Experiment No.1
Study various applications of NLP and Formulate the Problem Statement for Mini Project based on chosen real world NLP applications
Date of Performance:
Date of Submission:

Aim: Study various applications of NLP and Formulate the Problem Statement for Mini Project based on chosen real world NLP applications.

Objective: Understand the different applications of NLP and their techniques by reading and critiquing IEEE/ACM/Springer papers.

Theory:

1. Machine Translation

Machine translation is a process of converting the text from one language to the other automatically without or minimal human intervention.

2. Text Summarization

Condensing a lengthy text into a manageable length while maintaining the essential informational components and the meaning of the content is known as summarization. Since manually summarising material requires a lot of time and is generally difficult, automating the process is becoming more and more popular, which is a major driving force behind academic research.

Text summarization has significant uses in a variety of NLP-related activities, including text classification, question answering, summarising legal texts, summarising news, and creating headlines. Additionally, these systems can incorporate the creation of summaries as a middle step, which aids in shortening the text.

The quantity of text data from many sources has multiplied in the big data era. This substantial body of writing is a priceless repository of data and expertise that must be skillfully condensed in order to be of any use. A thorough investigation of NLP for automatic text summarization has been necessitated by the increase in the availability of documents. Automatic text summarising is the process of creating a succinct, fluid summary without the assistance of a human while maintaining the original text's meaning.

3. Sentiment Analysis

Sentiment analysis, often known as opinion mining, is a technique used in natural language processing (NLP) to determine the emotional undertone of a document. This is a common method used by organisations to identify and group ideas regarding a certain good, service, or concept. Text is mined for sentiment and subjective information using data mining, machine learning, and artificial intelligence (AI).

Opinion mining can extract the subject, opinion holder, and polarity (or the degree of positivity and negative) from text in addition to identifying sentiment. Additionally, other scopes, including document, paragraph, sentence, and sub-sentence levels, can be used for sentiment analysis.

Businesses must comprehend people's emotions since consumers can now communicate their views and feelings more freely than ever before. Brands are able to listen carefully to their customers and customise their products and services to match their demands by automatically evaluating customer input, from survey replies to social media chats.

4. Information Retrieval

A software programme that deals with the organisation, storage, retrieval, and evaluation of information from document repositories, particularly textual information, is known as information retrieval (IR). The system helps users locate the data they need, but it does not clearly return the questions' answers. It provides information about the presence and placement of papers that may contain the necessary data. Relevant documents are those that meet the needs of the user. Only relevant documents will be pulled up by the ideal IR system.

5. Question Answering System (QAS)

Building systems that automatically respond to questions presented by humans in natural language is the focus of the computer science topic of question answering (QA), which falls under the umbrella of information retrieval and natural language processing (NLP).

Abstract:

This NLP project aims to perform sentiment analysis on Amazon US reviews specifically for Mobile Electronics products using the DistilBERT language model. Sentiment analysis is the process of determining the emotional tone behind a piece of text, in this case, customer reviews. By employing DistilBERT, a powerful transformer-based language model, we seek to extract valuable insights from the reviews to understand customers' sentiments towards various mobile electronics products available on Amazon US. The results of this analysis will help businesses and consumers gain a better understanding of customer feedback and make informed decisions based on the sentiment expressed in these reviews.

Methodology:

1. Data Collection: We will collect a large dataset of Amazon US reviews for Mobile Electronics products. The dataset will include various attributes such as product ratings, review body(text), and other relevant metadata.
2. Data Preprocessing: The collected data will undergo preprocessing steps, including text cleaning, tokenization, and removal of stop words and special characters. Additionally, we will handle any missing or erroneous data points to ensure the quality and integrity of the dataset.
3. DistilBERT Model: We will utilize the DistilBERT language model, a lightweight version of BERT, to conduct the sentiment analysis. DistilBERT offers similar performance to BERT while being computationally more efficient, making it suitable for this project.
4. Fine-Tuning: To make the DistilBERT model task-specific, we will fine-tune it on our Mobile Electronics reviews dataset. This process involves adjusting the model's weights using the labeled sentiment data to optimize its performance for sentiment analysis.
5. Training and Evaluation: We will split the dataset into training and testing sets to train the fine-tuned DistilBERT model. The model's performance will be evaluated on the testing set using metrics to assess its effectiveness in sentiment classification.

6. Sentiment Analysis: Once the model is trained and evaluated, we will apply it to new, unseen reviews for Mobile Electronics products to predict their sentiment. The sentiment will be classified into positive or negative categories.

Technical Terms:

1. Sentiment Analysis: The process of determining the sentiment expressed in a piece of text, i.e., whether it is positive or negative.
2. Transformer: A transformer is a deep learning architecture that utilizes self-attention mechanisms to process sequential data efficiently, revolutionizing various natural language processing tasks.
3. Hugging Face: A popular organization that develops and maintains a repository of pre-trained language models, including DistilBERT, making it accessible and easy for everyone to use NLP models for various tasks.
4. DistilBERT: A lightweight, distilled version of the BERT (Bidirectional Encoder Representations from Transformers) language model, known for its efficiency and competitive performance in natural language processing tasks.
5. Tokenization: The process of breaking down text into smaller units called tokens, which can be words, subwords, or characters, to facilitate language model processing.