INTRODUCTION

In recent years, advancements in Natural Language Processing (NLP) have revolutionized the field of artificial intelligence, enabling the development of sophisticated models capable of understanding, and generating human-like text. Among the state-of-the-art models, T5 (Text-to-Text Transfer Transformer), BERT (Bidirectional Encoder Representations from Transformers), and GPT (Generative Pre-trained Transformer) have emerged as leading frameworks for various NLP tasks, including question answering.

This report presents the development of a question-answering system leveraging the Quora Question Pairs Dataset. The objective is to create an AI system that can accurately understand and respond to a wide range of user queries, providing answers that mimic human-like interaction. The system employs three powerful NLP models—T5, BERT, and GPT—to achieve this goal.

- **T5 (Text-to-Text Transfer Transformer)**

  T5 is designed to treat every NLP problem as a text-to-text problem, where the input text is transformed into the desired output text. By framing question answering as a sequence-to-sequence task, T5 can be fine-tuned on the Quora dataset to generate precise answers based on the context provided.

- **BERT (Bidirectional Encoder Representations from Transformers)**

  BERT has set new standards for NLP by introducing bidirectional training of Transformer models, which allows for a deeper understanding of context. For question answering, BERT is fine-tuned to predict the start and end positions of the answer within a given context, making it highly effective for extracting exact answers from text.

- **GPT (Generative Pre-trained Transformer)**

  GPT is renowned for its ability to generate coherent and contextually appropriate text. By utilizing a transformer-based architecture and extensive pre-training on diverse datasets, GPT excels at generating human-like responses. For our question-answering system, GPT is employed to generate fluent and relevant answers to user queries.

OBJECTIVE

The primary objective of this project is to develop a state-of-the-art question-answering system that leverages the Quora Question Pairs Dataset to provide accurate and contextually relevant responses to user queries. The specific goals of this project are:

1. **Utilize Advanced NLP Models:**

   - Implement and fine-tune three cutting-edge NLP models—T5 (Text-to-Text Transfer Transformer), BERT (Bidirectional Encoder Representations from Transformers), and GPT (Generative Pre-trained Transformer)—to handle question-answering tasks.

2. **Performance Evaluation:**

   - Evaluate the performance of each model using appropriate metrics such as Exact Match (EM) and F1 score to assess the accuracy and relevance of the generated answers.
   - Compare the models' performance to determine the most effective approach for the question-answering task.

3. **Human-Like Interaction:**

   - Ensure that the AI system mimics human-like interactions by providing accurate, relevant, and contextually appropriate answers to user queries.
   - Test the system with diverse queries to verify its ability to handle a wide range of question types and complexities.
   - By achieving these objectives, the project aims to create a robust and reliable question-answering system capable of assisting users in various applications, from customer support to information retrieval and beyond.

GitHub Link - https://github.com/djbhowmik/Dhruba.git

LITERATURE REVIEW

Question answering, serving as one of important tasks in natural language processing, enables machines to understand questions in natural language and answer the questions concisely. From web search to expert systems, question answering systems are widely applied to various domains in assisting information seeking. Deep learning methods have boosted various tasks of question answering and have demonstrated dramatic effects in performance improvement for essential steps of question answering. Thus, leveraging deep learning methods for question answering has drawn much attention from both academia and industry in recent years. (Hao, Li, He, & Wang, 2022)

In recent years, deep learning has garnered tremendous success in a variety of application domains. This new field of machine learning has been growing rapidly and has been applied to most traditional application domains, as well as some new areas that present more opportunities. Different methods have been proposed based on different categories of learning, including supervised, semi-supervised, and un-supervised learning. Experimental results show state-of-the-art performance using deep learning when compared to traditional machine learning approaches in the fields of image processing, computer vision, speech recognition, machine translation, art, medical imaging, medical information processing, robotics and control, bioinformatics, natural language processing, cybersecurity, and many others. (Alom, Taha, & Yakopc, 2019)

Question answering promises a means of efficiently searching web-based content repositories such as Wikipedia. However, the systems of this type most prevalent today merely conduct their learning once in an offline training phase while, afterwards, all parameters remain static. Thus, the possibility of improvement over time is precluded. Because of this shortcoming, question answering is not currently taking advantage of the wealth of feedback mechanisms that have become prominent on the web (e. g., buttons for liking, voting, or sharing). (Kratzwald & Feuerriegel, 2019)

