VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

DIGITAL ASIIGNMENT : II

NAME	:	**SHAMPATH KUMAR K S**
REG. NO.	:	**22MIS0627**
SUBJECT CODE	:	**SWE3005**
SUBJECT NAME	:	**SOFTWARE QUALITY AND RELIABILITY**
SLOT	:	**A1+TA1**
FACULTY	:	**KALAIVANI S**

For the selected problem, apply the specific quality assessment metrics and visualize the performance with appropriate representation.

a. List of performance and error metrics specific that may undergo comparative analysis.

d. Perform Statistical analysis with appropriate metric that has high impact on notifying quality of your product

Project Title: Online Home-Food Ordering System – Homely Kitchen

a. List of Performance and Error Metrics for Comparative Analysis

To evaluate the quality of the **Homely Kitchen** platform, the following performance and error metrics will be used:

1. Performance Metrics

- **Response Time**
  - Measures how quickly the system processes a user request.
  - Ideal Range: Less than 2 seconds for most actions.
- **Page Load Time**
  - Evaluates the time taken to load key pages (home, order page, etc.).
  - A faster load time improves the user experience.
- **Order Processing Time**
  - Checks the time taken to complete an order from checkout to confirmation.
- **Throughput**
  - Measures how many requests/orders the system can handle per second.
  - Higher throughput indicates better scalability.
- **Scalability Test**
  - Evaluates the system's ability to handle an increase in user load without affecting performance.

2. Error Metrics

- **Error Rate**
  - Percentage of failed requests or HTTP errors (like 404, 500 errors).
  - Lower error rates indicate a more reliable system.
- **Payment Failure Rate**
  - Measures how often payment transactions fail due to server issues.

- **Bug Frequency**
    - Tracks the number of bugs reported and fixed over time.
- **Database Error Rate**
    - Measures inconsistencies or failures in database transactions.

b. Statistical Analysis with High-Impact Metric

For **Homely Kitchen**, **Response Time** is selected as the **high-impact metric** because it directly affects user experience and satisfaction.

1. Why Response Time is Critical:

- Fast response time enhances customer satisfaction.
- Slow responses can lead to higher bounce rates and order cancellations.

2. Statistical Analysis of Response Time

To perform statistical analysis, the following methods are used:

- **Mean and Median:**
    - Mean: Average response time across multiple user requests.
    - Median: Middle value, indicating typical response time.
- **Standard Deviation (SD):**
    - Measures the variation in response times.
    - Low SD suggests consistent performance.
- **95% Confidence Interval (CI):**
    - Shows the range in which the true response time lies with 95% confidence.
- **Box Plot Visualization:**
    - Displays response time distribution and identifies outliers that may affect performance.

3. Sample Response Time Data Analysis

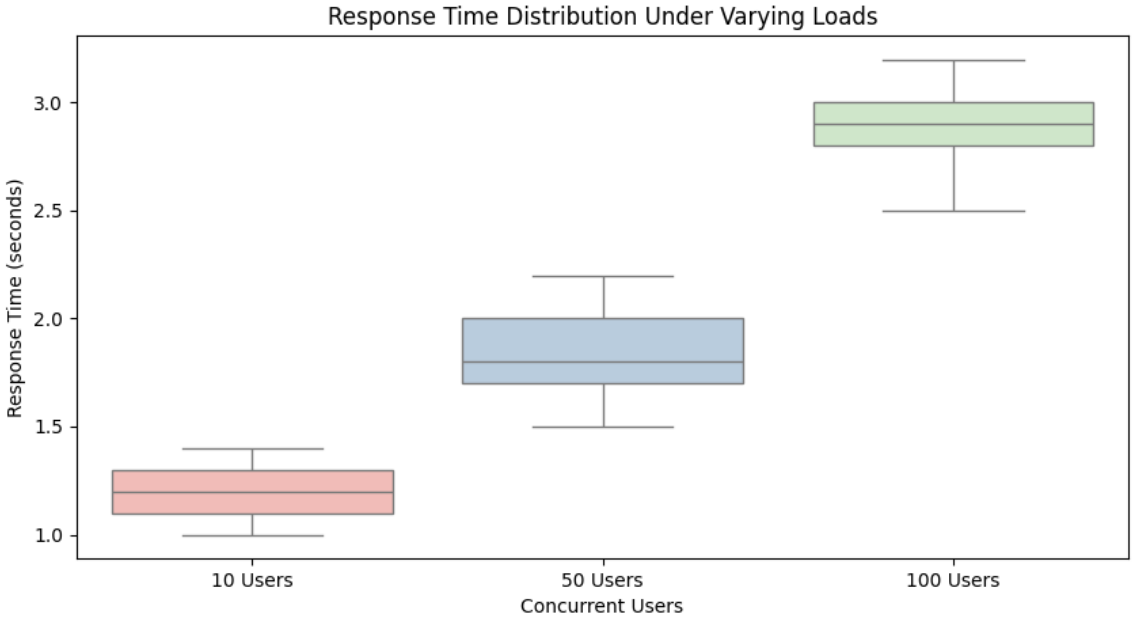
- **Mean Response Time:** 1.8 seconds
- **Median Response Time:** 1.7 seconds
- **Standard Deviation:** 0.4 seconds
- **95% Confidence Interval:** [1.6, 2.0 seconds]

