Name: Lance Abhishek Raj **Email: lanceabhishek321@gmail.com**

Introduction :

Fine-tuning pretrained models like GPT-3 and BERT has revolutionized the field of natural language processing (NLP) by providing state-of-the-art performance on a wide range of language-related tasks. Pretrained models are trained on massive amounts of text data using deep learning techniques, and they can be fine-tuned to perform specific NLP tasks by further training them on a smaller, task-specific dataset. Fine-tuning allows these models to adapt to new contexts and tasks, and has led to significant advances in areas such as language translation, sentiment analysis, and question-answering. In this introduction, we will explore the basics of fine-tuning pretrained models like GPT-3 and BERT, their advantages, and their potential applications in NLP. The task that we will be performing is text classification on AG news Dataset.

Data set used :

The AG News dataset is a collection of news articles, labeled into four categories: World, Sports, Business, and Science/Technology. The dataset contains approximately 120,000 news articles from over 2,000 news sources. It was originally created for research in the field of natural language processing (NLP) and has since become a widely used benchmark for text classification tasks.

The AG News dataset is relatively balanced, with each category containing roughly the same number of articles. The articles were collected from a variety of sources and cover a broad range of topics within each category. The dataset is available for free and has been used by researchers and machine learning practitioners for a variety of purposes, including training and evaluating text classification models.

Overall, the AG News dataset is a valuable resource for those interested in NLP and text classification. Its balanced distribution of categories and large size make it an ideal benchmark for evaluating machine learning models on text classification tasks.

Preprocessing Steps:

1. The raw dataset included 3 columns “Title” of type str , “Description” of type str and “Class Index” of int. Here we first combined Title and Description and named that column “Summary”.
2. Converting Them into word tokens and Removal of punctuations form the “Summary” Column along.
3. Along with removal of punctuations even stop words were removed. NLTK corpus stopwords was used.

About the Architecture/model and fine tuning :

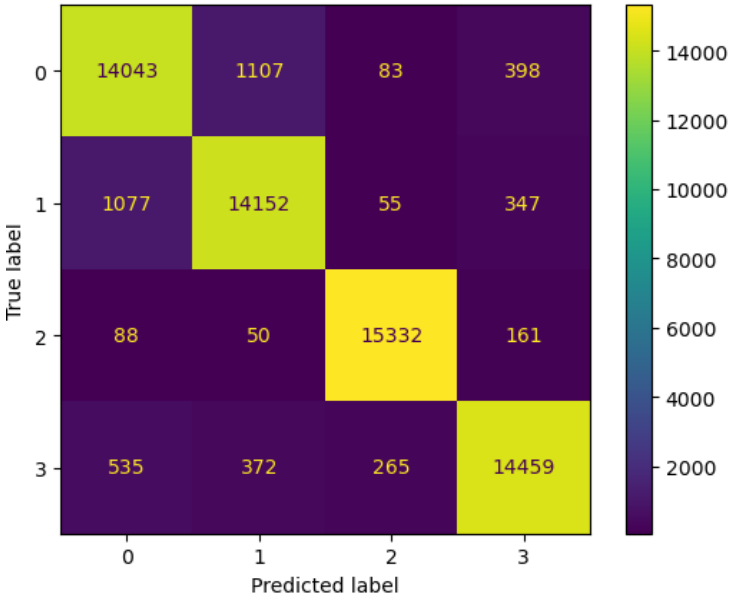
The BERT model is a pre-trained language model developed by Google that uses a transformer architecture to process text in both directions and learn general language representations. Hugging Face provides a range of pre-trained language models, including the BERT-based-uncased model, which is trained on a large corpus of text using an uncased vocabulary. The model is trained to predict missing words in sentences and determine whether two sentences are logically connected. It can be fine-tuned on specific NLP tasks by adding a task-specific output layer and training on a task-specific dataset, allowing it to adapt to specific nuances and vocabulary while still leveraging its pre-trained language representations.

The BERT (Bidirectional Encoder Representations from Transformers) model was trained on a large corpus of text using a technique called "masked language modeling" and "next sentence prediction". The uncased version of BERT uses the same training process as the cased version, with the only difference being the vocabulary used.

Results and Evaluation Metrics:

Accuracy is enough to get the performance of the model as the dataset class distribution was fully even. The performance metrics noted were. As follows for bert uncased model.

Confusion Matrix:



Class	Business	SCI/tech	Sports	world
precesion	0.8920155	0.90249346	0.97438831	0.94103482
recall	0.89840701	0.90538033	0.98087135	0.92502079
F score	0.89519985	0.90393459	0.97761908	0.93295909
Support	15631	15631	15631	15631

From the Above test results we can see the f1 score of each class is above 90% or close to 90% but there are few classes that can be improved. Like Business as it's not like other class above 90.