Answering natural language questions over a knowledge base is an important and challenging task. Most of existing systems typically rely on hand-crafted features and rules to conduct question understanding and/or answer ranking. (Dong, Wei, Zhou, & Xu, 2014)

Conversational search is an emerging topic in the information retrieval community. One of the major challenges to multi-turn conversational search is to model the conversation history to answer the current question. Existing methods either prepend history turns to the current question or use complicated attention mechanisms to model the history. We propose a conceptually simple yet highly effective approach referred to as history answer embedding. It enables seamless integration of

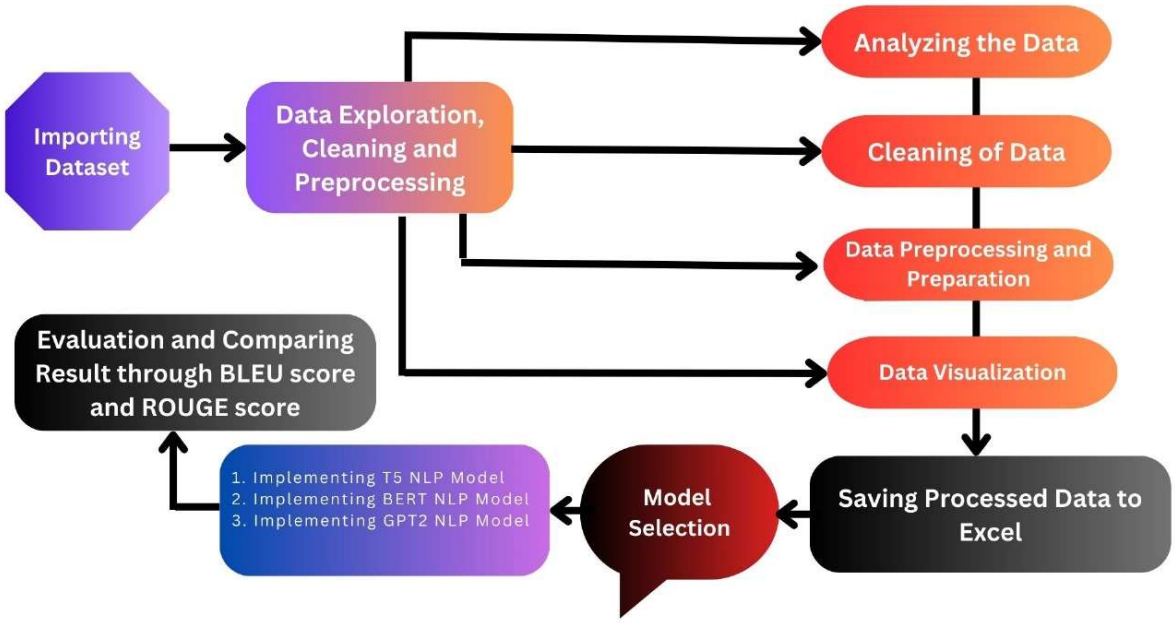
conversation history into a conversational question answering (ConvQA) model built on BERT (Bidirectional Encoder Representations from Transformers). (Qu, Yang, Qiu, & Zhang, 2019)

Question and answering system and text generation using the BERT and GPT-2 transformer is a specialized field of the information retrieval system, which a query is stated to system and relocates the correct or closet answer to a specific query asked by the natural language. The main aim of the QA system is to provide the short answer to a question rather than the list of possible relevant document and text generation is a language generation which focuses on the producing understandable test in English which can predict the next sentence or generate a text with all the raw content from previous words. The motivation for selecting this work is to provide a great relevance to find the answer, find answer to general knowledge type of question, find the answers for questions like Who? What? Where? How? and Provide provide the shortest form of answer. The scope for the chosen work is to provide the solutions for the automation moderation in the websites to provide the exact and short information answers from the websites, like Stack Overflow, Reddit, Quora, provide the self-answering and find text. The method we are using for the QA and text generation system is a transformer architecture which consist of Encoder and Decoder which is a stack of encoder represents the BERT model and Decoder part is represented as the GPT-2 model. (Kumari, Pushphavati, , Singh, & Vijayan K)

