Data Collection and Preprocessing Phase

Date	Nov 30, 2024
Team ID	739891
Project Title	Unlocking the Minds: Analyzing Mental Health with NLP
Maximum Marks	6 Marks

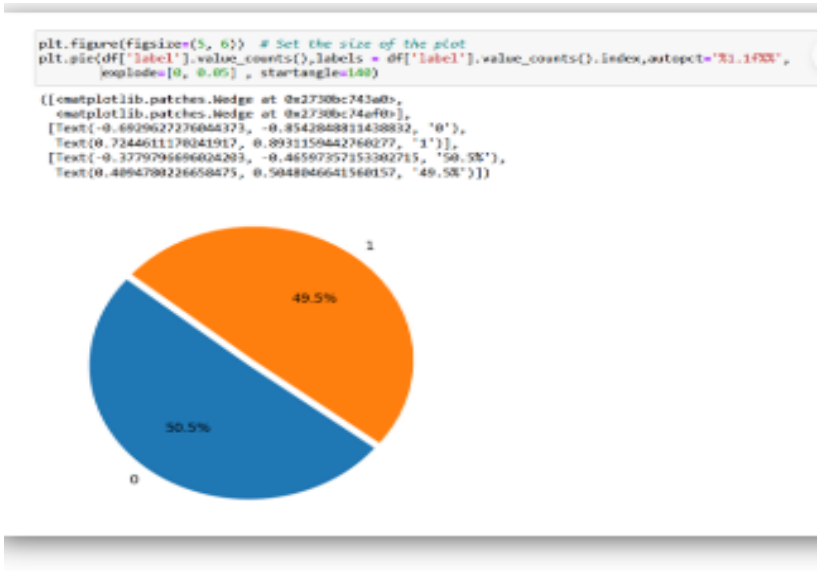
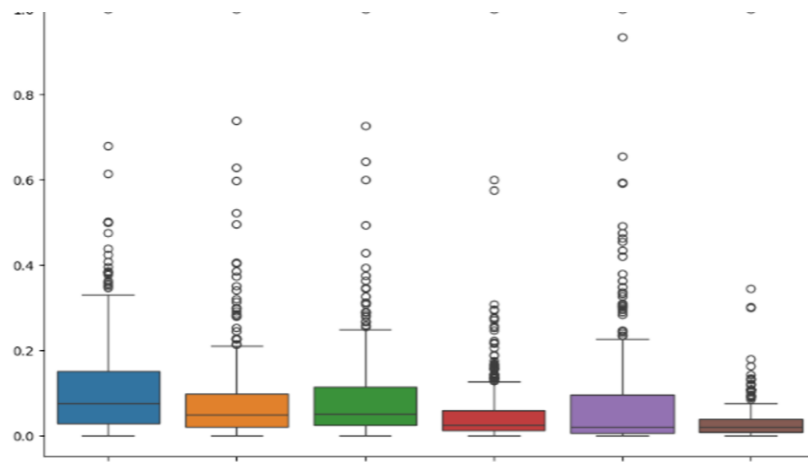
Data Exploration and Preprocessing Template

