CS472 Final Project Proposal

Problem: Stress Analysis in Social Media

In today's modern lifestyles, stress permeates every aspect, and when it exceeds healthy limits, it can give rise to a multitude of health problems. Consequently, it becomes imperative to cultivate an awareness of our individual stress levels. In this project, I plan to utilize a dataset obtained from Kaggle, which consists of comments from Reddit related to stress. These comments have been assigned binary labels of 0 or 1. My objective is to explore different machine learning models and assess their performance in predicting individuals' stress levels based on text analysis.

DataSet:

Here is the link to my dataset:

https://www.kaggle.com/datasets/ruchi798/stress-analysis-in-social-media

1. Instead of using all the dataset, I will cut the training data into 3 parts: ⅓ for training set, ⅓ for validation set , ⅓ for testing set
2. I will only use the columns including:
   index(auto index), subreddit, text, label

Model:

1. Text preprocessing for machine learning
   (1)bag of words(BoW)
   (2) TF-IDF
   (3) Word2vec
2. Classification Model
   (1)Logistic Regression
   (2)SVM

More Analysis(if possible): to be continued…