CSE 445 Section 05
Project Group 09
Project Name: Collect a diverse dataset of public social media posts from various platforms. Develop a dynamic clustering model that identifies emerging trends and topics in these posts.

Sadia Islam (2212030042)

Week 1 (Oct 12 – Oct 18)
Dataset Exploration and Selection: In the first week, I focused on exploring multiple open-source datasets to find suitable data for predicting trending posts. I examined several social media text datasets from sources such as Kaggle, GitHub, HuggingFace, and Reddit that contained user posts, captions, and engagement-related information. The main objective was to identify datasets that could represent both trending and non-trending content effectively. After careful analysis, I shortlisted two datasets, Reddit Posts and Combined Captions Cleaned, which aligned well with the project's objective. Both contained high-quality text samples relevant to social media trends. These datasets were merged into a single collection that included post captions and trend-related labels (1 for trending, 0 for not trending). This dataset would later be used to train and test a classification model capable of predicting whether a post could become trending.

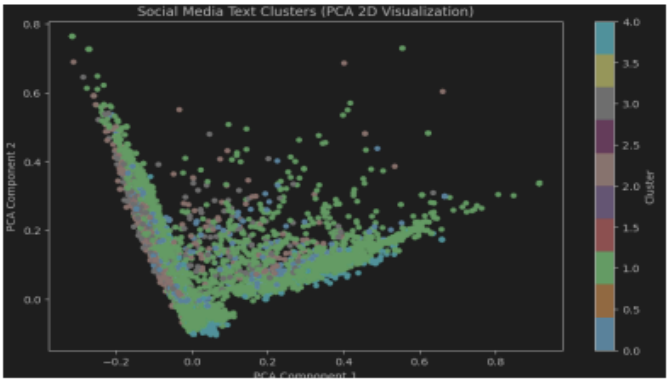
Week 2 (Oct 19 – Oct 25)
Data Cleaning and Preprocessing: During the second week, I concentrated on cleaning and preparing the data for BERT model training. I removed duplicates, filled missing values, and normalized all text data to lowercase to maintain consistency. Special characters, links, and unnecessary symbols were also removed to ensure a clean textual dataset. Next, I divided the dataset into training (80%), validation (10%), and testing (10%) subsets. This ensured that the model could be properly trained and evaluated. Using the BERT tokenizer from the Hugging Face Transformers library, I converted the text data into numerical input IDs and attention masks, which BERT requires. Each sequence was truncated or padded to a fixed length of 128 tokens for consistency.

Week 3 (Oct 26 – Nov 1):
Model Training and Runtime Challenges: In the third week, I began fine-tuning the BERT base model for binary classification using the preprocessed dataset. I used the AdamW optimizer with a learning rate of 2e-5, and the model was planned to train for 3 epochs. However, during the training phase, I encountered several runtime issues in Google Colab, mainly due to limited session time and GPU disconnections. As a result, the training process could not be fully completed across all epochs. The early results from the first epoch showed promising accuracy and stable loss reduction. I saved the intermediate model checkpoints to continue training later without restarting from scratch. I also began testing smaller batches of text samples to verify that the model correctly tokenized and predicted basic trend patterns.

Habiba Alam Raisa (2231272642)

Week-1 (Sep 21 – Sep 27): I was not assigned to any group. I joined a group the following week.
Week-2 (Sept 28 – Oct 4): I collected 2 datasets from Kaggle. One dataset has data from
Instagram and another from Reddit.
Week-3 (Oct 5 – Oct 11): Started data processing. Dropped irrelevant columns. Renamed some
columns and then merged 2 different datasets into one.
Week-4 (Oct 12 – Oct 18): Cleaned the data by removing URLs, punctuations (but I kept
hashtags due to training purpose), extra space, etc.
Week-5 (Oct 19 – Oct 25): Created a balanced dataset by sampling from Reddit and Instagram
data. Used TF-IDF vectorization to convert text data into numerical features.
Week-6 (Oct 26 – Nov 1): Performed K-means clustering on the TF-IDF features.

Week-7 (Nov 2– Nov 5): Tried a 2D visualization of the clusters using t-SNE.



Shakela Mostofa Priya
2132049642

Week 1 : Project Introduction and Familiarization During the first week of the semester, I focused on understanding the course structure and reviewing the preliminary machine learning concepts outlined in the lecture materials. Although the project had not yet been formally assigned, I began revisiting foundational topics such as: Vector operations and probability basics, Evaluation metrics (accuracy, precision, recall, F1-score), Confusion matrices and their interpretation. This preparatory work helped me build a solid foundation for the upcoming project tasks.

Week 2: Project Selection and Topic Research. In the second week, our group was assigned the project, and we selected the topic: "Collect a diverse dataset of public social media posts from various platforms. Develop a clustering model that identifies emerging trends and topics in these posts."

My contributions during this week included: Researching the problem domain and understanding the real-world applications of trend detection in social media studying clustering algorithms, with a focus on k-Means clustering Reviewing related concepts such as: Unsupervised learning Cluster initialization and centroid update methods Evaluation of clustering quality I also began exploring potential data sources and API access points for collecting social media data.

Week 3 : Dataset Collection and Preprocessing During the third week, I took the lead in collecting a diverse dataset of public social media posts. This involved: Identifying and accessing public datasets from platforms such as Twitter and Reddit Using Python libraries like tweepy and praw to gather posts based on hashtags and keywords Collecting metadata such as post content, timestamp, user info, and engagement metrics The dataset currently includes over 10,000 posts from multiple platforms, ensuring diversity in content and context. I also began preprocessing the text data by: Removing special characters, URLs, and stop words Applying tokenization and lemmatization Exploring feature extraction techniques such as TF-IDF and word embeddings.

Week 4 : Model Planning and Initial Setup In the fourth week, I started planning the clustering model implementation. My work included: Finalizing the feature representation for the social media posts Studying Principal Component Analysis (PCA) for dimensionality reduction and visualization Drafting a preliminary workflow for the clustering pipeline: Text preprocessing and feature extraction Dimensionality reduction (if needed) Applying k-Means clustering Evaluating results using the elbow method and silhouette scores I also reviewed the concept of anomaly detection to consider its potential use in identifying outlier trends. Next Steps In the coming weeks, I plan to: Implement the k-Means clustering model using the preprocessed dataset Experiment with different values of k and use the elbow method to determine the optimal number of clusters.

Tanushree Das (22122250452)

Week 1 (Sep21 – Sep27): I was not assigned any group.I joined group-9 on 12 October.
Week 2 (Oct12 – Oct 18): I collected 2 database from Kaggle. Which one is caption data of Instagram and another one is Reddit.
Week 3 (Oct19 – Oct25): Started preprocessing Data and removed missing entries,filtered non-English posts and dropped unnecessary columns and then merged 2 different dataset in one.
Week 4 (Oct26 – Nov01): Cleaned text data by removing (URLS, hashtages, emojis due to my train purpose).Standardized post text for analysis.
Week 5 (Oct02 – Nov05): Documented project progress and findings.My next step train these two dataset and applying text analysis in my future work.