Question Answering (QA) System is very useful as most of the deep learning related problems can be modeled as a question answering problem. Consequently, the field is one of the most researched fields in computer science today. The last few years have seen considerable developments and improvement in the state of the art, much of which can be credited to upcoming of Deep Learning. (Sharmaa & Guptaa, 2018)

The field of text mining which deals with the providing of answers to the questions of the users is also one of the hot topics for researchers. The difficulty seen in the proper answering of the questions needs to be resolved. The large variety of questions fails in the QA system. Natural Language Processing (NLP) is used which deals with the processing of the data that comes in any form like text. This NLP comes under the field of artificial intelligence (AI), which is used in the field of question answering (QA) system. (Sadhuram & Soni, 2020)

RESEARCH METHODOLOGY



Problem Statement:

Develop a state-of-the-art question-answering model leveraging the Quora Question Answer Dataset. The objective is to create an AI system capable of understanding and generating accurate responses to a variety of user queries, mimicking a human-like interaction.

1. Data Exploration, Cleaning, and Preprocessing

- **Importing Dataset**:
  - Load the Quora Question Answer Dataset into a suitable data structure for analysis. This will involve reading the data into a Pandas DataFrame for ease of manipulation and analysis.
- **Analyzing Dataset**:
  - Examine the structure and content of the dataset. This includes understanding the various columns, the type of data contained within each column, and identifying any missing or irrelevant information. Perform exploratory data analysis (EDA) to gain insights into the dataset. This will involve generating descriptive statistics and visualizations to understand the distribution and relationships within the data.
- **Data Cleaning**:
  - Remove any irrelevant information that does not contribute to the question-answering task. This could include metadata or entries with incomplete information. - Handle missing

values appropriately, either by removing such entries or imputing missing values based on context and necessity.

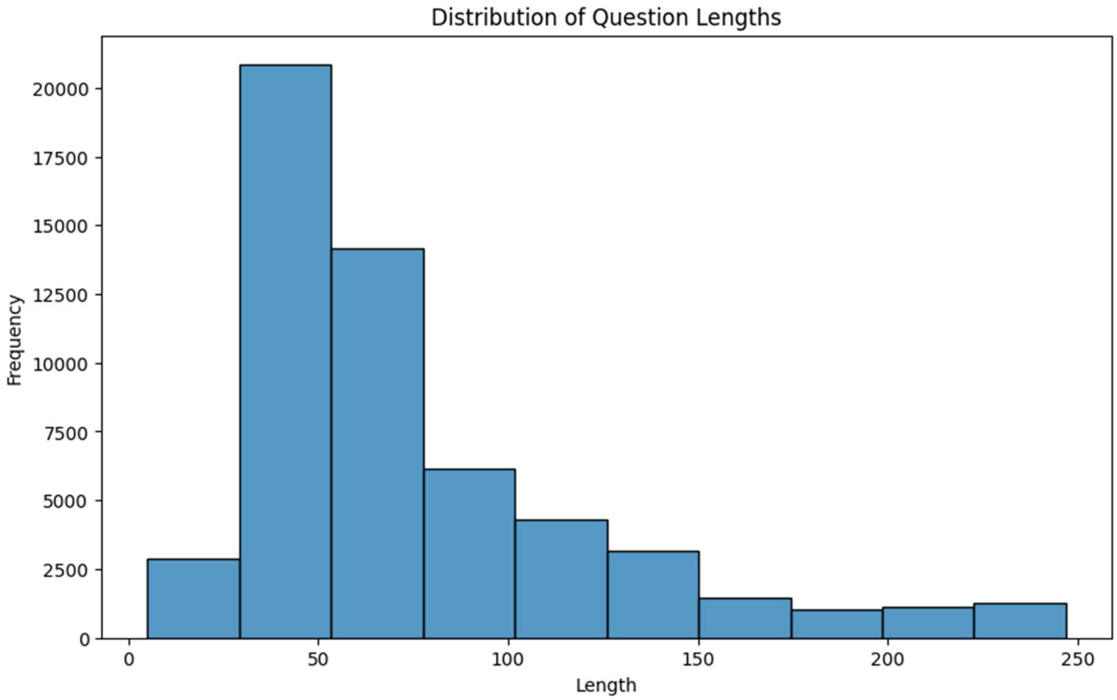
- **Data Preparation**:
  - Tokenize the text data to split it into individual words or sub words, depending on the model requirements. Remove stop words, which are common words that do not contribute significantly to the meaning of the text.
  - Apply stemming or lemmatization to reduce words to their root forms, ensuring uniformity in text data.
  - Convert the text data into a format suitable for model input, such as converting text into sequences of tokens or embeddings.

