NLP Experiment 2

Aim: Study Various Applications of NLP and Formulate the Problem Statement for Mini Project

Theory:

1. Introduction to Natural Language Processing (NLP)

Natural Language Processing (NLP) is a field of Artificial Intelligence that focuses on the interaction between computers and human language. It enables machines to understand, interpret, and generate human language in a meaningful way. NLP plays a crucial role in many real-world applications by processing large amounts of natural language data and extracting valuable information from it.

2. Applications of NLP

NLP has a wide range of applications, many of which are embedded into everyday technologies:

- **Machine Translation**: Automatic translation of text from one language to another, such as Google Translate. It involves complex techniques to maintain the context, grammar, and nuances of the original text.
- **Text Categorization**: Automatically categorizing or classifying text into predefined categories (e.g., email spam filtering, news classification). It uses methods such as supervised learning to label documents.
- **Text Summarization**: Creating concise summaries of long pieces of text while preserving the core message. Summarization can be extractive (selecting important parts) or abstractive (generating a new summary).
- **Chatbot Development**: Virtual assistants that simulate human conversation to provide customer service, technical support, or personal assistance. Modern chatbots leverage deep learning models to understand user input and deliver relevant responses.
- **Plagiarism Detection**: Detecting similarities in text content to ensure originality, widely used in academic and professional environments. NLP-based models can identify reworded text and paraphrased content.
- **Spelling & Grammar Checkers**: Systems that detect and correct spelling mistakes, grammatical errors, and syntactical issues in text. These systems use both rule-based and machine-learning approaches for improving text quality.
- **Sentiment/Opinion Analysis**: Extracting and analyzing opinions or sentiments expressed in text, widely used in market analysis, customer reviews, and social media monitoring.

- **Question Answering Systems**: Answering user queries by retrieving relevant information from databases or documents. Such systems range from simple FAQ retrieval to complex models that provide specific answers to complex questions.

- **Personal Assistants**: NLP-based digital assistants like Siri, Alexa, and Google Assistant help users perform tasks through voice commands, such as setting reminders, searching the web, or controlling smart home devices.

- **Tutoring Systems**: AI-driven tutoring platforms that provide personalized learning experiences, assess student progress, and offer tailored educational content.

3. Literature Review of Recent Advancements in NLP

In recent years, NLP has witnessed significant advancements due to the rise of deep learning models like transformers. Key innovations include:

- **Transformers and Attention Mechanisms**: The introduction of the transformer architecture (e.g., BERT, GPT) revolutionized NLP by improving model performance in tasks like translation, summarization, and question answering. Attention mechanisms allow the model to focus on different parts of the input data, improving contextual understanding.

- **Pre-trained Language Models**: Large pre-trained models, such as GPT-3 and BERT, have become the foundation of many NLP applications. These models are trained on vast datasets and can be fine-tuned for specific tasks, drastically reducing the need for task-specific training data.

- **Transfer Learning in NLP**: Transfer learning, where a model trained on one task is fine-tuned for another, has been highly effective in NLP. This allows smaller datasets to benefit from pre-trained knowledge, leading to more robust solutions.

- **Multilingual Models**: New models capable of handling multiple languages simultaneously, such as mBERT, enable multilingual machine translation, cross-lingual tasks, and more without requiring language-specific data.

4. Problem Statement for Mini Project

Based on the study of various applications, the mini-project will focus on **developing a text summarization tool** that can generate concise summaries of legal documents. Legal texts tend to be lengthy, and summarization tools can help extract key information efficiently for legal professionals. The project will aim to explore both extractive and abstractive summarization techniques using pre-trained models like BERT or GPT-2.

Objective: To create an efficient and accurate text summarization tool for legal documents using state-of-the-art NLP techniques, with a focus on improving the quality of the generated summaries.

Our problem statement for the NLP project is a Text classification problem. We have a dataset which consists of:

- Synopsis of a movie
- Genre of the movie

The synopsis for each movie will labelled with its genre. We aim to develop a model that can correctly predict the Genre of a movie when given a synopsis.

File Name: train.csv

Details:

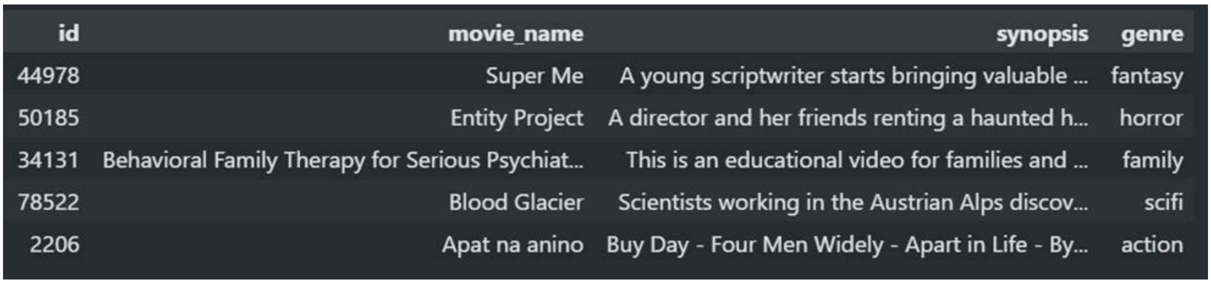
Number of rows: 54,000

Columns:
Id, movie, name, synopsis

Column Description:
id: ID of the movie
movie_name: Name of the movie
genre: Genre of the movie

Primarily the “synopsis” column will be used for all NLP experiments while the “genre” column will be used for labelling the data in the NLP Mini project.

Sample data:



id	movie_name	synopsis	genre
44978	Super Me	A young scriptwriter starts bringing valuable ...	fantasy
50185	Entity Project	A director and her friends renting a haunted h...	horror
34131	Behavioral Family Therapy for Serious Psychiat...	This is an educational video for families and ...	family
78522	Blood Glacier	Scientists working in the Austrian Alps discov...	scifi
2206	Apat na anino	Buy Day - Four Men Widely - Apart in Life - By...	action

5. Conclusion

NLP has become a cornerstone of modern AI applications, with use cases in various domains such as language translation, customer service, and legal assistance. This project will contribute to the growing field by tackling the challenge of text summarization for legal texts, enabling users to process information faster and more effectively.