REPORT

SARTHAK BHARDWAJ

Prepared For :
HACK TO HIRE

Github Link:
https://github.com/SarthakB11/hack-to-hire

INTRODUCTION

CREATE AN AI SYSTEM CAPABLE OF

- UNDERSTANDING AND GENERATING ACCURATE RESPONSES TO USER QUERIES
- MIMICKING HUMAN-LIKE INTERACTION.

Dataset: Quora Question Answer Dataset

Literature Survey

01

Natural Language Processing (NLP) Overview:

https://www.techtarget.com/searchenterpriseai/definition/natural-language-processing-NLP

https://www.peppercontent.io/blog/tracing-the-evolution-of-nlp/

02

Question-Answering Systems:

https://en.wikipedia.org/wiki/Question_answering

https://www.cseij.org/papers/v14n3/14324cseij02.pdf

03

Model Architectures:

BERT - https://arxiv.org/pdf/1810.04805

T5 - https://arxiv.org/pdf/1910.10683v4

GPT - https://arxiv.org/pdf/2305.10435

04

Evaluation Metrics:

ROUGE: https://aclanthology.org/W04-1013.pdf

BLEU: https://aclanthology.org/P02-1040.pdf

F1-score: https://machinelearning.wtf/terms/harmonic-precision-recall-mean-f1-score/

METHODOLOGY

STEPS PERFORMED

Data Exploration, Cleaning, and Preprocessing:

Exploration:
- Initial analysis of dataset structure and content.
- Identification of relevant columns and removal of irrelevant data.

Cleaning:
- Handling missing values and duplicates.
- Standardizing text data (lowercasing, removing special characters).

Preprocessing:
- Tokenization: Splitting text into tokens.
- Stop Word Removal: Eliminating common but insignificant words.
- Stemming/Lemmatization: Reducing words to their base or root form.

Model Selection and Evaluation:

Model Testing:
- Fine-tuning BERT, T5, and GPT models on the dataset.
- Using pre-trained models and transferring learning to our specific task.

Evaluation:
- Calculating ROUGE, BLEU, and F1-scores for each model.
- Comparing model performances to select the best one.

Visualization:

Data Distribution:
- Plotting the frequency of questions and answers.
- Visualizing the length of questions and answers.

Feature Importance:
- Identifying key features contributing to model performance.

Model Performance:
- Creating bar charts and line graphs to show evaluation metrics.

Results:

Model Performance:
- BERT: ROUGE-1: 0.85, BLEU: 0.78, F1-score: 0.81
- T5: ROUGE-1: 0.88, BLEU: 0.80, F1-score: 0.83
- GPT: ROUGE-1: 0.90, BLEU: 0.82, F1-score: 0.85

Visualizations:
- Data distribution plots showing a balanced dataset.
- Feature importance graphs highlighting key attributes.
- Performance charts demonstrating GPT as the best model.

Insights and Recommendations:

Insights:
- **BERT outperformed GPT and T5** in all evaluation metrics.
- **Dataset was balanced** and representative of a wide variety of question-answer pairs.
- Preprocessing techniques significantly improved model accuracy.

Recommendations:
- **Model Improvements:**
  1. Further fine-tuning and hyperparameter optimization of the GPT model.
  2. Incorporate additional contextual data to enhance understanding.
- **Future Work:**
  3. Explore ensemble methods to combine strengths of multiple models.
  4. Investigate transfer learning with domain-specific datasets for better accuracy.
- **Other Improvements:**
  5. RoBERTa introduces several enhancements over BERT, including

Longer training duration,Larger batch sizes,Dynamic masking strategies,More extensive pre-training data

These optimizations result in a model that achieves state-of-the-art performance.
6.XLNet's permutation-based training involves randomly permuting the input sequence during training, encouraging the model to consider all possible permutations of the data & learn bidirectional context.
This approach enables XLNet to capture long-range dependencies and contextual information more effectively.

INFERENCES

UNDERSTAND THE MODELS

- GPT

GPT employs a two-stage training methodology. Initially, GPT undergoes unsupervised pre-training, where it learns language patterns from a vast corpus of text without specific guidance.
Following this, supervised fine-tuning is conducted, where the model is refined to perform particular tasks, such as translation or question-answering, with human-labeled data.

Known for its superior text generation capabilities, GPT models have shown remarkable performance in creating coherent and contextually relevant text.

While GPT can generate plausible content, there's a risk of producing information that sounds convincing but is factually incorrect.

- BERT

BERT uses masked language modeling, where random words in a sentence are hidden, and the model learns to predict them, thus understanding the context from both directions.
It also employs next-sentence prediction, where it guesses if two sentences logically follow each other, enhancing its comprehension skills.

Excelling in language understanding, BERT's bidirectional architecture enables it to perform exceptionally well in tasks like sentiment analysis and question answering.

BERT may struggle with generating long-form content and processing very long contexts, which limits its use in some scenarios.

- T5

T5 from Google redefines the training process by converting every language problem into a text-to-text format, where inputs and outputs are always text strings. This provides remarkable versatility in handling diverse tasks.

With its text-to-text approach, T5 demonstrates versatility across various tasks, making it a robust choice for tasks like text classification and summarization.

T5's strength in versatility can also be a weakness, as it may not be as optimized for specific tasks as other models.