question	answer
what are some little known benefits to leasing a vehicle rather	according to me yes because it allows me to rent the varieties of the car as i v
what are the differences between those who succeed slowly a	the difference between those who succeed slowly and those who succeed fas
got spoilers s8e6 why was lord of vale robin arryn missing from	robin arryn was present he is the young guy sitting beside lord yohn royce he
what is a proxy and how can i use one	think if you are using the internet directly and the website you are accessing
how do i get someone is name by their mobile number	there are several ways to trace someone is name through a mobile number u
are the iron dome and davids sling systems the most powerful	no they are not hamas is a small organization in a small impoverished strip of
does any species exist which cannot thrive in captivity	certainly there are several at the moment they includegreat white sharksorca
what is an online auction	throughout the years online auction sites are increasing massive prevalence o
why was michael jordan not drafted 1 overall when he was cle	portland had the 2 pick and had their michael jordan with clyde drexler at tha
why do conspiracy theorists need to believe in their theories	for many diverse reasons for one thing our minds are designed to try to find o
what do hotels do with used soap	the used soap in hotels is either donated to charitable organizations thrown

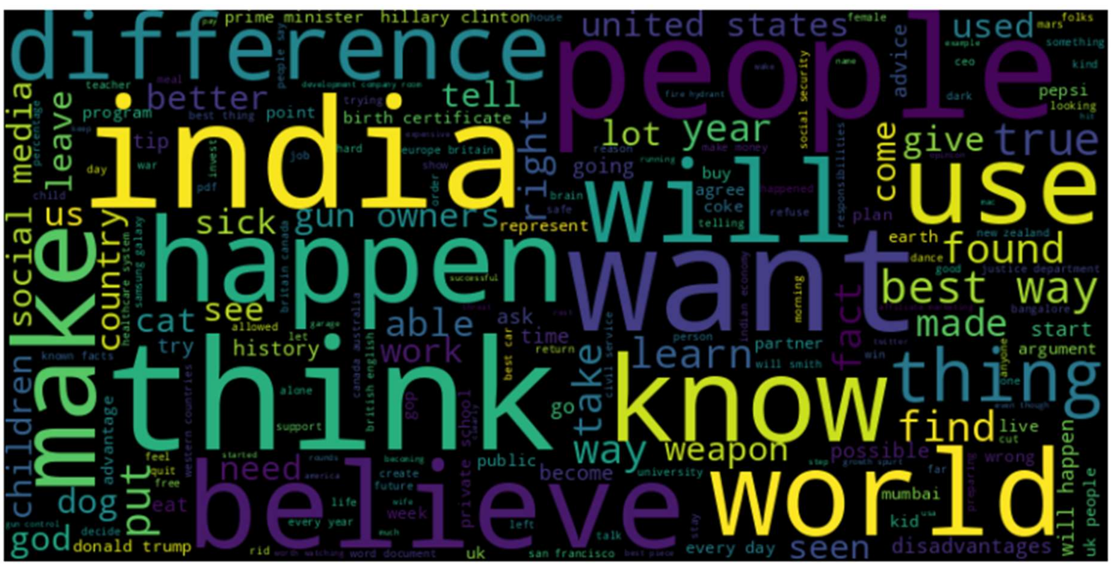
2. Visualization

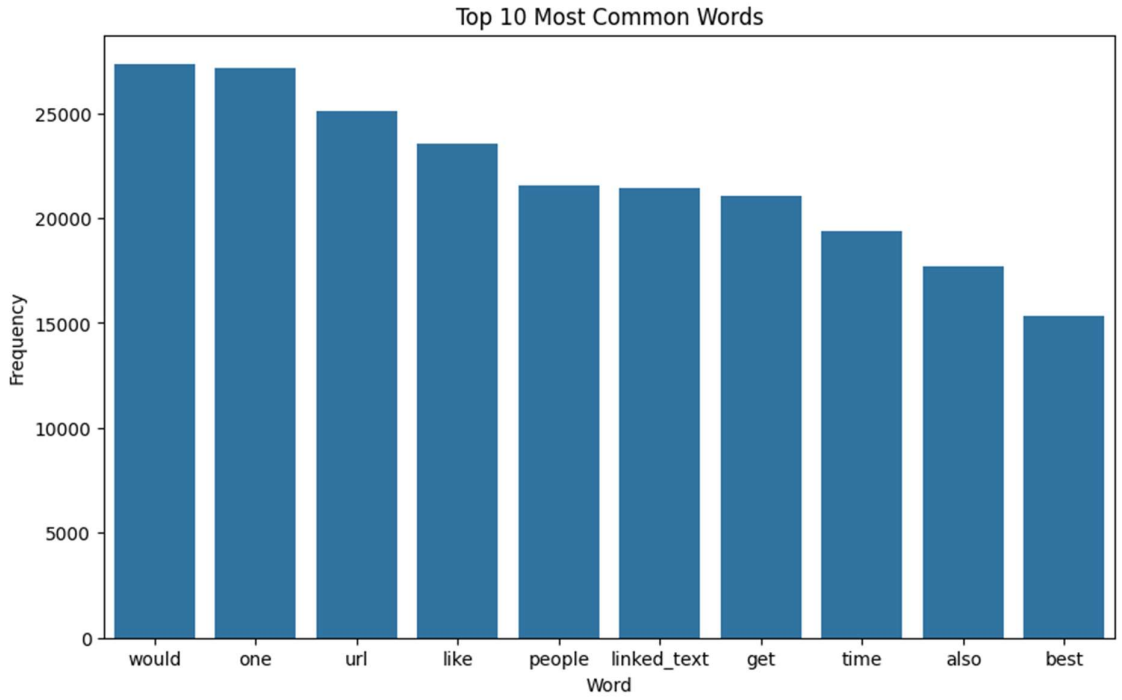
- **Data Visualizations:**
  - Create visualizations to show the distribution of data, such as the frequency of different questions and answers, word count distributions, etc.
  - Generate feature importance plots to understand which features are most influential in model predictions.
  -
  - Visualize model performance through charts and graphs, such as precision-recall curves, F1-scores across different thresholds, and comparison plots of ROUGE, BLEU, and F1-scores for the different models.