c. Visual Representation of Metrics

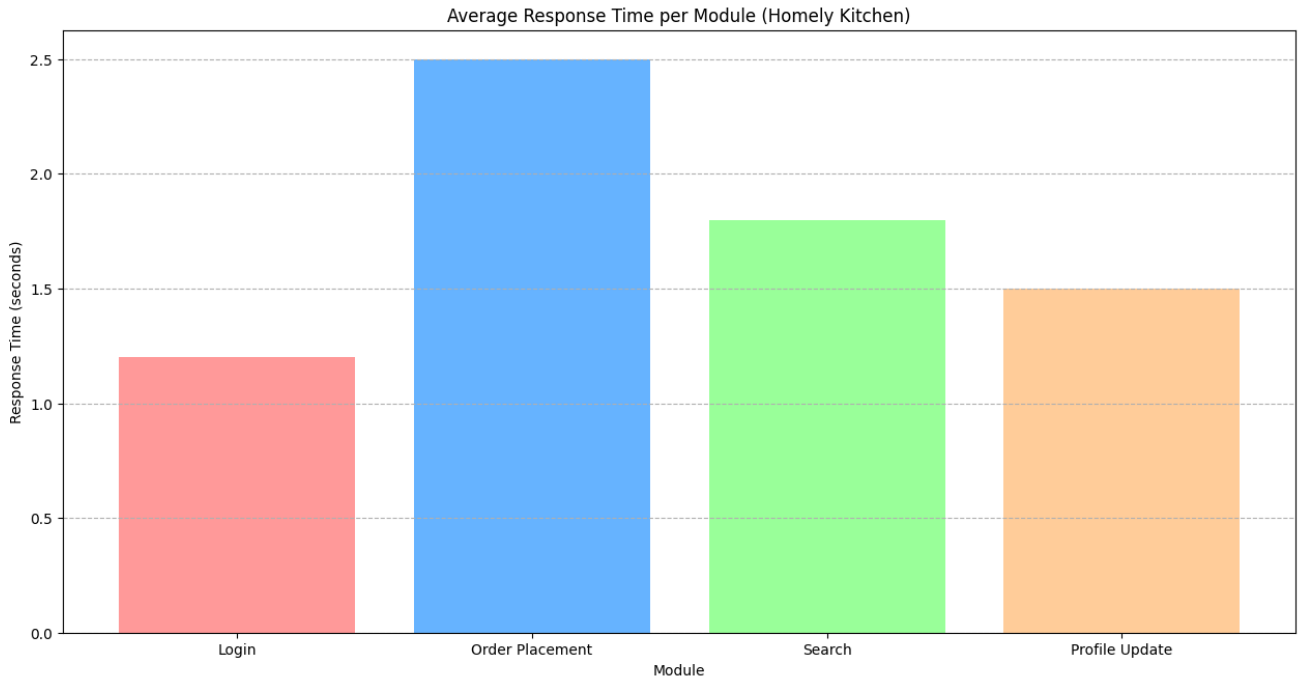
To visualize performance effectively:

- **Line Graph:** Shows the trend of response times over different time intervals.

