Project Initialization and Planning Phase

Date	JUNE 2024
Team ID	740103
Project Title	The Language Of Youtube: A Text Classification Approach To Video Descriptions
Maximum Marks	3 Marks

Project Proposal (Proposed Solution) template

This project proposal outlines a solution to address a specific problem. With a clear objective, defined scope, and a concise problem statement, the proposed solution details the approach, key features, and resource requirements, including hardware, software, and personnel.

Project Overview	
Objective	The objective of "The Language of YouTube: A Text Classification Approach to Video Descriptions" is to develop a robust and efficient text classification system to analyze and categorize the vast array of video descriptions on YouTube. By leveraging natural language processing (NLP) techniques, this approach aims to automate the classification of video descriptions into predefined categories, such as genre, content type, or target audience
Scope	The scope of "The Language of YouTube: A Text Classification Approach to Video Descriptions" encompasses the exploration and application of text classification techniques to analyze and categorize video descriptions on YouTube like Data Collection, Preprocessing, Feature Extraction etc

Problem Statement	
Description	YouTube stands out as a dominant platform, hosting millions of videos across a vast array of topics and languages. Each video is accompanied by a description, which serves as a crucial component for categorizing and understanding the content. However, the sheer volume and diversity of these descriptions present significant challenges for automated text classification systems. The primary problem is the effective classification of video descriptions into relevant categories based on their textual content
Impact	In "The Language of YouTube: A Text Classification Approach to Video Descriptions," several impactful challenges and issues might be faced: Variability in Language Use, Data Quality Issues, Multilingual and Cross-Linguistic Challenges, Ambiguity and Contextual Interpretation, Ethical Considerations and Bias etc
Proposed Solution	
Approach	To approach the text classification of YouTube video descriptions effectively, you need a systematic methodology that integrates data collection, preprocessing, model training, and evaluation. Here's a structured approach to tackle this task: • **Normalization:** Convert text to lowercase to ensure consistency. • **Tokenization:** Split descriptions into words or tokens to facilitate analysis. • **Removing Noise:** Eliminate URLs, special characters, and unnecessary whitespace. • **Handling Stop Words:** Remove common but noninformative words using stop words lists. • **Stemming/Lemmatization:** Reduce words to their root forms to standardize text. • **Language Detection:** Filter descriptions based on language to ensure relevance to the target language.

Resource Type	Description	Specification/Allocation
Hardware		
Computing Resources	CPU/GPU specifications, number of cores	e.g., 2 x NVIDIA V100 GPUs
Memory	RAM specifications	e.g., 8 GB
Storage	Disk space for data, models, and logs	e.g., 1 TB SSD
Software		
Frameworks	Python frameworks	e.g., Flask , sklearn , metrics
Libraries	Additional libraries	e.g., scikit-learn, pandas, numpy
Development Environment	IDE, version control	e.g., s, Git , spyder, Google co lab
Data		

Key Features	Real-time Prediction: These predictions are made available through an API, allowing integration with dashboards and alert systems for stakeholders. Adaptive Learning: The model will continually learn from new data, improving its accuracy. Scalability: Designed to handle large volumes of transactions without compromising performance.

Resource Requirements

Data	Source, size, format	e.g., Kaggle dataset, 500 images , CSV