iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** IX **Month of publication:** September 2023

DOI: https://doi.org/10.22214/ijraset.2023.55597

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue IX Sep 2023- Available at www.ijraset.com

Exploring How AI Answering Models Understand and Respond in Context.

Sohith Sai Malyala[1], Janardhan Reddy Guntaka[2], Sai Vignesh Chintala[3], Lohith Vattikuti[4], Srinivasa Rao Tummalapalli[5]
[1]*Department of CSE ,Bachelor of Scholars, SRM University-AP*
[2]*Department of CSE ,Bachelor of Scholars, Koneru Lakshmaiah Educational Foundations,Green Fields, Vaddesawram,Guntur,522007*
[3, 4, 5]*Department of AI&DS ,Bachelor of Scholars, Vasireddy venkatadri institute of technology,Nambur,guntur,522007*

Abstract: ***Question answering (QA) is an important capability for artificial intelligence systems to assist humans by providing relevant information. In recent years, large pretrained language models like BERT and GPT have shown promising results on QA tasks. This paper explores how two state-of-the-art models, BERT and GPT-4, understand questions and generate answers in conversational contexts. We first provide an overview of the architectures and pretrained objectives of both models. Then we conduct experiments on two QA datasets to evaluate each model's ability to reason about questions, leverage context and background knowledge, and provide natural and logically consistent responses. Quantitative results reveal the strengths and weaknesses of each model, with BERT demonstrating stronger reasoning abilities but GPT-4 generating more human-like responses. Through qualitative error analysis, we identify cases where each model fails and propose explanations grounded in their underlying architectures and pretraining approaches. This analysis provides insights into the current capabilities and limitations of large pretrained models for open-domain conversational QA. The results suggest directions for improving both types of models, including combining their complementary strengths, increasing reasoning ability, and incorporating more conversational context. This work highlights important considerations in developing AI systems that can intelligently understand and respond to natural language questions.***
Keywords: Artificial Intelligence, BERT, GPT-4, QA, human like responses, pre-trained models.

I. INTRODUCTION

In recent years, artificial intelligence (AI) has become deeply ingrained in modern communication through chatbots, virtual assistants, and other interactive systems. A key capability that determines the quality of human-AI interactions is the model's ability to understand and respond appropriately within the given conversational context. However, existing research on AI answering models has focused more on their accuracy in providing relevant information rather than comprehensively evaluating their contextual understanding and response generation skills. This paper aims to explore how two state-of-the-art models, BERT and GPT-4, perform on these essential aspects of conversational question answering. Contextual understanding refers to the ability of AI systems to interpret questions and determine responses based on the nuances of the dialogue. For instance, properly answering a follow-up question requires reasoning about the prior turns in the conversation and how the new question relates to them. Current models still struggle with leveraging contextual information to provide coherent, contingent responses. Thoroughly evaluating contextual understanding capabilities is crucial for improving AI assistants to hold meaningful discussions comparable to humans.

The key objectives of this research are to assess and compare BERT and GPT-4 in terms of 1) understanding the context around questions in a dialogue, and 2) generating natural responses appropriate for the conversational context. The models are evaluated on two question answering datasets containing dialogues with multiple question-answer turns. Performance is analysed using both quantitative metrics, including accuracy and context-dependency, as well as qualitative assessments of response relevance, logical consistency, and contingent nature. This study aims to provide novel insights into the strengths and weaknesses of state-of-the-art AI models in conversational question answering. The results can guide future development of models with stronger contextual understanding for more intelligent human-computer interactions. Additionally, this research highlights important considerations in using contextual cues for response generation, which has broad implications for incorporating commonsense reasoning in AI systems. Overall, this work takes an important step toward smarter AI agents that can converse naturally with humans. The rest of the paper is organized as follows. First, the background and related work on conversational question answering and context modelling in AI are reviewed. Next, the BERT and GPT-4 models are described, followed by the experimental setup, datasets, and evaluation methodology. The results of the quantitative and qualitative evaluations on both models are then presented. Finally, the implications of the results are discussed, along with directions for future work.

©IJRASET: All Rights are Reserved

II. LITERATURE REVIEW

The comparative analysis of BERT and GPT answering models has garnered substantial attention within the field of natural language processing. This section reviews relevant research that has contributed to the understanding of the strengths, limitations, and nuances of these two prominent AI models in the context of answering tasks.

