**INDIGO HACK TO HIRE**

**Case Study Report**



**Submitted By –**

Name – Shouvik Pramanik

Role - Data Science (4010)

**TABLE OF CONTENTS**

1. Introduction
2. Literature Survey
3. Methodology
4. Results
5. Conclusion

**INTRODUCTION**

In this study, we investigate the capabilities of various transformer-based models for question answering tasks using the Quora Question Answer Dataset (Quora-QuAD). Question answering (QA) systems are now used in a wide range of applications, including customer service and instructional resources. However, different models have distinct strengths and weaknesses, making it critical to understand their performance and how they can be effectively combined.

The primary goal of this research is to preprocess the Quora-QuAD dataset, evaluate the performance of three advanced models (BERT, GPT-2, and T5), and offer techniques to improve QA systems using hybrid approaches.

* BERT (Bidirectional Encoder Representations from Transformers) is used because of its powerful context awareness and extraction capabilities.
* GPT-2 (Generative Pretrained Transformer 2) is a tool for producing inventive and fluent replies.
* T5 (Text-To-Text Transfer Transformer) is used for its versatility in converting input text into a desired output format, hence improving the refinement process.

By integrating these models, we want to build a more robust QA system capable of handling a wide range of questions more successfully. In addition, we suggest enhancements like as continuous learning, explainability, and hybrid modeling to increase the system's performance and dependability.  
  
This project demonstrates a complete approach to developing and testing a complex QA system, as well as insights regarding transformer-based models' strengths and prospective advancements.

**LITERATURE SURVEY**

Research Paper – 1

Title - Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

Authors - Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu

Published By - Google, Mountain View, CA 94043, USA

Summary –

Colin Raffel et al.'s paper "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer" investigates the effectiveness of transfer learning in natural language processing (NLP) using a novel unified framework that views all text-based tasks as text-to-text problems. This method allows a single model to do multiple tasks reliably, such as summarization, question answering, and text classification. Their study is around the "Colossal Clean Crawled Corpus" (C4), a new large-scale, cleaned dataset obtained from Common Crawl data. The authors conducted comprehensive experiments to test various pre-training objectives, model architectures, datasets, and transfer methodologies, indicating that their approach can obtain cutting-edge results on a variety of NLP benchmarks.

Key findings show that combining the text-to-text framework with the high-quality C4 dataset improves model performance significantly. The study reveals that larger models and thorough fine-tuning procedures result in significant improvements across a wide range of tasks. It underlines the significance of balancing pre-training with fine-tuning to avoid overfitting, particularly in low-resource environments. The authors hope that by publishing their dataset, pre-trained models, and code, they would be able to support future study and development, showing the potential of unified techniques and large-scale pre-training to push the bounds of NLP capabilities.

Research Paper – 2

Title - BERT: A Review of Applications in Natural Language Processing and Understanding

Authors - Koroteev M.V

Published By - Financial University under the government of the Russian Federation, Moscow, Russia

Summary –

M.V. Koroteev's "BERT: A Review of Applications in Natural Language Processing and Understanding" investigates BERT's transformative impact on natural language processing (NLP). BERT, introduced by Google AI in 2018, represents a significant improvement above previous models such as word2vec and GloVe, owing mostly to its bidirectional training strategy. This method allows BERT to grasp context from both directions in a sentence, resulting in unparalleled precision in a variety of NLP tasks like as text classification, question answering, and named entity recognition. Using transformers, BERT learns deep text representations using a masked language model technique, predicting randomly masked words while taking into account both left and right context.

The review underscores BERT's superiority over other models like GPT and ELMo, attributing its edge to the bidirectional context understanding. Nonetheless, the text admits obstacles such as training's computational complexity and "catastrophic forgetting" during fine-tuning. Despite these difficulties, BERT has quickly become the industry standard for many NLP applications, because to its capacity to be fine-tuned for specific requirements. This adaptability has led to extensive use in a variety of disciplines, including conversational AI, machine translation, and information retrieval. The work finishes by recommending future research areas, such as improving BERT efficiency and investigating multitask learning, to improve its performance and application.

Research Paper – 3

Title – Improving Language Understanding by Generative Pre-Training

Authors - Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever

Published By - OpenAI

Summary –

This work describes a novel strategy to improve natural language understanding that combines unsupervised pre-training with supervised fine-tuning. The authors pre-train a large language model on a broad corpus of unlabeled text using a Transformer architecture before fine-tuning it for specific supervised tasks. This strategy seeks to use the huge volumes of available unlabeled text data to improve performance on a range of language understanding tasks, particularly in domains where labeled data is rare.