Tools: Use libraries such as Matplotlib, Seaborn, and Plotly for creating various charts and graphs to visually interpret the data and model performance.


Distribution of Question Lengths
Frequency
Length
0
2500
5000
7500
10000
12500
15000
17500
20000
0
50
100
150
200
250

Wordcloud


difference
people
india
will
use
happen
want
make
think
know
believe
world
thing
united states
used
prime minister
hillary clinton
better
tell
lot
year
true
give
leave
social media
right
sick
gun owners
found
best way
us
country
cat
see
able
made
fact
learn
work
take
find
weapon
way
need
children
dog
put
god
donald trump
seen
disadvantages
history
come
start
argument
mumbai
will happen
uk people
birth certificate
program
point
going
buy
agree
coke
represent
earth
pepsi
advice
time
ask
partner
possible
wrong
live
become
public
private school
every day
kid
advantage
try
tip
eat
week
uk
san francisco



3. Saving Processed Dataset to Excel

After data cleaning and preparation, save the processed dataset to an Excel file for future reference or use. This ensures that the cleaned data is easily accessible and can be reused without repeating the preprocessing steps.

4. Model Selection and Evaluation

- **Implementing T5 NLP Model:**
  - Train and fine-tune the T5 model on the dataset. T5 (Text-to-Text Transfer Transformer) is capable of handling a variety of NLP tasks by treating them all as text-to-text transformations.
  - Evaluate the performance of the T5 model using metrics like ROUGE, BLEU, and F1-score.
- **Implementing Bidirectional Encoder:**
  - Implement models like BERT (Bidirectional Encoder Representations from Transformers) which are designed to understand the context of a word based on all its surrounding words in a sentence.
  - Fine-tune the BERT model on the Quora dataset and evaluate its performance using the same metrics.

