Forecast Depression Level and Risk of Suicide

INTRODUCTION

Depression is a common illness worldwide. People with depression could suffer greatly and function poorly in daily life, and it can lead to suicide. Many people experiencing depression tend to express their feelings on social networks. This project uses a dataset containing 500 redditors' posts with 5-label depression classification from subreddit r/depression. Psychiatrists label each reddit user based on posts they sent following the guidelines outlined in Columbia Suicide Severity Rating Scale (C-SSRS). This dataset then has 500 rows and 3 columns. Each row of the dataset represents a user and three columns are the user index, posts under this user and a depression label developed by exports. The labels are "Supportive" - no sign of suicide, "Indicator" - having pessimistic character or has family history of suicide, "Ideation" - having thoughts of suicide, "Behavior" - having historical self-harm or planning to commit suicide, and "Attempt" - Previous known suicide attempt. We would like to use this dataset to forecast depression level and risk of suicide from anylizing texts from a person.

METHOD

The first step is data preprocessing. We tokenize sentences to words from the original dataset.Then, we remove punctuations and change uppercase to lowercase from posts. After that, we use lemmatization and stemming to get an organized dataset. The second step is feature extraction. The feature extraction methods we used are TF-IDF (term frequency-inverse document frequency), word frequency, n-gram, sentiment feature extraction, exclamation marks extraction, problem marks extraction, and capitalized letters extraction. After this step, we can get many useful features. Then, we split the dataset into five sets for five fold cross-validation, 80% data used to train the model, and 20% data used to test the model. Finally, we choose four different models which are Linear Regression, Naive Bayes, Decision Tree, and Random Forest. We would choose one best fit from these four models.

RESULT

Table 1: using PCA reduce dimension to 300, no max feature, no grid search CV

Model Name	Accuracy
Linear regression	0.10
Naive Bayes	0.15
Random Forest	0.31
Decision Tree	0.25

Table 2: after max feature=20, No PCA, random forest and decision tree used grid search CV(using example parameter)

Model Name	Macro F1 Score	Micro F1 Score	Weighted F1 score	Best parameters
Linear regression(no grid)	0.09	0.14	0.09	/
Naive Bayes(no grid)	0.20	0.28	0.26	/
Random Forest	0.21	0.32	0.27	'criterion': 'gini', 'max_depth': 30, 'n_estimators': 38
Decision Tree	0.18	0.33	0.26	'criterion': 'entropy', 'max_leaf_nodes': 3, 'min_samples_split': 2

Table 3: after max feature=50, No PCA, random forest and decision tree used grid search CV(using example parameter)

Model Name	Macro F1 Score	Micro F1 Score	Weighted F1 score	Best parameters
Linear regression(no grid)	0.11	0.15	0.13	/
Naive Bayes(no grid)	0.21	0.25	0.26	/
Random Forest	0.22	0.36	0.30	'criterion': 'gini', 'max_depth': 5, 'n_estimators': 17
Decision Tree	0.19	0.33	0.25	'criterion': 'gini', 'max_leaf_nodes': 13, 'min_samples_split': 2

Table 4: after max feature=100, No PCA, random forest and decision tree used grid search CV(using example parameter)

Model Name	Macro F1 Score	Micro F1 Score	Weighted F1 score	Best parameters
Linear regression(no grid)	0.13	0.19	0.18	/
Naive Bayes(no grid)	0.21	0.24	0.25	/
Random Forest	0.24	0.37	0.30	'criterion': 'entropy', 'max_depth': 55, 'n_estimators': 31
Decision Tree	0.22	0.33	0.28	'criterion': 'gini', 'max_leaf_nodes': 4, 'min_samples_split': 2

RESULT ANALYSIS

We got a dummy classifier score of 0.23. DummyClassifier is a classifier that makes predictions using simple rules. It can be used as a baseline to compare with real classifiers. Then, we used PCA, max feature and grid search to generate our results. Max feature is creating a feature matrix out of the most given number of frequent words across text documents. While using PCA to reduce dimensions to 300, we found that the linear regression model has the lowest accuracy, and the random forest model has the highest accuracy, shown in Table 1. Then, we tried to use max-feature and grid search. When we set the max feature value to 20 and apply grid search to random forest and decision tree, the accuracy for random forest and decision tree are high and close to each other, which means that both models perform well when the max feature is 20. When we set the max feature to 50 and use grid search for the last two models, the f1 score for random forest is the highest. Similarly, if we set the max feature to 100, we found that the random forest model still has the highest f1 scores. Therefore, the random forest model has the best performance among 4 models, but the decision tree's accuracy will approach the random forest or even be higher, when we have a small max feature value. The linear regression model has low scores for all the four cases. Furthermore, our group found a paper that was published this year that uses the same dataset we use. We got 0.37 (Micro F1) and 0.30 (Weighted F1), which is higher than the best result in this paper. However, one way to get a better performance on this dataset can be having more data. The current dataset has 500 data points with 5 labels. Increasing the number of datapoints can lead to better model predictions.

CONCLUSION

In the application, we can let the user enter a social media address or username, and it will then automatically capture the user's social media posts and give their level of depression. This model can help people who are concerned that they may be depressed. If the classification results indicate the possibility of depression, then the user is advised to go to the hospital for further treatment and seek professional advice from their doctor.
This model could help to support doctors in making more structured decisions. Currently, the clinical diagnosis of depression is based on the ICD-10 or DSM-V diagnostic criteria for depression, combined with patient interviews, scales and physician experience.

This approach is only applicable to one-on-one testing and can easily lead to misdiagnosis due to subjective factors such as patient cooperation and physician proficiency. Due to the lack of patient awareness and early screening tools, patients may have reached major depression after treatment.

Therefore, we hope to use this model as an early screening tool to detect depression as early as possible rather than discover depression in the late stage.

REFERENCE

[1] Sun, Peng, X., & Ding, S. (2017). Emotional Human-Machine Conversation Generation Based on Long Short-Term Memory. Cognitive Computation, 10(3), 389–397. https://doi.org/10.1007/s12559-017-9539-4
[2] Dong J, Wei W, (2019)The application of machine learning in depression, 2020, Vol. 28, 266-274, doi: 10.3724/SP.J.1042.2020.00266
[3] J, Trueman, T. E., & A K, A. (2021). Suicidal risk identification in social media. Procedia Computer Science, 189, 368–373. https://doi.org/10.1016/j.procs.2021.05.106