Machine Learning for Psychological Assessments

Sam Spell
Department of Computer Science
University of Missouri
Columbia, MO 65201
sesxr2@missouri.edu

James Tipton
Department of Computer Science
University of Missouri
Columbia, MO 65201
jctppc@missouri.edu

Abstract

An estimated 21.0 million adults in the United States had at least one major depressive episode. This number represents 8.4% of all U.S. adults. The use of machine learning can predict depressive features on social media and other text sources. By analyzing patterns in social media text classified by the model as showing signs of depression, we hoped to extract patterns of text connected to depressive messaging (if they exist). By developing a Support Vector Machine, and training it with depression-related strings from social media, we created a highly accurate and precise model, which we then could use to predict depressive features in other datasets. We used the model to conduct analysis on political messages, movie reviews, and COVID-19 sentiment datasets; we found that depressive features were approximately constant in all of the testing classes, except for political leaning. Compared to a Random Forest model trained in the same fashion, we found a higher prevalence of false negatives and lower prevalence of false positives in classification. This resulted in much lower overall percentages of depressive features among the same datasets, though still with nearly identical ratios.

Keywords SVM, Random Forest, Machine Learning, Depression

Introduction

Depression is a large mental health issue in America with an estimated 21 million adults in the United States having at one major depressive episode. Along with the already high prevalence for Americans, the prevalence of major depressive episodes was higher among adult females (10.5%) compared to males (6.2%) (National Institute of Mental Health, 2023). Depression can also lead to an increased chance of suicide, with depression being the leading cause of more than two-thirds of suicides every year (Chiong et al., 2021). Social media and other opinion sharing platforms could potentially give an insight into the sentiment of writing. The goal of this project is to use the training on depression classification data to gauge whether the model can give an insight into other social media platforms and the potential depressive sentiment. A possible use of the model created in this project is the potential to highlight textual posts on social media as aligning with depressed feelings, and give users a mental health warning.

This was an exploratory project into working with and understanding the application of machine learning tools to string format data, and different classification techniques. One machine learning method implemented in the project was a Support Vector Machine (SVM). This tool was used initially because our group had learned it was used often when it came to machine learning with string-formatted data. The SVM tool was part of the Science Kit library (sci-kit), and was used with the Python language. In order to analyze multiple machine learning methods for a better analysis, the Science Kit Random Forest Classification (RF) machine learning model was also used to train and predict the classification of other data sets.

Method

Method: design

The main model we decided to use was a Support Vector Machine (SVM) model. We also decided on using a Random Forest (RF) model for the sake of comparison, and because RF models excel in areas where SVMs do not. The SVM model type is effective in high-dimensional spaces, making them suitable for tasks involving many features. They also employ a kernel function to capture complex relationships, enabling a less intensive classification. Additionally, SVM models are memory efficient, while RF models are very memory intensive for high-dimensional datasets. RF models utilize ensemble learning which is a combination of learning algorithms in a decision tree to make more accurate and robust predictions, and provides a measure of feature importance, indicating which features have the most significant impact on the predictions. The RF model is also resilient to overfitting, which the SVM may have been susceptible to (i.e. the model gives accurate predictions for training data but not for new data). The SVM is also sensitive to parameter tuning. SVM has several parameters (e.g., kernel type, regularization parameter) that need to be carefully tuned for optimal performance. Improper parameter selection can lead to suboptimal results.

We utilized Python with Science Kit Learn, components from Natural Language Toolkit, and other standard data processing libraries to create and train our models.

Method: procedure

In preparing data for our machine learning model, multiple steps were followed to take string data and format it to where it could be assessed by the machine learning model. The first step in data preparation was ensuring the data only had one feature, consisting of a string, and a classification associated with it. Once that data was chosen and met the previous requirement, stopwords were removed from the string. This process helped to remove words from the strings that generally do not provide insight into the sentiment of a text, as well as making the string length shorter to improve processing and model training time. After removing stop words, a process of stemming and lemmatization was done to the string data. This process can potentially help to fix spelling errors in data, as well as associated similar words together that may have the same root. At this point the data was still stored in a matrix of string variables, which do not always feed into a machine learning model automatically. To ensure that the data would work with multiple machine learning models, the string data was converted into a string frequency vector, which was another tool that was part of the sci-kit library.

With data prepared for a machine learning model, the data was split into a training and testing data set, with a random 33% of the data being removed from the model training. The SVM was trained on the depression string data set using the SVM and RF machine learning models. Once the model was trained, it was used to predict the tests data features, and then compared to the labels associated with these selected test features. In the analysis of the model training, multiple tools were used to assess the interpretability and effectiveness of each model. The different metrics that were chosen were the accuracy, precision,

Results

Both the SVM and the RF model were able to successfully predict depressive texts with high precision. As seen in Table 1, the random forest model had an accuracy (Pearson R^2 correlation coefficient) of 0.951, while the SVM had an accuracy of 0.9506. For the precision of the models, the RF scored higher than the SVM, with a value of 0.986 compared to 0.949. The recall values were reversed, with RF receiving a value of 0.913 while SVM had an overall 0.951. The last metric used was the F1 score, where the RF model had a value of 0.948 and the SVM had a score of 0.950. The confusion matrices of both models showed an even distribution of false positives and false negatives from the SVM model, and higher false negatives and much lower false positives from the RF model (see Table 2)

	SVM Model	RF Model
Accuracy	0.951	0.9511
Precision	0.949	0.986
Recall	0.951	0.913
F1 Score	0.950	0.948

Table 1

SVM and RF model success metrics after training and testing. Results are from Science Kit default functions.

$$\begin{vmatrix} 1281 & 16 \\ 109 & 1146 \end{vmatrix} \qquad \begin{vmatrix} 1281 & 16 \\ 109 & 1146 \end{vmatrix}$$

Table 2

Confusion matrices of the RF model (left) and SVM model (right). Entries (1,1) and (2,2) represent true positives and negatives, respectively; entries (1,2) and (2,1) represent false positives and negatives, respectively.

