## Proposal Title: The Pulse of Public Opinion: A Sentiment Analysis of Instagram Comments

**Group Members:**

2022-bs-se-123 - Yousma Khayyam 2022-bs-se-073 - Esha Mazhar **Group Number:** 6

**Subject:** Natural Language Processing (NLP)

# Introduction

In today's digital era, social media platforms serve as powerful reflections of society, capturing a wide range of human emotions, thoughts, and opinions. Instagram, as one of the most popular social media platforms, holds a vast repository of user-generated content, including comments that reflect individual and collective sentiments. These comments, which range from casual expressions to deeply personal opinions, offer a unique opportunity to explore public opinion and emotional trends.

Sentiment analysis, a critical aspect of Natural Language Processing (NLP), can be used to understand the emotional tone behind text data. By examining the comments on Instagram posts, we can analyze the underlying emotions, be it positive, negative, or neutral, and gain valuable insights into public perceptions, concerns, and attitudes. This proposal outlines a project that leverages sentiment analysis techniques on Instagram comments to investigate human emotions and public sentiment in a structured and meaningful way.

# Objective

The primary objective of this project is to conduct sentiment analysis on Instagram comments using a variety of NLP techniques. Through this analysis, we aim to:

* Identify the prevailing sentiments (positive, negative, or neutral) in Instagram comments.
* Explore the emotional tone behind comments on posts from public figures, brands, or general users.
* Investigate how sentiment varies across different types of posts (e.g., personal, promotional, or event-based).
* Contribute to the field of social media sentiment analysis and enhance the understanding of online human emotions.

# Scope

The scope of the project includes:

* **Data Collection:** Instagram comments will be scraped using an appropriate API or web scraping techniques (in accordance with Instagram's terms and conditions).
* **Data Preprocessing:** Text data will be cleaned and prepared using tokenization, stemming, and lemmatization techniques to standardize and normalize the text.
* **Feature Extraction:** We will apply methods like Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) to extract meaningful features from the text data.
* **Sentiment Analysis:** Machine learning models such as Support Vector Machines (SVM), Random Forest Classifiers, and deep learning techniques like Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) will be used to classify the sentiments of Instagram comments.
* **Performance Evaluation:** The performance of the models will be evaluated using metrics like accuracy, precision, recall, and F1 score.

# Methodology

This project will follow a systematic approach using the following steps:

## Data Collection:

* + Instagram comments will be collected from a sample of posts using Instagram’s API or web scraping methods. We will target posts from various categories, such as personal accounts, brands, and public figures.

## Text Preprocessing:

* + **Tokenization:** Breaking down the comments into individual words or tokens.
  + **Stemming and Lemmatization:** Reducing words to their root form to ensure uniformity.
  + **Removing Stop Words:** Eliminating common, unimportant words (such as "the," "and," etc.) to focus on meaningful content.
  + **Normalization:** Converting all text to lowercase to avoid treating words like “Good” and “good” as separate terms.

## Feature Extraction:

* + **Bag-of-Words (BoW):** Creating a representation of the text where each word is treated as a feature.
  + **TF-IDF (Term Frequency-Inverse Document Frequency):** A statistical measure used to evaluate the importance of a word in a document relative to the entire corpus.

# Expected Outcomes

* A comprehensive sentiment analysis of Instagram comments, offering a clear picture of the emotional tone behind user interactions.
* Insights into how different types of posts generate different sentiments from audiences.
* A comparison between traditional machine learning techniques (SVM, Random Forest) and modern deep learning methods (RNN, CNN) for sentiment classification.
* A valuable contribution to the growing field of social media sentiment analysis, with potential applications in brand management, social media monitoring, and public opinion research.