Interim Report  
  
We began by outlining our Data Mining group project plan, defining the stages from data preprocessing to model training, evaluation, and a future user application. Then, we imported all the required Python libraries for data handling, visualization, and machine learning.

Next, we moved into Exploratory Data Analysis (EDA). For this, we referenced some websites (<https://www.ibm.com/think/topics/exploratory-data-analysis>, <https://www.kaggle.com/code/nikhil7280/goemotions-basic-eda>) and Naïve Bayes Spam.ipynb file from the first individual project assignment. We loaded three GoEmotions CSV files and inspected their structures to verify row counts and column formats. After confirming the datasets, we merged all three files into one combined dataset, ending up with over 211K text samples.

With the full dataset ready, we examined the distribution of emotion labels. We visualized the frequency of each emotion and observed that *neutral* was the most common label, indicating class imbalance. To better understand sentiment tendencies, we loaded a sentiment-mapping JSON file and grouped the emotions into three sentiment categories: positive, negative, and ambiguous, and plotted their distribution. Finally, we analyzed the length of the text samples, noting that most texts were roughly 70 characters long, giving us insight into input size for modeling.  
  
After EDA, we transitioned into the preprocessing stage. The raw GoEmotions dataset contains many binary emotion annotation columns, so we first identified all emotion indicator fields and wrote a helper function to map each row to its active emotion labels. Using this mapping, we transformed the dataset into a simplified two-column format consisting of text and class, where each text sample is paired with a comma-separated list of its assigned emotions (defaulting to “neutral” when none were present).  
  
With the dataset simplified, we applied a series of text-cleaning steps to prepare the inputs for modeling. We normalized text encoding issues, removed non-English characters, converted text to lowercase, removed URLs and @mentions, stripped emojis, deleted punctuation, removed stopwords, and trimmed surrounding whitespace.