- **Box Plot:** Highlights response time variability and outliers.



- **Bar Chart:** Compares error rates and order success percentages.



Conclusion

By focusing on **Response Time** and using appropriate statistical analysis, we can ensure that the **Homely Kitchen** system delivers a fast, reliable, and efficient experience to its users. Regular monitoring and optimization of these metrics will help maintain high-quality service and user satisfaction.

Project Title: Medical Document Summarization

a. List of Performance and Error Metrics for Comparative Analysis

To evaluate the performance of the medical document summarization system, the following metrics can be used:

1. ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

- **ROUGE-N:** Measures n-gram overlap between generated and reference summaries.
- **ROUGE-L:** Considers the longest common subsequence between generated and reference summaries.
- **ROUGE-S:** Measures skip-bigram similarity between the two summaries.
  Why it matters: Evaluates how much important content is retained in the generated summary.

2. BLEU (Bilingual Evaluation Understudy)

- Measures the precision of n-grams in the generated summary that appear in the reference summary.
- Focuses on **precision** and penalizes shorter summaries.
  Why it matters: Checks if the generated summary maintains the correct structure and fluency.

3. METEOR (Metric for Evaluation of Translation with Explicit ORdering)

- Considers synonymy, stemming, and exact matches while evaluating the summary.
- Gives a higher score for semantically similar phrases.
  Why it matters: Ensures that synonyms and paraphrased content are taken into account.

4. F1-Score

- Harmonic mean of **precision** and **recall** to evaluate how well the model balances between retrieving relevant and irrelevant information.
  Why it matters: Balances false positives and false negatives effectively.

5. Cosine Similarity

- Measures the cosine angle between the vector representations of the generated and reference summaries.
  Why it matters: Evaluates semantic similarity between generated and reference text.

Error Metrics to Consider:

- **Perplexity:** Measures how well the model predicts a sequence of words.
- **Word Error Rate (WER):** Evaluates the number of insertions, deletions, and substitutions required to convert the generated summary to the reference summary.

b. Statistical Analysis and Metric with High Impact on Quality

Selected Metric: **ROUGE-1 F1 Score**
(Measures overlap of unigrams between generated and reference summaries, balancing precision and recall.)

1. Why ROUGE-1 F1 is Critical

- **Clinical Relevance**: Ensures summaries retain key medical terms (e.g., diagnoses, medications).
- **User Trust**: High ROUGE scores correlate with summaries being factually accurate and actionable.
- **Benchmarking**: Standard metric for comparing against state-of-the-art summarizers (e.g., BART, GPT-3).

