Aim: Literature for Natural Language Processing Application.

Objective: To develop design and analysis ability in students to develop the NLP application in real world scenario by studying a recent Research journal paper
Also develop technical writing skills in students.

Theory:
This assignment asks student to study and understand recent journal paper which is based on application in real world problems.
Write your own report on paper which you have studied.

Title: Large Language Models and their comparison
Aim: The aim of this literature survey is to provide a comprehensive review of large language models in natural language processing. It will focus on analyzing the architectures, training techniques, and applications of prominent models such as GPT and BERT. The survey aims to compare the performance, limitations, and impact of these models, while identifying emerging trends and open challenges in the field.

Introduction:
LLMs are based on neural networks, which are a type of machine learning algorithm that can process data quickly and accurately. They are trained on large datasets of text, such as books, articles, and websites, and use this data to learn patterns and relationships between words and phrases. This allows them to generate natural-sounding sentences and paragraphs that are similar to those written by humans. Unlike other AI technologies, LLMs do not require pre-programmed rules or labeled data, which makes them more flexible and adaptable to different tasks.

LLMs work by using a type of neural network called a transformer, which is designed to process sequential data such as text. The transformer consists of multiple layers of processing units that can learn patterns and relationships between words and phrases in a text dataset. During training, the transformer is fed large amounts of text data and learns to predict the next word in a sentence based on the words that came before it. This process allows the transformer to learn the underlying structure of language and generate natural-sounding text. Once the LLM is trained, it can be used to generate text in response to a given prompt or question, or to perform other natural language processing tasks such as sentiment analysis or language translation.

Model Architecture:
BERT (Bidirectional Encoder Representations of Transformers):
BERT, short for Bidirectional Encoder Representations from Transformers, is a groundbreaking language model introduced by Google in 2018. It is a deep learning model based on the transformer architecture, which revolutionized the field of Natural Language Processing (NLP) by enabling efficient bidirectional language representation learning. The key innovation of BERT lies in its ability to capture contextual information by considering

both the left and right context of each word in a sentence. Architecture of BERT can be divided into following blocks for better understanding:

Transformer Encoder: At the heart of BERT is the transformer encoder, which consists of multiple layers of self-attention mechanisms and feed-forward neural networks. Each layer processes the input text through a series of attention and feed-forward sub-layers.
Tokenization: BERT tokenizes the input text into smaller subword tokens using WordPiece tokenization. This helps handle out-of-vocabulary (OOV) words and reduces the vocabulary size.
Input Representations: BERT takes variable-length text as input, and to ensure consistent input size, it adds special tokens: [CLS] (classification) at the beginning and [SEP] (separator) between sentences. Additionally, for sentence-pair tasks, it separates the sentences using the [SEP] token.
Pre-Training Objectives: BERT is pre-trained on a large corpus of text using two unsupervised learning tasks:
Masked Language Modeling (MLM): Randomly masks some words in the input sentences and tasks the model to predict the masked words based on the surrounding context. This encourages BERT to understand the bidirectional context of words.
Next Sentence Prediction (NSP): For sentence-pair tasks, BERT is trained to predict if the second sentence follows the first in the original text. This helps BERT understand the relationships between sentences.
Transformer Layers: Each transformer layer in BERT comprises self-attention and feed-forward neural network sub-layers. The self-attention mechanism enables each word to attend to all other words in the sentence, capturing both contextual and syntactic information.
Attention Masking: To ensure that words can only attend to preceding words in the transformer's self-attention mechanism, BERT employs masking. It masks out attention to tokens in the future (right-side context) during pre-training and fine-tuning.
Embedding Concatenation: During pre-training, BERT learns contextual embeddings for each token in the input sentence. These embeddings are concatenated and fed into the next stage for downstream tasks.
Pooling and Classification: For classification tasks, BERT uses the [CLS] token's final hidden state as an aggregated representation of the whole sentence. It then applies a classification layer on top of this representation for specific task predictions.

text-davinci 003:

text-davinci-003 is one of the models in the OpenAI API that can understand and generate natural language or code. It is part of the GPT-3 family, which are large multimodal models that can solve difficult problems with greater accuracy than any of our previous models, thanks to their broader general knowledge and advanced reasoning capabilities.
text-davinci-003 is an improved version of text-davinci-002, which was one of the InstructGPT models that could handle natural language instructions. text-davinci-003 has the following improvements over text-davinci-002:

It produces higher quality writing. This will help your applications deliver clearer, more engaging, and more compelling content.
It can handle more complex instructions, meaning you can get even more creative with how you make use of its capabilities now.

It's better at longer form content generation, allowing you to take on tasks that would have previously been too difficult to achieve.
text-davinci-003 has a maximum token limit of 8,192 tokens, which means it can generate up to about 2,000 words at a time. It was trained on data up to September 2021. You can access text-davinci-003 through the OpenAI API and Playground at the same price as the other Davinci base language models.

Comparison between BERT and GPT-3 models:
Their architecture distinguishes GPT-3 from BERT in a prominent manner. As previously indicated, GPT-3 follows an autoregressive approach, whereas BERT employs a bidirectional strategy. While GPT-3's predictive process accounts solely for the left context, BERT takes into consideration both the left and right context. This distinction renders BERT more suitable for tasks like sentiment analysis and natural language understanding, where grasping the complete context of a sentence or phrase is pivotal.

Another dissimilarity between these two models pertains to their training datasets. Although both models underwent training with extensive text data from sources like Wikipedia and books, GPT-3's training incorporated a whopping 45TB of data, while BERT was trained on 3TB. Consequently, GPT-3 enjoys access to a greater volume of information compared to BERT. This could confer advantages in particular tasks such as summarization or translation, where abundant data can prove advantageous.

Lastly, there exists a divergence in terms of model size. While both models boast substantial sizes (GPT-3 with 1.5 billion parameters and BERT with 340 million parameters), GPT-3's magnitude significantly surpasses its predecessor due to the significantly larger training dataset size (470 times larger than the one used for training BERT).

Comparison between BERT and GPT-3 in terms of their capabilities:
Both GPT-3 and BERT have demonstrated strong performance across a range of NLP tasks, including question answering, summarization, and translation. The level of accuracy varies based on the specific task.

However, GPT-3's advantage lies in certain tasks like summarization or translation, where its larger training dataset provides it an edge. This increased data availability bolsters its performance over its predecessor.

Conversely, BERT excels in tasks such as sentiment analysis or NLU. Its bidirectional approach enables it to consider both left and right context when making predictions. In contrast, GPT-3's prediction scope is limited to the left context for words or phrases in a sentence.

Conclusion:
In essence, GPT-3 and BERT have established their significance as valuable resources for executing diverse NLP tasks, displaying different levels of precision. Nevertheless, their contrasting architectures and training dataset sizes render them more effective for distinct tasks. To illustrate, GPT-3 excels in tasks like summarization or translation, whereas BERT's

strengths lie in sentiment analysis or NLU. Ultimately, the selection between these two models hinges on the particular requirements of your task and the goals you aim to achieve.