The suggested approach delivers cutting-edge performance on nine of the twelve tasks tested, including natural language inference, question answering, semantic similarity, and text categorization. The use of the Transformer architecture rather than LSTMs, task-aware input transformations during fine-tuning, and an auxiliary language modeling target during fine-tuning are all significant innovations. The findings demonstrate significant advances over earlier methodologies, particularly in commonsense reasoning, question responding, and textual entailment tasks.

The analysis of the model suggests that transferring more layers from the pre-trained model enhances performance, and the model has some zero-shot learning capabilities. Importantly, the technique works across datasets of various sizes, ranging from tiny (5,700 examples) to large (550,000 examples). This study shows that unsupervised pre-training of a language model can significantly improve performance on a variety of language understanding tasks with minimal task-specific architecture changes, potentially opening up new avenues for improving NLP systems with limited labeled data.

Research Paper – 4

Title - Question Answering Systems: A Systematic Literature Review

Authors - Sarah Saad Alanazi, Nazar Elfadil, Mutsam Jarajreh, Saad Algarni

Published By - Computer Engineering Department Fahad Bin Sultan University Tabuk, Saudi Arabia

Summary –

The paper conducts a systematic literature analysis of Question Answering Systems (QAS), examining works published after January 2018 to assess the current status of QAS research, identify existing gaps and limits, and investigate successful design strategies. A study of 80 relevant studies revealed that QAS research employs a wide range of methodologies, with many researchers developing novel systems from the ground up. There is a positive tendency toward integrating multiple components, however the area faces difficulties in objectively comparing different systems.  
  
The main gaps and limits noted are the specialization of many QAS rather than establishing general-purpose systems, flaws in the underlying models and algorithms, and the need for vast amounts of high-quality training data, particularly for deep learning. Standard datasets and question formats limit real-world applicability, and several studies lacked complete reviews. The review identifies effective design methodologies that emphasize syntax and context, use word encoding and knowledge systems, and combine deep learning techniques with machine learning and AI components. The report suggests using more modular ways to improve collaboration, performing more extensive evaluations in real-world contexts, and addressing the constraints of specialized systems and reliance on standard datasets. Despite these obstacles, QAS research is making great progress and remains a vibrant and diversified area.

**METHODOLOGY**

1. Data Pre-processing

The dataset is first cleaned up by converting it to lowercase and deleting non-word characters, unnecessary spaces, and newline characters. Standardization ensures consistency. Tokenization is then accomplished by splitting the text into words with NLTK, followed by stop word removal to remove common, non-essential terms. Stemming and lemmatization are used to reduce words to their base forms and assure consistency.

2. Setting Up the Model

Three transformer-based models are used: BERT (bert-large-uncased-whole-word-masking-fine-tuned-squad) for extraction, GPT-2 (gpt2) for text creation, and T5 (google-t5/t5-small) for versatile text transformation. These pre-trained models are loaded with Hugging Face Transformers, allowing us to take advantage of their capabilities in question-answering tasks.

3. Model Assessment

Each model is tested using a sample of ten dataset items. BERT collects answers from context, GPT-2 generates replies from questions, and T5 refines responses. ROUGE, BLEU, and F1 scores are used to evaluate the overlap with reference answers, text quality, and overall accuracy.

4. Examining the Results

The ROUGE, BLEU, and F1 ratings are used to evaluate the performance of BERT, GPT-2, and T5. This comparison aids in determining each model's strengths and shortcomings, as well as how well their solutions correspond to reference answers.

5. Proposed Improvements  
To improve performance, we recommend Hybrid Models that use BERT for extraction, T5 for refinement, and GPT-2 for innovative replies. Continuous Learning helps keep models up to date by incorporating user feedback and retraining. Explainability tools can visually represent model attention, making judgments more apparent. Data augmentation and domain-specific pre-training are also suggested for improving model accuracy and flexibility.  
  
6. Reporting & Visualization  
Data distribution, feature importance, and model performance are visualized using Matplotlib, Seaborn, and Plotly. These aid in comprehending the results and gaining actionable insights for further improving the question-and-answering system.

**CONCLUSION**

**BERT Model**

The BERT model, fine-tuned for question answering, yielded the following performance metrics:

* **ROUGE Scores**:
  + ROUGE-1: 0.198
  + ROUGE-2: 0.107
  + ROUGE-L: 0.192
  + ROUGE-Lsum: 0.185

These scores indicate a moderate level of overlap between the model’s responses and the reference answers, with ROUGE-1 being the highest, suggesting reasonable coverage of individual words and phrases.

* **BLEU Score**:
  + BLEU: 7.57e-12

The BLEU score is very low, indicating that the model's generated answers do not closely match the reference answers in terms of n-gram precision. This could be due to BERT's primary focus on extraction rather than generation.

* **F1 Score**:
  + F1: 0.197

The F1 score reflects the model’s accuracy, balancing precision and recall. The low score suggests that while BERT performs well in extracting relevant information, it may struggle with generating highly accurate and relevant answers.

