

I Couldn't Find a Therapist so I tried to Build One - Part II.

Capstone 2 Milestone Report

Kevin Chung
05/03/2021
Springboard

Table of Contents

1. Problem Statement & Project Proposal

- a. Introduction
- b. Project Plan
 - i. Scope
 - ii. Milestones
 - iii. Metrics

2. Data Wrangling & EDA Report

- a. Sourcing Data
- b. Wrangling Data
- c. EDA

Problem Statement & Project Proposal

Introduction

In my previous Capstone project, I identified three core limitations of my study. The following are those three limitations:

- Imbalanced classes & lack of data
- Lack of computation power to try out other hyperparameter settings
- Absence of comparative metrics -- currently only using BERT

For this portion of the Capstone, I will attempt to overcome these limitations in four ways:

1. Web scrape data from mental health support forums for all six classes to obtain more samples.
2. Move the computing environment to a Surface Studio equipped with a cuda GPU unit.
3. Compare results from 8 different non-deep learning models, and apply dimension reduction and two random sampling methods to normalize the data.
4. Fine tune BERT by training an attention head and running predictions.

After these four steps, I will compare all of the prediction results to determine which model performs the most reliably, and also determine which of the models is the most scalable for continued training and deployment.

Scope

This paper will compare several non-deep learning models and determine their efficacy as a multiclass text classifier. After running the initial comparisons, I choose one classifier to tune via grid searching and re-run the predictions. The goal of this process is to identify the best non-deep learning classifier for my dataset and compare its performance against the initial results from Capstone I.

Milestones

I have clearly defined in the above the limitations I must overcome to make this study more robust. The first and foremost milestone is to find and successfully scrape text data from mental health support forums, and preprocess it for exploratory analysis.

Once I have successfully obtained the data, I explore it in the same manner as I have previously in Capstone I, then I utilize slightly more advanced techniques to explore the feature space. The second milestone is achieved after the completion of this EDA.

The third milestone of this portion of the project is the comparison of the different classifiers, and to determine the best non-deep learning classifier for the dataset I introduced.

Metrics

I will leverage Scikit-Learn's classification report to obtain a high-level view of the results for each iteration. Since this is a multiclass classification task, observing the f1 score will be particularly helpful. The other metrics provided by the report will also serve as a good measure for selecting the classifier to tune. Then, I will employ a more visual approach by plotting the ROC AUC curve, as well as the confusion matrix to identify signs of performance issues and overfitting.

Data Wrangling & EDA Report

Sourcing Data

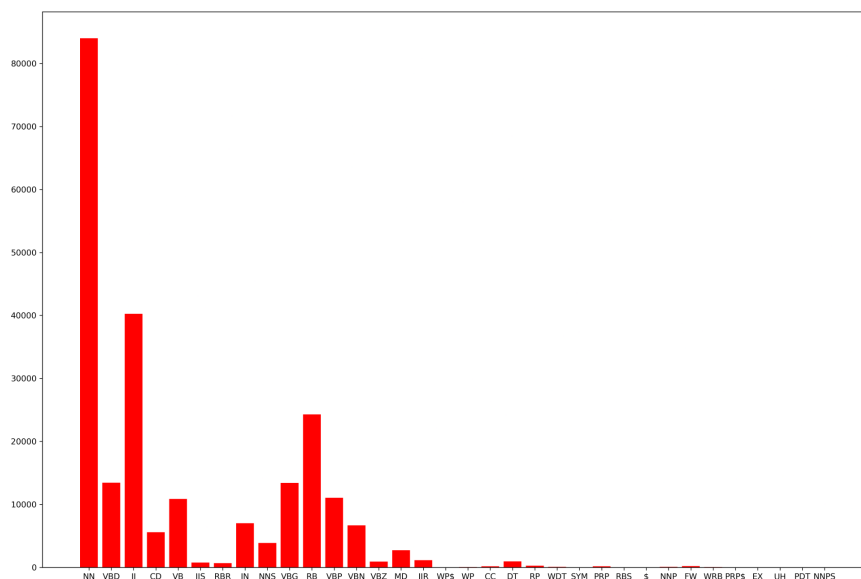
I identified a large mental health support forum called “mentalhealthforum”, which had groups for a variety of mental health conditions. The site was relatively easy to navigate, and it had groups for all six of my classes which I could not find elsewhere. A quick eye scan over the content seemed to be very rich in content; however, the number of available documents differed from one another significantly.

