**Project Title: "Predicting YouTube Video Popularity with Sentiment Analysis"**

**Members:** Xiao Nan (A69027384), Liang Xinyi (A69027258)

**Objective:**

YouTube videos are becoming increasingly popular for people busy at work seeking entertainment. Generating those with high popularity helps content creators earn much for their livings. In view of this, We aim to develop a machine learning model to predict the popularity of YouTube videos, measured by the "net likes" (difference between likes and dislikes). We will Incorporate sentiment analysis of the video title to explore the impact of emotional tone on viewer engagement. We may also categorize each video to explore the impact of video category on viewers’ preferences using Large Language Models.

**Dataset:**

Utilize the "YouTube Trending Videos" dataset, available on Kaggle. Dataset Link: <https://www.kaggle.com/datasets/datasnaek/youtube-new/>, which gives metadata for the Youtube videos like title and descriptions, as well as its like and dislike counts.

**Tasks:**

**Feature Selection:**

Identify relevant features, including video category, publish time, title length, and sentiment of the title. Calculate the "net likes" as the target variable, representing the overall sentiment towards the video.

**Data Preprocessing:**

Clean and preprocess the data, handling missing values and converting categorical features into suitable formats for machine learning models. Perform sentiment analysis on video titles, classifying them into positive/negative or assigning sentiment scores.

**Regression Modeling:**

Select and implement regression models (e.g., linear regression, random forests) to predict the "net likes" (or days taken to trend) based on the chosen features.

**Model Evaluation:**

Evaluate the model's performance using appropriate metrics, considering both predictive accuracy and interpretability. Implement feature importance analysis to understand the impact of each feature on video popularity.

**Creativity:**

We may Explore creative approaches to enhance prediction accuracy, considering mining more features from the dataset using the prevailing sentiment analysis techniques.

**Output:**

Present the predictive model and analysis in a comprehensive Data-Exploration Report adhering to project guidelines. Include visualizations and insights. Understand the indications of the “net likes” statistics (for example, on how we can infer video quality based on that) and indicate the potential usages of them.