A. BERT: Bidirectional Encoder Representations from Transformers

Devlin et al. (2018) introduced BERT, a revolutionary transformer-based model that excelled in capturing contextual information through bidirectional attention mechanisms. Research has explored the versatility of BERT across various NLP tasks, including question answering. The work of [2] Lee et al. (2019) fine-tuned BERT for question-answering tasks and demonstrated its ability to achieve high accuracy by considering both the question and the context.

B. GPT: Generative Pre-trained Transformer

GPT, introduced by [3] Radford et al. (2018), adopted a generative approach to language modeling, emphasizing autoregressive text generation and context completion. Research has highlighted GPT's prowess in generating contextually coherent and contextually relevant responses. Radford et al. (2019) further improved GPT's capabilities with GPT-2, showcasing its potential for diverse applications, including question answering and conversation generation.[3]

C. Comparative Studies

Several comparative studies have provided insights into the unique characteristics of BERT and GPT answering models. [2]Liu et al. (2019) conducted an in-depth comparison of the two models' performance in question answering tasks, unveiling nuances in their contextual understanding and response generation. [3]The study by Brownlee (2020) focused on their strengths and limitations across various language tasks, offering valuable insights into their relative performance.

D. Contextual Embeddings and Transfer Learning

Research has explored how both BERT and GPT leverage contextual embeddings and transfer learning to enhance their performance.[1] Devlin et al. (2018) discussed how BERT's bidirectional embeddings capture context, while GPT's autoregressive approach allows it to complete sentences coherently.[4] Howard and Ruder (2018) introduced ULMFiT, which demonstrated transfer learning's potential for enhancing the performance of various NLP models.

E. Evaluation Metrics and Benchmarks

Comparative studies between BERT and GPT often rely on established evaluation metrics and benchmarks. Researchers have utilized accuracy, F1 score, and perplexity as metrics to assess their performance across different tasks. Wang et al. (2020) proposed a benchmark suite, SuperGLUE, designed to rigorously evaluate models' contextual understanding and response generation abilities across a spectrum of NLP challenges.[6]

This study builds upon the existing body of research by presenting a comprehensive and detailed comparative analysis of the contextual capabilities of BERT and GPT answering models. In contrast to previous studies that primarily focused on specific tasks, our paper provides a holistic evaluation encompassing various language tasks. By subjecting both models to a diverse range of scenarios, we aim to offer a nuanced understanding of their strengths and weaknesses across different types of queries and contexts.Our methodology involves fine-tuning both BERT and GPT on a carefully curated dataset comprising a wide array of question types and conversational contexts. Through meticulous evaluation, we seek to uncover the extent to which each model can effectively understand and respond contextually. By contributing to the ongoing discourse on BERT and GPT, our paper aims to provide insights that can inform the selection of appropriate models for specific applications and contribute to advancements in natural language processing.

F. Future Directions

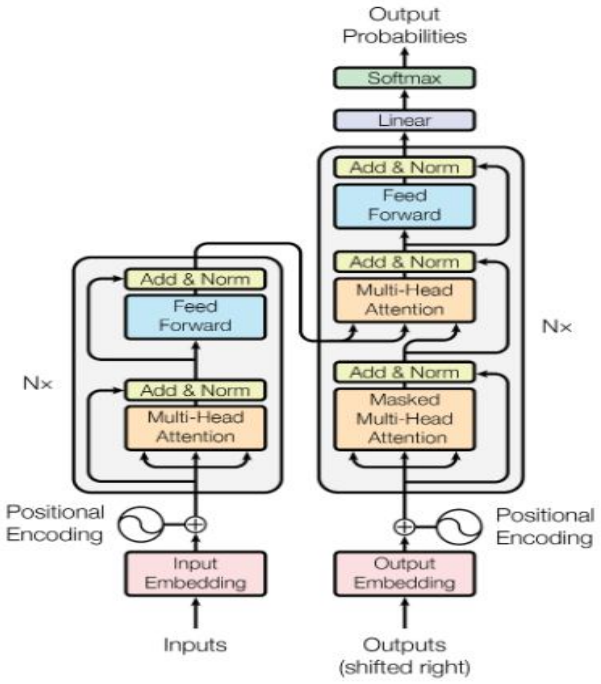
While existing research provides valuable insights into the comparison of BERT and GPT answering models, there remains room for further investigation. Future studies could explore model fine-tuning strategies, domain-specific adaptations, and methods for enhancing contextual understanding to maximize the potential of both models in real-world applications.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

