Sentence Outline

Hugo Pérez de Albéniz, Javier Sánchez, Beatriz Billón

November 18, 2024

1 Introduction

I. Objective

a) Use machine learning techniques to analyze and categorize Reddit posts.

1. Decision Trees classify Reddit messages by creating a series of simple, interpretable rules based on the extracted text features, enabling clear classification.
2. Random Forest combines multiple decision trees to enhance accuracy, reduce overfitting, and capture complex patterns.
3. Neural networks learn complex patterns from data through interconnected layers, making them effective for text classification.

b) The current objective is to classify messages by their subreddit, understanding this to be their category of mental health issues. This could be used in an therapy chat bot system, for example, to give targeted aid to patients tailored to their needs.

II. Data Features

a) Readability indicators: Assess the readability of the posts.

b) TF-IDF representations: Include Term Frequency-Inverse Document Frequency of the most commonly used terms in the posts.

2 Methodology

I. Data Loading and Preprocessing

a) Merged data from multiple CSV files into a single DataFrame.

b) Separated the dataset into features (X) and labels (y), excluding irrelevant columns such as author, date, and post.

c) Performed a train-test split with a 75%-25% ratio.

II. Cost-Complexity Pruning Decision Trees

a) Trained an initial decision tree (min_samples_leaf=0.02).

b) Applied cost-complexity pruning using ccp_alpha to balance bias and variance.

c) Evaluated various values of ccp_alpha through 5-fold cross-validation and visualized the trends in training and validation accuracy.

III. Random Forest

a) Implemented a Random Forest model with an initial configuration of 75 trees, min_samples_leaf=0.05, and max_features=6.

b) Calculated the Out-of-Bag (OOB) error for varying tree counts (20 to 200 trees).

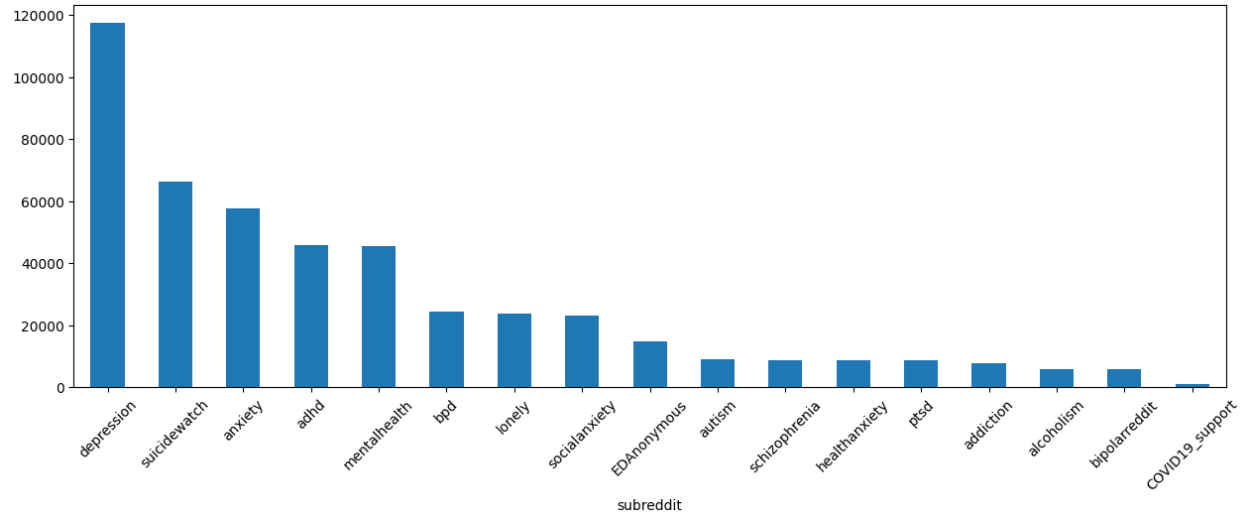
c) Conducted feature importance analysis to identify the most significant predictors.

3 Preliminary Results

I. Post Distribution by Subreddit

Identified very severe class imbalance between the subreddits which needs to be corrected to avoid over-prediction of overrepresented classes.

Number of reddit posts per subreddit:



II. Decision Tree with Pruning

Working with decision trees based on the currently used features (basic NLP that came with the dataset) yielded incredibly poor results. Maximum achieved accuracy of 0.33

III. Random Forest

We decided to ensemble the process with a random forest algorithm but this did not improve accuracy sufficiently.

Computation time was excessive, taking almost 2 minutes to fit each tree, without cross validation, given the large amount of data.

4 Preliminary Conclusions

I. Model Effectiveness

a) The preliminary investigation demonstrates how poorly the used supervised models identify trends in Reddit posts based on the utilised features.

II. Model Optimization

a) Model generalization is enhanced via hyperparameter optimization, such as adjusting ccp alpha and the number of trees.

5 Next Steps

I. Refinement of Models

a) Will refine the models adjusting some parameters

1. ccp_alpha (Decision Tree).
2. Numbers of trees.
3. max_features and min_samples_leaf (Random Forest).

II. Alternative Feature Extraction

Will try alternative feature extraction techniques, that might be more able to capture the meaning of the reddit messages, such as:

1. **Word embeddings** map words to high-dimensional vectors where similar words are closer. Techniques like Word2Vec use neural networks to predict context words, generating embeddings as learned weights.
2. **Bag-of-words** represents text as a frequency distribution of words (ignoring word order). It uses word frequencies for tasks like sentiment analysis, content classification, and spam detection.

This will help us identify the words or linguistic patterns that best distinguish mental health-related posts.

II. Exploring New Algorithms

a) Plan to explore other algorithms

    1. Neural Networks might capture more complex patterns.

b) Will compare this with the current models to assess whether they improve accuracy.

III. Model Comparison

a) Will compare the refined models with the current models to assess whether they improve accuracy.