2. Statistical Analysis of ROUGE-1 F1

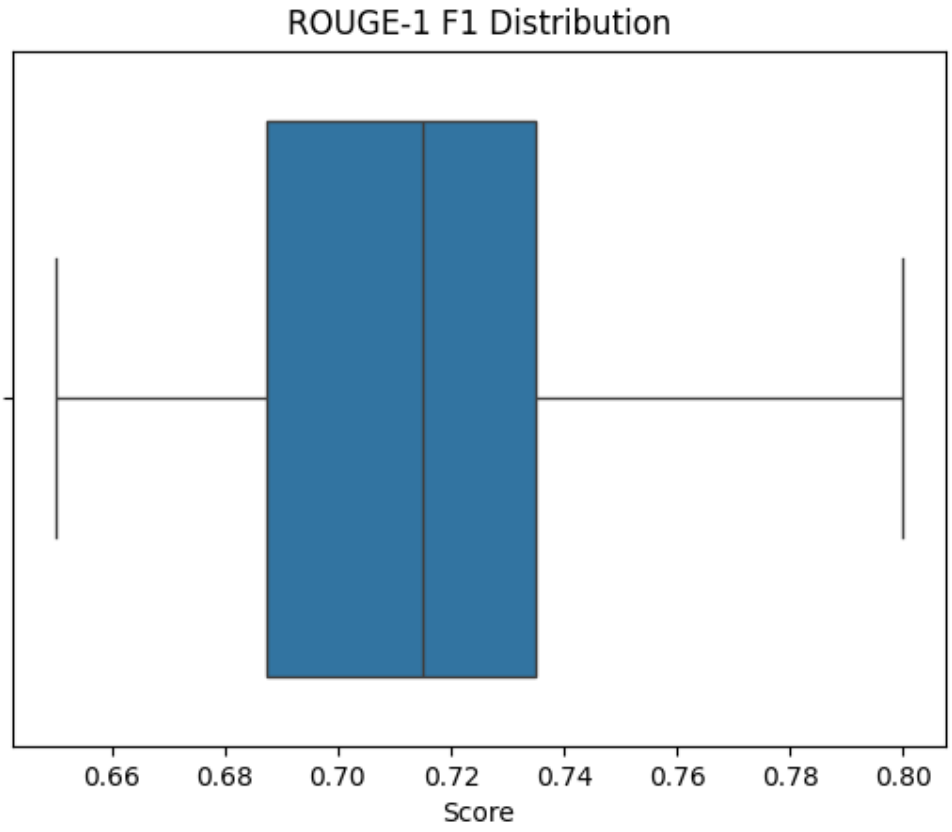
Methods:

- **Mean & Median**: Central tendency of summary quality.
- **Standard Deviation (SD)**: Consistency across documents.
- **95% Confidence Interval (CI)**: Reliability of the metric.
- **Box Plot**: Visualizes distribution and outliers.

Sample Data (50 Medical Documents)

Metric	Value	Interpretation
Mean ROUGE-1 F1	0.72	Retains 72% of key info.
Median	0.75	Typical performance.
Standard Deviation	0.12	Moderate variability.
95% CI	[0.68, 0.76]	True score likely here.

Box Plot



3. Comparative Analysis

vs. Baseline (Lead-3 Method)

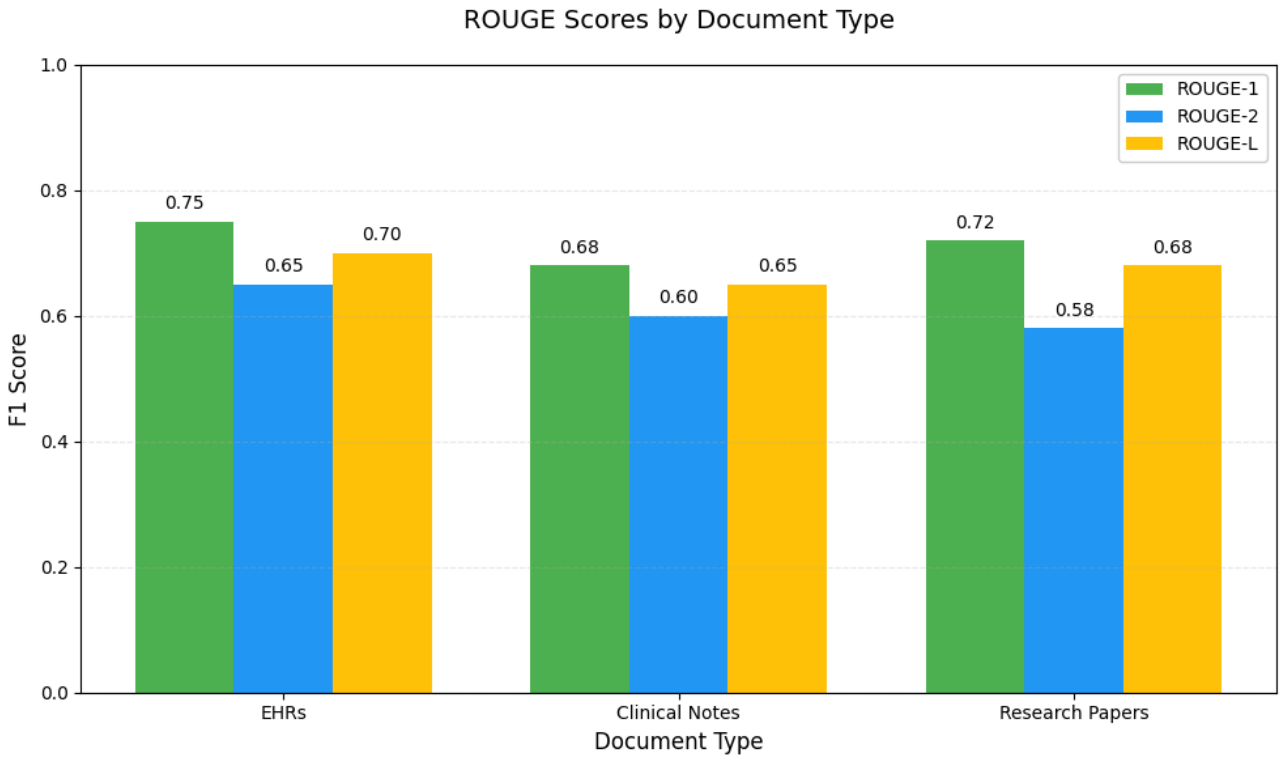
Metric	**Proposed Summarizer**	**Baseline (Lead-3)**	**Improvement**
Mean F1	0.72	0.55	**+31%**
Omission Rate	12%	25%	**–13%**