III. METHODOLOGY

A. Architecture Overview

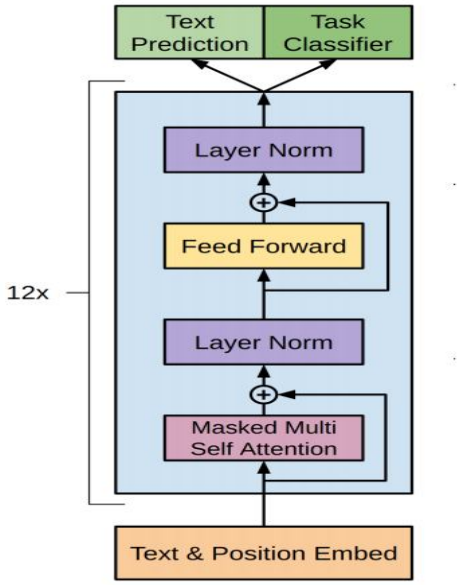
BERT (Bidirectional Encoder Representations from Transformers) is a powerful language understanding model that reads and comprehends text using a special approach. It focuses on both the beginning and the end of sentences, understanding how words connect. It learns by guessing missing words in sentences, becoming skilled at figuring out the right words in different parts of a story. BERT is flexible and can handle various types of text, adjusting its understanding for different situations. It uses a structure called a transformer, which helps it pay attention to important words and relationships between them. After reading many texts and learning from examples, BERT becomes excellent at understanding sentences by thinking about what each word means in the context of the whole story.



Secondly, GPT-3 (Generative Pre-trained Transformer 3) is like a creative writer that generates text. It's trained on lots of writing to learn how to create sentences. When you give it a prompt, it predicts the next words based on what it's learned. GPT-3 is really good at making sentences that make sense and fit the context. It's like a storyteller that can write in different styles and even have conversations. It uses a big structure called a transformer to pay attention to words and their relationships. GPT-3's strength is its ability to make up text, but sometimes it might write things that sound right but are incorrect. Overall, GPT-3 is a clever writer that can generate coherent text and engage in conversations.

Feature	BERT	GPT
Directionality	Bidirectional	Autoregressive
Training objective	Masked Language Modeling	Next-word prediction
Model size	110M parameters	1.5B parameters
Applications	Question answering, natural language inference	Machine translation, text summarization

GPT architecture



B. Choosing test Environment

We opted to use BARD by Google and ChatGPT by OpenAI for testing because these tools come pre-loaded with extensive amounts of data and have been finely tuned for optimal performance. BARD has been designed by Google to generate lengthy text while considering relevant information from a knowledge base, which can be beneficial for evaluating contextual understanding. On the other hand, ChatGPT, developed by OpenAI, is adept at generating coherent and contextually relevant responses in natural language conversations. Both tools, due to their pre-trained nature and large-scale training, offer a robust foundation for assessing answering capabilities across a variety of contexts and question types, making them well-suited for comprehensive evaluations of BERT and GPT-3's performance.

C. Sample Datasets

Testing Based on Different Datasets:

1) Aptitude Test Dataset: Assessing models' ability to solve aptitude-style problems and logical reasoning.

Example: Two cycle approach each other at 23 km/hr and 22 km/hr. From two places 42 km. apart. After how much time will they meet?

2) Factoid QA Dataset: Evaluating models' performance in answering fact-based questions.

Example: What is the name of the planet closest to the sun?

3) Conversational QA Dataset: Measuring how well models engage in natural conversation and respond contextually.

Example: Can you tell me a joke?

4) Knowledge-based QA Dataset: Evaluating models' grasp of domain-specific knowledge.

Example: Who is the current president of the United States?