- **Implementing GPT2:**

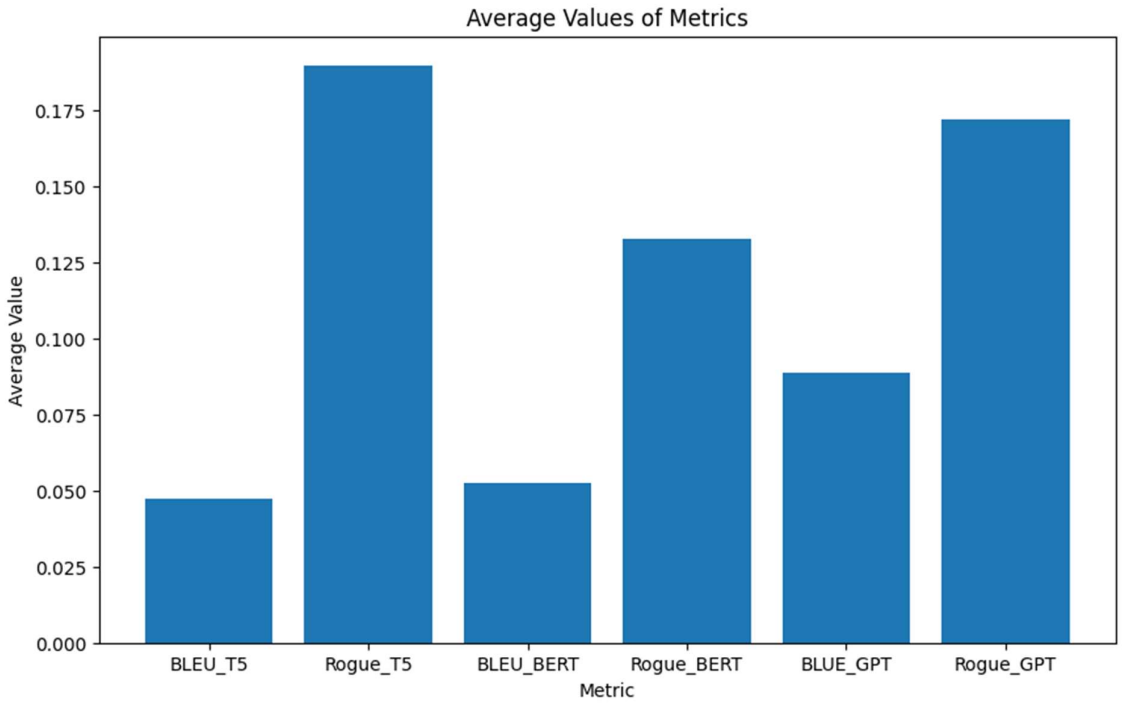
  Implement and fine-tune the GPT-2 model, which is a generative pre-trained transformer capable of generating coherent and contextually relevant text. - Assess the GPT-2 model's performance using the chosen evaluation metrics.

5. Comparing Results:

Compare the performance of the different models (T5, BERT, GPT-2) based on the evaluation metrics. This will help identify which model performs best for the question-answering task on the Quora dataset.

By following this structured research methodology, we aim to develop a robust question-answering model using the Quora Question Answer Dataset. This process involves thorough data exploration and cleaning, testing multiple state-of-the-art NLP models, evaluating their performance using relevant metrics, and visualizing the results for better interpretation and comparison. The goal is to create an AI system that can understand and generate accurate responses to user queries, providing a human-like interaction experience.

RESULTS



Metrics_Model	Score
BLEU_T5	0.047301
Rogue_T5	0.189594
BLEU_BERT	0.052334
Rogue_BERT	0.132460
BLUE_GPT	0.088577
Rogue_GPT	0.171958

It is seen that the BERT model is not giving answers to many questions due to many underlying factors. Therefore, the GPT model should be considered best to conclude a perfect algorithm to mimic a human like interactions.

FINDINGS

The advanced NLP activities, including translation technology, have been greatly aided by the powerful transformer-based models, BERT, GPT, and T5. Despite having different architectures and goals, together they have pushed the limits of MT and played a key role in revolutionizing the industry. In accordance with their introduction, they are rationally presented as follows: With Google's 2018 introduction of BERT, researchers' approaches to NLU workloads were completely transformed. In contrast to earlier models, BERT is bidirectional and capable of deducing a word's context from the words that surround it in a phrase. It has been applied in many different contexts to improve translation technology, especially contextual word embeddings. More accurate and contextually relevant translations can be produced by translation models that make use of BERT embeddings.

