Stress Detection Project using Machine Learning

Name: Karan Das

WEEK 4 Task:

In [6]:
```
import nltk
import re
from nltk. corpus import stopwords
import string
nltk. download( 'stopwords' )
stemmer = nltk. SnowballStemmer("english")
stopword=set (stopwords . words ( 'english' ))

def clean(text):
    text = str(text) . lower()  #returns a string where all characters are lower case. Symbols and Numbers are ignored.
    text = re. sub('\[.*?\]',' ',text)  #substring and returns a string with replaced values.
    text = re. sub('https?://\S+/www\. \S+', ' ', text)#whitespace char with pattern
    text = re. sub('<. *?>+', ' ', text)#special char enclosed in square brackets
    text = re. sub(' [%s]' % re. escape(string. punctuation), ' ', text)#eliminate punctuation from string
    text = re. sub(' \n',' ', text)
    text = re. sub(' \w*\d\w*' ,' ', text)#word character ASCII punctuation
    text = [word for word in text. split(' ') if word not in stopword]  #removing stopwords
    text =" ". join(text)
    text = [stemmer . stem(word) for word in text. split(' ') ]#remove morphological affixes from words
    text = " ". join(text)
    return text
df [ "text"] = df["text"]. apply(clean)
```

```
[nltk_data] Error loading stopwords: <urlopen error [Errno 11001]
[nltk_data]     getaddrinfo failed>
```

In [7]:
```
import matplotlib. pyplot as plt
from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator
text = " ". join(i for i in df. text)
stopwords = set (STOPWORDS)
wordcloud = WordCloud( stopwords=stopwords,background_color="white") . generate(text)
plt. figure(figsize=(10, 10) )
plt. imshow(wordcloud )
plt. axis("off")
plt. show( )
```

```
---------------------------------------------------------------------------
ModuleNotFoundError                       Traceback (most recent call last)
Cell In[7], line 2
      1 import matplotlib. pyplot as plt
----> 2 from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator
      3 text = " ". join(i for i in df. text)
      4 stopwords = set (STOPWORDS)

ModuleNotFoundError: No module named 'wordcloud'
```

In [8]:
```
from sklearn. feature_extraction. text import CountVectorizer
from sklearn. model_selection import train_test_split

x = np.array (df["text"])
y = np.array (df["label"])

cv = CountVectorizer ()
X = cv. fit_transform(x)
print(X)
xtrain, xtest, ytrain, ytest = train_test_split(X, y,test_size=0.33)
```

```
  (0, 7405)     1
  (0, 3278)     1
  (0, 9454)     1
  (0, 861)      1
  (0, 8359)     1
  (0, 3750)     1
  (0, 7214)     1
  (0, 8908)     1
  (0, 298)      1
  (0, 9749)     1
  (0, 4303)     1
  (0, 5034)     1
  (0, 5325)     1
  (0, 2188)     1
  (0, 5118)     1
  (0, 3265)     1
  (0, 2593)     3
  (0, 4188)     1
  (0, 5316)     1
  (0, 3697)     1
  (0, 8339)     1
  (0, 6861)     1
```