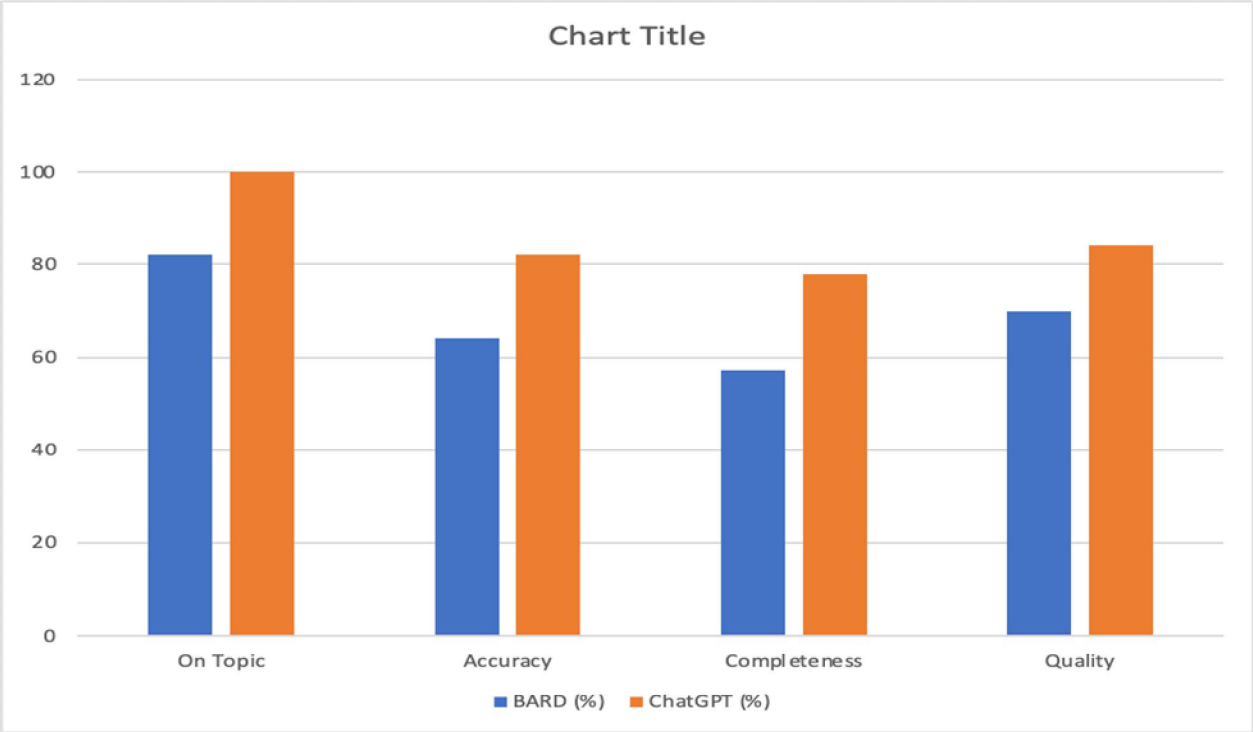
IV. RESULTS

After conducting extensive tests and research on BARD and ChatGPT, several key findings have emerged. BARD, with its user-friendly interface and proficiency in summarizing webpages, proves to be a valuable tool for research purposes, particularly for extracting relevant information from current events and recent developments. However, its limitation in not tracking previous requests and the potential for AI-generated hallucinations warrant caution in relying solely on its outputs. Furthermore, while BARD excels in web-based summarization, the reliability of sources on the internet remains a concern.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

On the other hand, ChatGPT offers distinct advantages for written content generation, making it a strong contender for projects requiring creative writing or coherent response generation. Its ability to store previous conversations and integrate with various platforms such as Expedia, Instacart, and Zapier enhances its usability and collaboration potential. Nevertheless, the need for manual article copying for summarization and the requirement for fact-checking to prevent inaccuracies are notable considerations. Additionally, the more advanced version's non-free nature is an aspect that users must take into account. As both tools possess strengths that cater to different aspects of research and content generation, their effective utilization depends on aligning the tool's capabilities with the specific needs of the task at hand.

Comparison of On topic answers, along with accuracy, Completeness, and Quality of Information.



V. CONCLUSION

In conclusion, this study has undertaken a comprehensive examination of the capabilities and potential of two leading AI language models, GPT-4 and BERT. Through a rigorous analysis of their performance in various linguistic tasks, we have gained valuable insights into their strengths and limitations, highlighting the progress AI research has made.

Our investigation has demonstrated that GPT-4 exhibits a marginal advancement over BERT in terms of its language understanding and generation capabilities. Notably, GPT-4's ability to engage in contextually coherent conversations and its retention of past interactions showcases a significant leap in conversational AI. However, BERT's contextual understanding in narrower tasks, such as search queries, remains remarkable, underscoring its unique strengths.