**GPT-2 Model**

The GPT-2 model, known for its generative capabilities, produced the following results:

* **ROUGE Scores**:
  + ROUGE-1: 0.160
  + ROUGE-2: 0.039
  + ROUGE-L: 0.110
  + ROUGE-Lsum: 0.108

GPT-2’s ROUGE scores are generally lower compared to BERT, especially in ROUGE-2, indicating less overlap of longer n-grams and a more limited ability to capture nuanced information.

* **BLEU Score**:
  + BLEU: 0.00069

The BLEU score is also quite low, showing that GPT-2’s responses have limited n-gram precision relative to the reference answers. This suggests that while GPT-2 generates creative text, it may not be as accurate or relevant in this context.

* **F1 Score**:
  + F1: 0.156

The F1 score for GPT-2 is lower than that of BERT, highlighting its relatively weaker performance in providing precise and relevant answers.

**T5 Model**

The T5 model, utilized for both generation and transformation, achieved the following metrics:

* **ROUGE Scores**:
  + ROUGE-1: 0.415
  + ROUGE-2: 0.233
  + ROUGE-L: 0.416
  + ROUGE-Lsum: 0.411

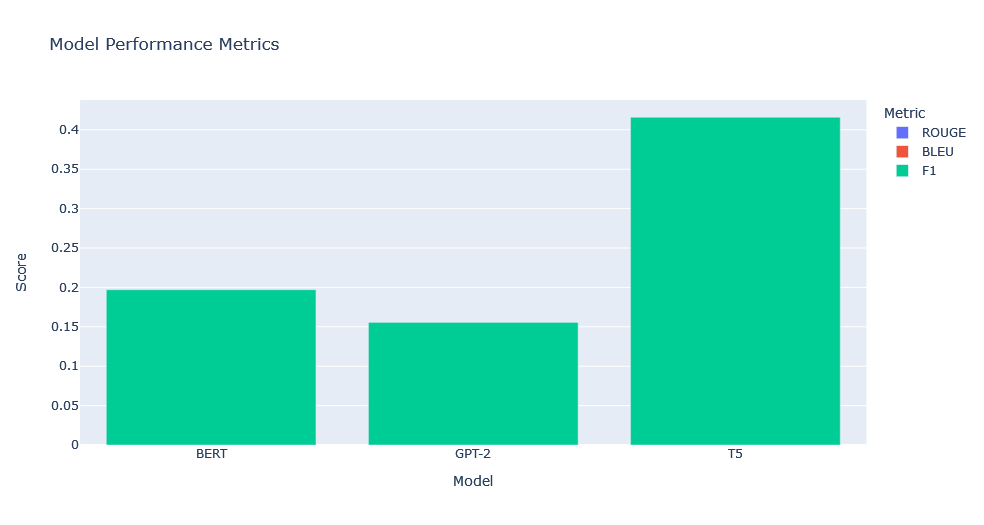
T5 demonstrated the highest ROUGE scores among the models, indicating strong performance in capturing both individual and longer n-gram overlaps with the reference answers.

* **BLEU Score**:
  + BLEU: 8.40e-05

Despite a low BLEU score, T5’s performance in generating responses is relatively better than GPT-2, although it still shows a gap compared to the reference answers in terms of n-gram precision.

* **F1 Score**:
  + F1: 0.416

The F1 score for T5 is the highest among the models, suggesting that T5 provides the most accurate and relevant answers in this evaluation. This reflects its capability to effectively refine and generate coherent responses.



**CONCLUSION**

The study of the BERT, GPT-2, and T5 models for question-answering tasks indicates distinct strengths and drawbacks for each. BERT excelled at extracting relevant information from the context, as evidenced by its superior ROUGE scores. However, its lower BLEU and F1 ratings show limits in producing extremely accurate and coherent results. BERT excels in identifying key details but struggles to provide responses that closely match reference answers.

In contrast, GPT-2, which is known for its generative skills, performed worse across all criteria. Its ROUGE, BLEU, and F1 scores were lower than those of BERT and T5, indicating that, while GPT-2 may generate creative and varied text, its replies lacked alignment and precision. This result demonstrates GPT-2's limitations in applications demanding a high degree of accuracy and relevance.

T5 outperformed the other models. Its better ROUGE and F1 scores demonstrate its stronger capacity to create and improve answers efficiently. T5's result demonstrates that it excels at delivering meaningful and correct replies, making it the best model for this question-answering task.  
  
According to the findings, a hybrid model approach could be useful for improving performance even further. Combining BERT's strength in information extraction, T5's capacity to improve answers, and GPT-2's creative production could capitalize on each model's distinct strengths, potentially enhancing overall performance. Furthermore, adding continuous learning methods to keep models up to current with user feedback and fresh data would aid in the long-term maintenance and improvement of accuracy.