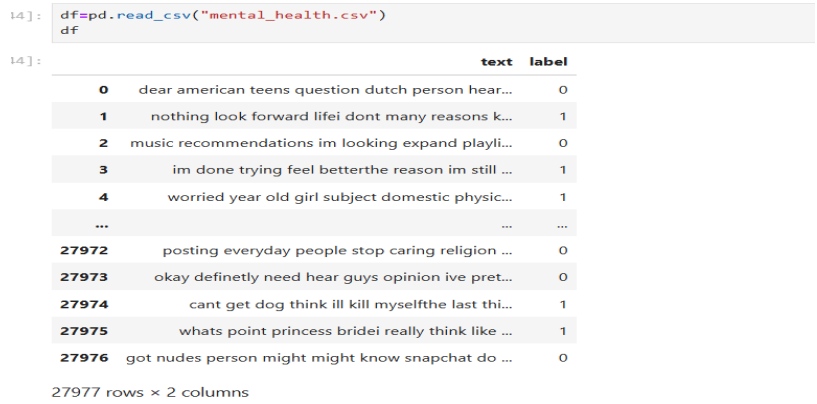
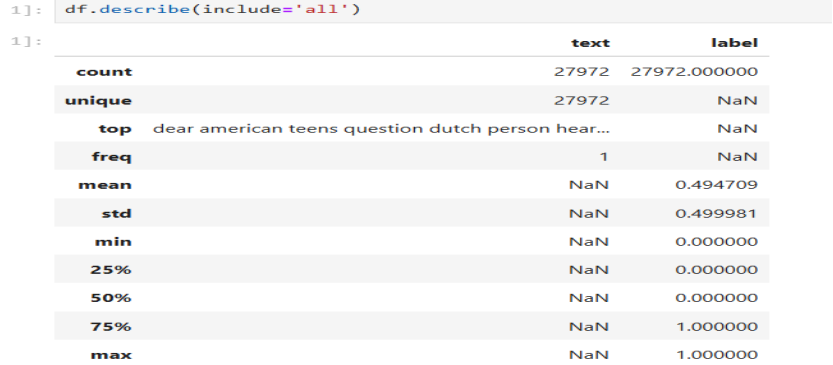
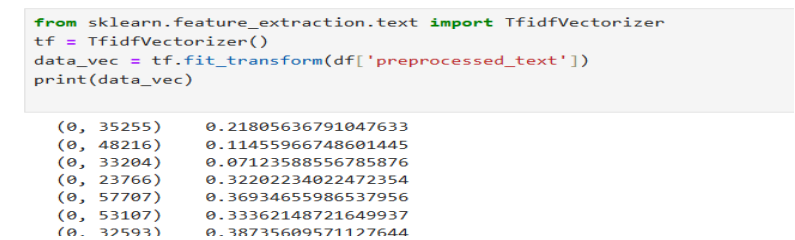
Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

Section	Description
Data Overview	

	text	label
0	dear american teens question dutch person hear...	0
1	nothing look forward lifei dont many reasons k...	1
2	music recommendations im looking expand playli...	0
3	im done trying feel betterthe reason im still ...	1
4	worried year old girl subject domestic physic...	1
...	...	...
27972	posting everyday people stop caring religion ...	0
27973	okay definetly need hear guys opinion ive pret...	0
27974	cant get dog think ill kill myselfthe last thi...	1
27975	whats point princess bridei really think like ...	1
27976	got nudes person might might know snapchat do ...	0

27977 rows × 2 columns

Univariate Analysis	
Bivariate Analysis	To find the relation between two features we use bivariate analysis. Here we don't have 2 numerical columns in our dataset.
Multivariate Analysis	In simple words, multivariate analysis is to find the relation between multiple features. Here we don't have multiple features in our dataset.
Outliers and Anomalies	
Data Preprocessing Code Screenshots	

Loading Data	 14]: df=pd.read_csv("mental_health.csv") df 14]: text label 0 dear american teens question dutch person hear... 0 1 nothing look forward lifei dont many reasons k... 1 2 music recommendations im looking expand playli... 0 3 im done trying feel betterthe reason im still ... 1 4 worried year old girl subject domestic physic... 1 ... ... 27972 posting everyday people stop caring religion ... 0 27973 okay definetly need hear guys opinion ive pret... 0 27974 cant get dog think ill kill myselfthe last thi... 1 27975 whats point princess bridei really think like ... 1 27976 got nudes person might might know snapchat do ... 0 27977 rows × 2 columns
Handling Missing Data	 1]: df.describe(include='all') 1]: text label count 27972 27972.000000 unique 27972 NaN top dear american teens question dutch person hear... NaN freq 1 NaN mean NaN 0.494709 std NaN 0.499981 min NaN 0.000000 25% NaN 0.000000 50% NaN 0.000000 75% NaN 1.000000 max NaN 1.000000
Data Transformation	 from sklearn.feature_extraction.text import TfidfVectorizer tf = TfidfVectorizer() data_vec = tf.fit_transform(df['preprocessed_text']) print(data_vec) (0, 35255) 0.21805636791047633 (0, 48216) 0.11455966748601445 (0, 33204) 0.07123588556785876 (0, 23766) 0.32202234022472354 (0, 57707) 0.36934655986537956 (0, 53107) 0.33362148721649937 (0, 32593) 0.3873560957112764 4
Feature Engineering	Attached the codes in final submission
Save Processed Data	-