TEXT CLASSIFICATION ON QUORA DATASET (USING NLP)

**Abstract:**

This project focuses on classifying questions on Quora using **Natural Language Processing (NLP).** The goal is to identify **duplicate or spam** questions to improve user experience. We will use NLP techniques like tokenization, word embeddings (TF-IDF, Word2Vec), and machine learning models to Analyze question similarities. The model will be trained on a Quora dataset and evaluated for accuracy. This project can help in better content moderation and efficient search results on platforms like Quora.

**INTRODUCTION:**

Every day, people post thousands of questions on online platforms. Some of these questions can be misleading, offensive, or not asked in good faith, which can create a negative experience for others. Checking and removing such questions manually takes a lot of time and effort. This project aims to automate the process by using AI and language processing techniques to quickly identify and filter out inappropriate questions, making online spaces safer and more reliable.

**TECHNOLOGIES USED:**

* Natural Language Processing (NLP)
* Machine Learning (ML)
* Text Classification Models

**PROGRAMMING LANGUAGES:**

* Python
* HTML
* CSS
* Flask (if deployed as a web application)

**TOOLS:**

* NLTK & spaCy (for text preprocessing)
* Scikit-learn (for ML models)
* TensorFlow & Keras (for deep learning models)
* Pandas & NumPy (for data handling)
* Google Colab / Jupyter Notebook (for model training and testing)

**METHODOLOGY:**

* DATA COLLECTION
* DATA PREPROCESSING
* MODEL DEVELOPMENT & TRAINING
* PREDICTION & DEPLOYMENT

**DATA COLLECTION:**

* The dataset consists of Quora questions labeled as sincere or insincere, allowing the model to learn patterns in text classification.
* The dataset is obtained from publicly available Quora datasets for NLP research.

**DATA PREPROCESSING:**

* Removing stop words, punctuation, and special characters.
* Tokenization, stemming, and lemmatization.
* Converting text into numerical form using TF-IDF, Word Embeddings (Word2Vec, GloVe).

**MODEL DEVELOPMENT & TRAINING:**

* A **Logistic Regression / Random Forest** model trained on the dataset.

**PREDICTION & DEPLOYMENT:**

* The trained model predicts whether a given Quora question is **sincere or insincere**.
* The model can be deployed as a web API or integrated into an existing system for content moderation.

**RESULTS:**

* The NLP-based model successfully classifies questions with **high accuracy**, making it useful for automated moderation.
* It helps detect **harmful or misleading** questions more efficiently than manual moderation.

**CONCLUSION:**  
This project provides an automated and scalable solution for detecting insincere questions using NLP. By implementing text classification techniques, platforms can improve content quality, reduce moderation efforts, and enhance user experience.