I used BeautifulSoup and requests to obtain the datasets. By iterating over the page count available in each of the forum groups (each group had a different number of pages available) I was able to create a nested for loop which clicked in each page and grab all of the specified text documents within that page. The biggest challenge of this phase was going through the html scripts to identify classes that indicated the correct documents on the web page.

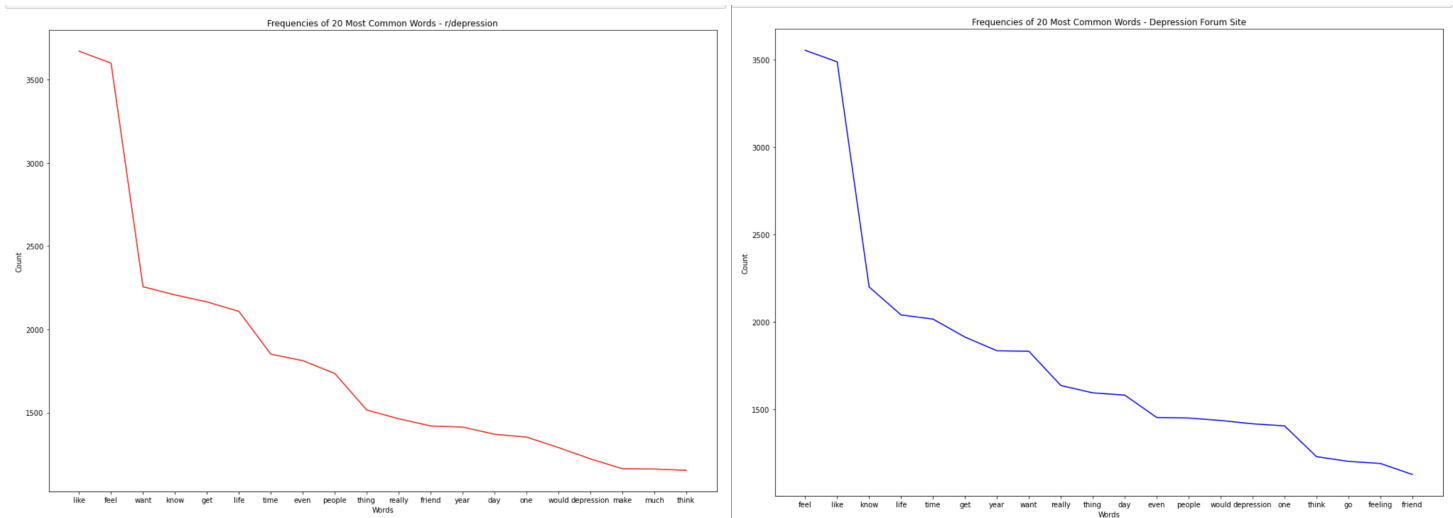
Wrangling Data

I followed the same recipe to prepare the data for exploratory analysis. The steps involved tokenization, removing stopwords, and lemmatization. Once the data had been preprocessed, I proceeded to take a quick look into the corpus for visual insights.

I generated a count of the 20 most frequent words that showed up, searched for concordances with the word “want”, then found quadgram collocations. I also POS tagged every document to visualize choices in diction, and how they vary by class, by plotting each tags’ frequency throughout the dataset. The below are the outputs for the depression forum.



Down to word choice and term frequency ratios, the corpuses' resemblance to each other are uncanny even to the naked eye. I also plotted and compared the top 20 word frequencies from both the subreddits and the forum sites. The below were the results:



The left image is the word frequency from r/depression subreddit, and the right is from the forum site. I repeated this comparison for each of the classes and found similar results.

I then combined the two datasets into a master dataset, called "master_rf.csv". This helped to nearly double the amount of examples available for training, thus taking care of the first of the three limitations.