In predicting the number of strings classified as depressing, we analyzed three different string form data sources. The first data source was a collection of political tweets. When predicted with the trained models, both the RF and the SVM classified a higher percentage of the liberal leaning texts as depressing, as compared to the conservative leaning texts. The second data set that was predicted was a data set on movie reviews from IMDB. The percentages of both the positive reviews and negative reviews classified as depressing were very similar (33.7% negative and 40.1% positive) for the SVM model. Similar results but at a lower percentage were retrieved from the RF model. Similarly, for the third data set, covid tweets classified as neutral, negative, and positive resulted in similar depressive percentages. For the SVM model, the percent classified as depressive for positive, negative, and neutral were 38.2%, 39.2%, and 39.2% respectively (See Figure 1). However, for the RF model, the percent classified as depressive for positive, negative, and neutral were 9.1%, 8.7%, and 9.8% respectively.

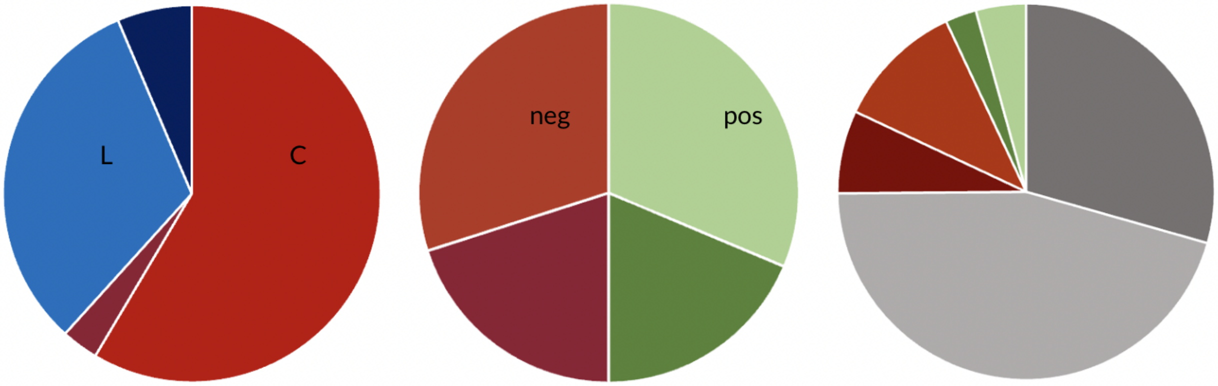

Figure 1

Political Dataset (left), IMDb Dataset (center), and CORD-19 Dataset (right) showing ratios of depressive to non-depressive text.

Discussion

Comparing the analysis of the two machine learning models we utilized, it is important to discuss which was more useful in achieving our goal. The two models had very similar accuracy. The main difference between the two models was that the RF model predicted the depression classification with higher precision. The RF machine learning model made predictions with a 104% increase in precision over the SVM model, though they both had a 95% accuracy. The SVM model was more prone to predicting false positives, whereas the random forest predicted more false negatives. In this case, predicting more false negatives is most likely more meaningful, because it means that less text will get predicted as having depressive features, and some of the barely classified strings may be looked over.

The ratios of depressive features in both the CORD-19 and IMDb datasets were nearly identical within each class of the datasets (i.e. the same ratio of depressive text for positive reviews versus negative reviews, and the same ratio for positive, negative, and neutral coronavirus-related texts). Although, within the two classes of the political dataset, we found a much higher ratio of depression-related text in reddit posts labeled as liberal than those labeled as conservative. As to why this discrepancy exists, we do not yet know, because it's hard to interpret what features of the texts the preprocessing and machine learning models recognize as important.

Inability to interpret the output of the models is a severe disadvantage to working with both SVM and RF models, as we can't discern the important features of depressive text that the preprocessing and model predictions reveal. Future work would be required to accurately discern specific words or phrases that increase probability of depressive sentiment in text.

Future work on this project includes testing the trained model on another string based depression dataset. This we can use to compare the high accuracy of the model to a completely new dataset. The goal of that comparison would be to show the interpretability of the model on new data, which is the use of the project.

References

Chiong, Raymon, et al. “A textual-based featuring approach for depression detection using machine learning classifiers and social media texts.” *Computers in Biology and Medicine*, vol. 135, no. 104499, 2021. *ScienceDirect*, https://www.sciencedirect.com/science/article/abs/pii/S0010482521002936.

Gajare, N. (2017). Liberals vs Conservatives on Reddit (13,000 Posts). Kaggle. https://www.kaggle.com/datasets/neelgajare/liberals-vs-conservatives-on-reddit-13000-posts

InfamousCoder. (2021). Depression Reddit (cleaned). Kaggle. https://www.kaggle.com/datasets/infamouscoder/depression-reddit-cleaned

InfamousCoder. (2021). Mental Health Social Media. Kaggle. https://www.kaggle.com/datasets/infamouscoder/mental-health-social-media

National Institute of Mental Health. (2023, March 8). Major depression. Retrieved from https://www.nimh.nih.gov/health/statistics/major-depression