Detailed Team Contributions towards the Final Project

EDS 6346: Data Mining for Engineers

- **Vikram Koti Mourya Vangara**, PSID: 2315018

  Vikram was responsible for implementing Sentence-BERT (SBERT) embeddings to represent the textual data in a high-dimensional semantic space. Additionally, he conducted clustering metric evaluations to assess the performance of clustering algorithms applied to the SBERT-embedded data.

- **Gayathri Seelam**, PSID: 2297215

  Gayathri led the data preprocessing efforts, which included cleaning and preparing the 20 Newsgroups by dataset for analysis. The tasks like converting text to lowercase, removing all non-alphabetic characters and removing extra whitespaces were involved. She also took the lead on compiling and making the final presentation, which included the methodology, results, and interpretations.

- **Neha Reddy Jakka**, PSID: 2296660

  Neha focused on traditional feature extraction by developing the TF-IDF vectorization pipeline. She also implemented the K-Means++ clustering algorithm using both of the TF-IDF vectors and the S-BERT Embeddings, optimizing the clustering performance.

- **Aniketh Bharat**, PSID: 2381419

  Aniketh was in charge of dimensionality reduction and visual analysis. He applied PCA and t-SNE to both TF-IDF and SBERT embeddings to visualize cluster separability in 2D space. Additionally, he led the performance comparison and bringing the overall conclusion. He was also responsible for verifying the quality and consistency of the final Jupyter notebook, ensuring that the code, visualizations, and narrative flowed logically and professionally for submission.