Choudhury and De (2014) studied the behavior patterns of users on reddit mental health discourse forums and found that users frequently made "throw-away" accounts to make one anonymous post and then delete their account to ensure that they leave no trace of their identity. De Choudhury and De's 2014 findings suggest that anonymity is a key feature of reddit that attracts users, and possibly allows these forums to be effective in providing peer-to-peer support. If people on reddit mental health discourse forums learned that their posts will certainly make their way to the hard drive of thousands of scientists without any consent, the sense of privacy and anonymity which may make reddit helpful for mental health support could be shattered and studies like ours are not respectful to people's rights to give consent for their data be used for science. This could be addressed by the creation of a reddit consent feature in which people who do not want their data scraped can indicate that in the account settings. Then, the government can enforce laws that only accounts who gave consent can have their data scraped.

Systems like ours may be extended to try to classify people's mood or mental health status based on their text, with a different data set that ensures carefully controlled labels of people's mood and mental health status. Using such systems for targeted advertising poses a severe risk of exploiting vulnerable people and causing a decline in people's mental health by companies who are more concerned about profit than offering products and ideas which are conducive to well-being. Addressing this problem requires a complete transformation of how the internet works and the business models of web industry leaders.

Lastly, machine learning models are error prone and will often incorrectly classify people's mood and mental health status, which can therefore bring about more harm than good by misdiagnosis if the machines predictions are taken too seriously in clinical settings (Benton, Coppersmith, & Dredze, 2017). Uncertainty quantification about predictions can help with this issue, but not completely resolve it. Ultimately, decisions like sending a child to psychiatrist cannot be made by an error prone machine, and must be informed by human reason and human empathy as well.

	With "Anxiety" Feature	Without "Anxiety" Feature
Experiment 1	88.2%	85.7%
Experiment 2	86.6%	85.6%
Experiment 3	77.6%	73.8%

Figure 8. Classification accuracy from the three experiments with and without the "Anxiety" LIWC Feature

General Discussion & Conclusion

Our experiments demonstrated that a multi-class lasso logistic regression model, with LIWC and

NRCLex features, has strong performance in discriminating posts in our three classes in experiments 1 and 2. Experiment 1 provides an extension of Shen and Rudzicz (2017) and Fatima et al. (2019), demonstrating the robustness of the logistic regression model in the multi-class setting, with adequate data points, and with the data set splits controlled for author. Experiment 2 provides a further extension of this previous work, demonstrating the robustness of logistic regression when the training data is strictly before the testing data in time.

Experiment 3 demonstrates that logistic regression does not have robust performance when the training data is pre-COVID outbreak while the testing data is post-COVID outbreak. Experiment 3 suggests that the patterns of communication across the various subreddits before the COVID outbreak have changed during the COVID outbreak. As a result, the model learned on the training data did not generalize as well to the testing data in experiment 3. In order to further evaluate what brought about this discrepancy in performance in experiment 3, future work could take an approach of topic clustering posts similar to (Park et al., 2018). Topic modeling may demonstrate that what we had taken as daily life controls group now communicated more similar to a mental health forum due to the turbulent COVID outbreak. Furthermore, topic modeling may demonstrate a homogeneity in the topics across all the various subreddits, due to the outbreak consuming people's attention.

To further evaluate the communication differences in our three groups, subsequent studies may conduct a performance comparison analysis for feature removals other than the LIWC "anxiety" and "health" features. Such an analysis would demonstrate which features are important to the performance, and therefore provide more insights into how the communication patterns across the three groups differ.

References

Benton, A., Coppersmith, G., & Dredze, M. (2017). Ethical research protocols for social media health research. In *Proceedings of the first acl workshop on ethics in natural language processing* (pp. 94–102).

Chancellor, S., & De Choudhury, M. (2020). Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine*, *3*(1), 1–11.

De Choudhury, M., & De, S. (2014). Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Proceedings of the international aaai conference on web and social media* (Vol. 8).

Fatima, I., Abbasi, B. U. D., Khan, S., Al-Saeed, M., Ahmad, H. F., & Mumtaz, R. (2019). Prediction of postpartum depression using machine learning techniques from social media text. *Expert Systems*, *36*(4), e12409.

Fox, S. (n.d.). *Peer-to-peer healthcare: Many people—especially those living with chronic or rare diseases—use online connections to supplement professional medical advice. report, pew internet and american life project, 28 february 2011.*

Hwang, K. O., Ottenbacher, A. J., Green, A. P., Cannon-Diehl, M. R., Richardson, O., Bernstam, E. V., & Thomas, E. J. (2010). Social support in an internet weight loss community. *International journal of medical informatics*, *79*(1), 5–13.

MacLeod, A. K., & Byrne, A. (1996). Anxiety, depression, and the anticipation of future positive and negative experiences. *Journal of abnormal psychology*, *105*(2), 286.

Mohamad, N. A., Ali, Z., Noor, N. M., & Baharum, A. (2016). Multinomial logistic regression modelling of stress level among secondary school teachers in kubang pasu district, kedah. In *Aip conference proceedings* (Vol. 1750, p. 060018).

Park, A., Conway, M., & Chen, A. T. (2018). Examining thematic similarity, difference, and membership in three online mental health communities from reddit: a text mining and visualization approach. *Computers in human behavior*, *78*, 98–112.

Shen, J. H., & Rudzicz, F. (2017). Detecting anxiety through reddit. In *Proceedings of the fourth workshop on computational linguistics and clinical psychology—from linguistic signal to clinical reality* (pp. 58–65).

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, *29*(1), 24–54.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

White, M., & Dorman, S. M. (2001). Receiving social support online: implications for health education. *Health education research*, *16*(6), 693–707.