How to Improve the model/suggestion:
The Model can be Improved with the help of active learning. Pool based active learning.
In this process we select the most informative samples while training. Instead of randomly selecting samples. This helps us in building a model with accuracy with less samples.

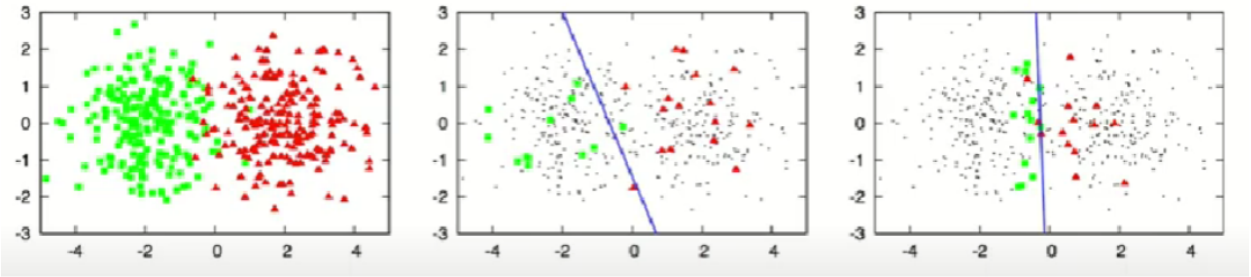


Fig 3.2.1: Fully annotated dataset

Fig 3.2.2: After active learning

Fig 3.2.3: After active learning, with proper selection of samples.

Active learning visualizing.

As you can see in the above image we are selecting the samples that are close to the decision boundary. This avoids making a weak classifier because this reduces the chances of selecting an outlier.

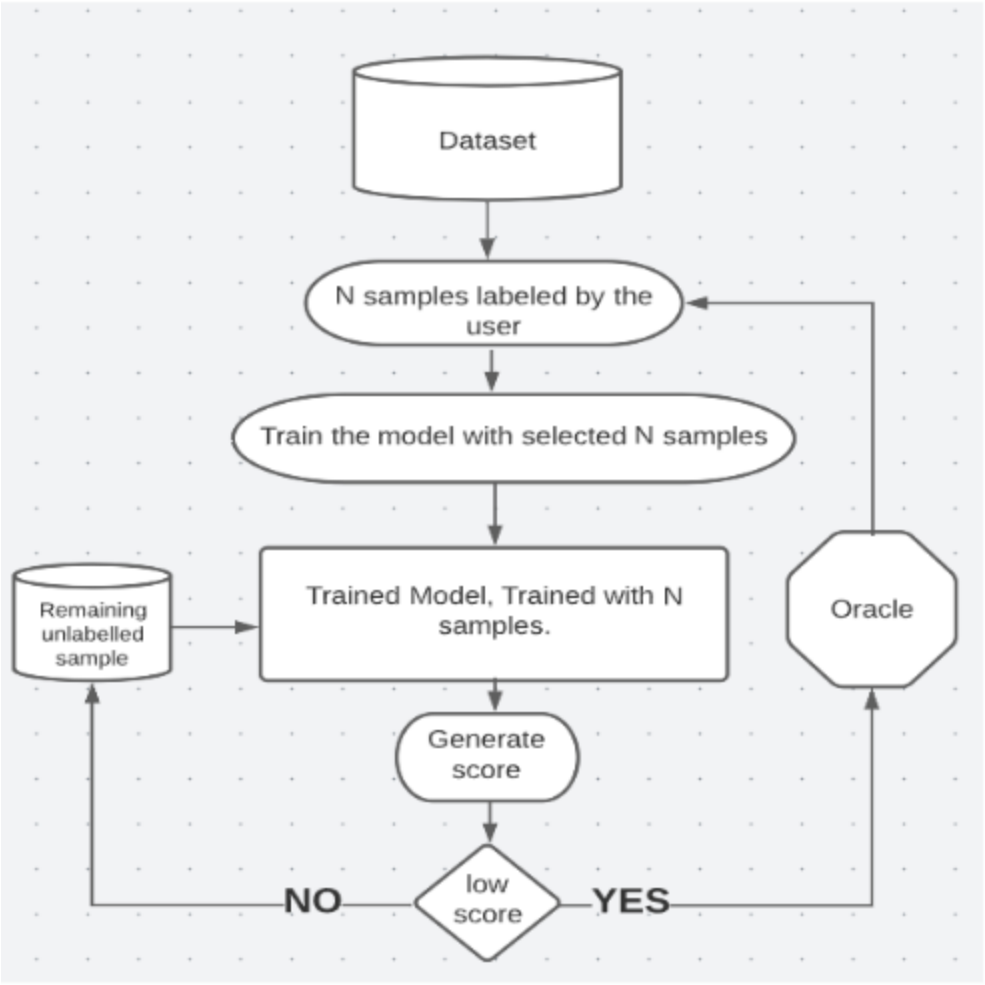


Fig 3: Base Block Diagram of Active Learning