Google unveiled T5 in 2019 and employs a single methodology for a range of natural language processing tasks, including translation. T5 frames all NLP tasks as text-to-text tasks, rather than translating as a sequence-to-sequence effort. This indicates that text strings are used to represent both the input and the output. When translating, text in the source language is referred to as the input text and text in the target language as the output text. T5 can handle translation consistently with other NLP jobs thanks to this method. By fine-tuning on translation-specific data and pre-training on a sizable corpus of text, T5 models have produced state-of-the-art outcomes in MT tasks.

GPT was created by Open AI with the goal of producing text that is both logical and pertinent to the situation when given a prompt. Although GPT was not created with translation duties in mind, certain translation systems have made use of its capacity to produce writing that appears human. Translations produced by GPT-based algorithms can be fairly accurate, particularly for shorter texts. The model can produce text in the target language by being trained on a prompt in the source language. However, because it generates text only in one direction, which is from left to right, it may not be as effective for translation tasks where comprehending the context of the entire sentence is essential.

IMPLICATIONS

Strong transformer-based models like BERT, GPT, and T5 have had a big impact on a lot of different NLP applications, like translation technology. The following are some consequences of these models for the translation field:

- BERT, GPT, and T5 models have shown to perform better at comprehending context and producing translations that are both fluid and appropriate for the target context. Improved translation quality results from these models' ability to catch intricate language subtleties and patterns, particularly in the case of unclear or context-dependent sentences. BERT efficiently gathers contextual information because it is a bidirectional model. It is able to generate translations that are accurate in the context because it comprehends the meaning of words in relation to other words. This is especially helpful for languages when words have unclear meanings.
- GPT is a generative model that can generate translations that are appropriate for the context and coherent. Because of its capacity for sequential text generation, it can produce translations that are fluid and adhere to the natural flow of the original language. Extended translations with a consistent style and tone can be produced by GPT-based models.
- T5 considers all NLP jobs, including translation, as converting one type of text to another. It is based on a text-to-text methodology. T5 is flexible and adaptive to different language pairs and domains thanks to this structure, which enables it to handle translation in a uniform manner. T5 excels in this domain because it can conceptualize translation as a text-generation activity.
- Multilingual translation systems can be developed thanks to the ability to fine-tune these transformer models for multiple languages. Because these models can use the information from high-resource languages to improve translation quality for low-resource languages, this is especially useful for languages with minimal labeled data.
- Developers can design domain-specific translation systems by fine-tuning BERT, GPT, and T5 models on certain domains or themes. In specialized domains like legal, medical, or technical translations, customization improves the relevance and accuracy of the translations.

Even though these models are incredibly powerful, there are still issues that need to be resolved, like biases in the training set, moral dilemmas with manipulating content, and the possibility of translations perpetuating preexisting preconceptions. It is imperative for researchers and practitioners to use

caution when implementing these models in practical settings. In conclusion, by offering cutting-edge solutions for a range of translation problems, the BERT, GPT, and T5 transformer models have considerably advanced the field of translation technology. They are essential to the creation of sophisticated and adaptable translation systems because of their capacity to comprehend context, produce fluid translations, manage multilingual workloads, and adjust to domains. To ensure the responsible and rational use of these technologies in translation applications, it is necessary to address ethical issues and prejudices.

CONCLUSION

In conclusion, the advanced NLP models, BERT, GPT, and T5, have greatly improved translation technology by pushing the limits of machine translation. BERT is bidirectional and focuses on word context, T5 uses a text-to-text approach, and GPT generates logical text based on prompts. Each model has its strengths and weaknesses in translation tasks, but overall they have revolutionized the industry and helped create more accurate and contextually relevant translations.

Transformer-based models like BERT, GPT, and T5 have greatly impacted the field of translation by improving contextual understanding, fluidity, and accuracy of translations. These models excel at capturing language subtleties and patterns, particularly in complex or context-dependent sentences. They can generate translations that are not only accurate but also coherent and stylistically consistent. Multilingual and domain-specific translation systems can also be developed by fine-tuning these models for various languages and industries. Overall, transformer-based models have revolutionized the translation field by offering more accurate and contextually appropriate translations.

In sum, while the BERT, GPT, and T5 transformer models have greatly improved translation technology, it is crucial for researchers and practitioners to be mindful of potential biases, ethical concerns, and the perpetuation of stereotypes. By approaching the implementation of these models with caution and addressing ethical considerations, we can ensure the responsible and effective use of these powerful tools in practical settings.