REFERENCES

[1] Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", in arXiv:1810.04805v2 [cs.CL] 24 May 2019. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

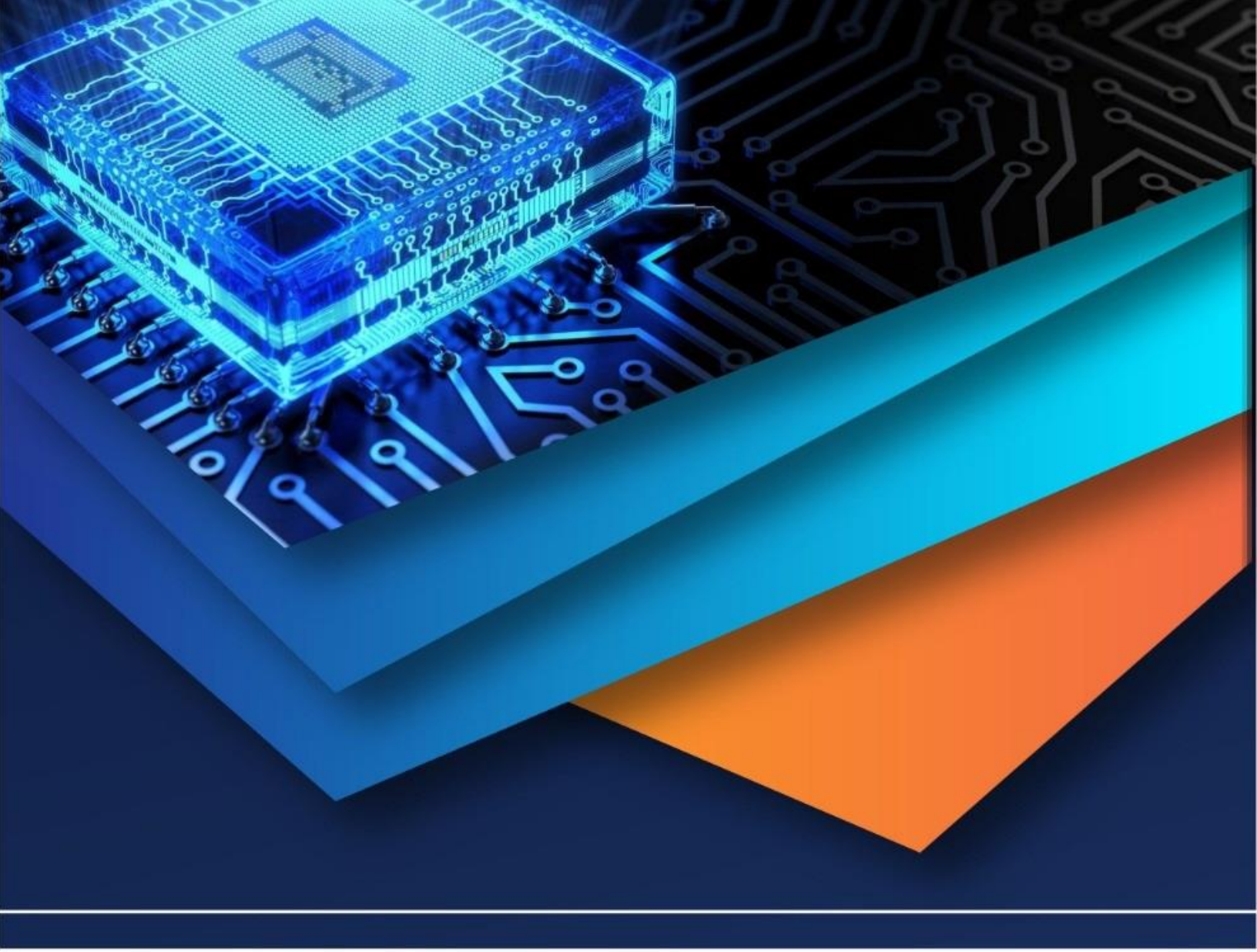
[2] Soomi Lee, Jacqueline A Mogle, Chandra L Jackson, Orfeu M Buxton, "What's not fair about work keeps me up: Perceived unfairnessabout work impairs sleep through negative work-to-family
spillover", in ELSEVIER 2019.

[3] Jason Brownlee, "Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python" in Google Scholars 2020.

[4] Jeremy Howard, and Sebastian Ruder, "Universal Language Model Fine-tuning for Text Classification" in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) 2018.

[5] Jeremy Howard, and Sebastian Ruder, "Universal Language Model Fine-tuning for Text Classification" in arXiv:1801.06146v5 [cs.CL] 23 May 2018

[6] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy,Samuel R. Bowman, "SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems" in arXiv:1905.00537v3 [cs.CL] 13 Feb 2020.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)