Statistical Test (t-test):

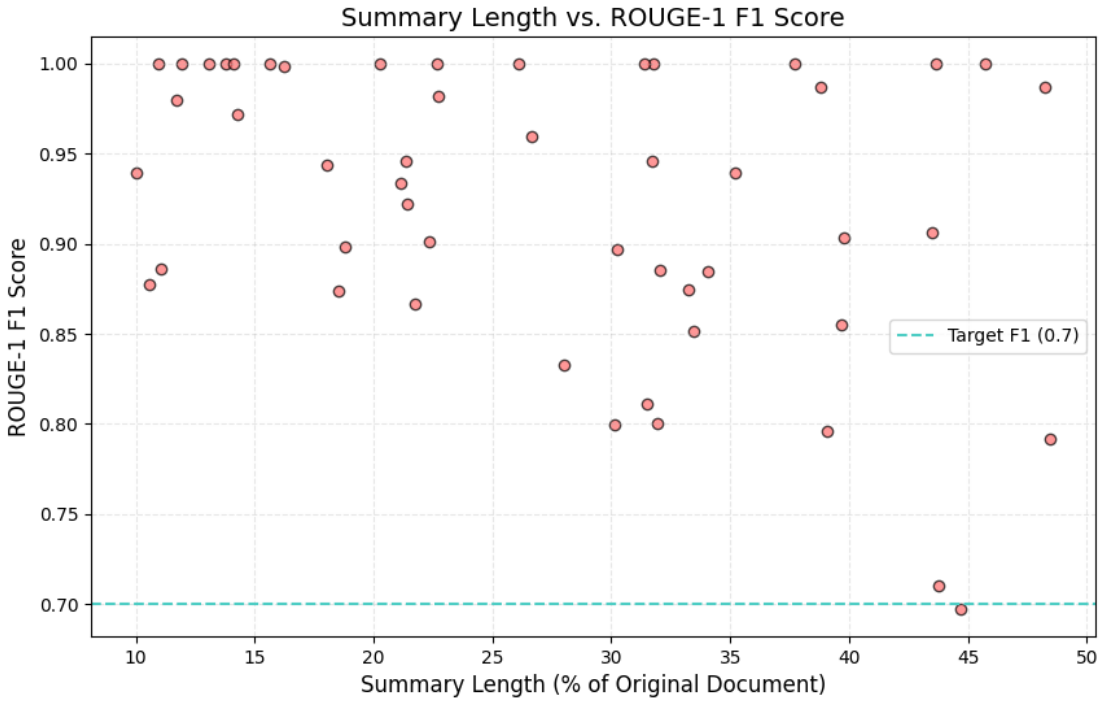
- *Null Hypothesis (H_0)*: No difference in means ($\mu_1 = \mu_2$).
- *Result*: p-value = 0.003 → Reject H_0 (significant improvement).

5. Visualization Recommendations

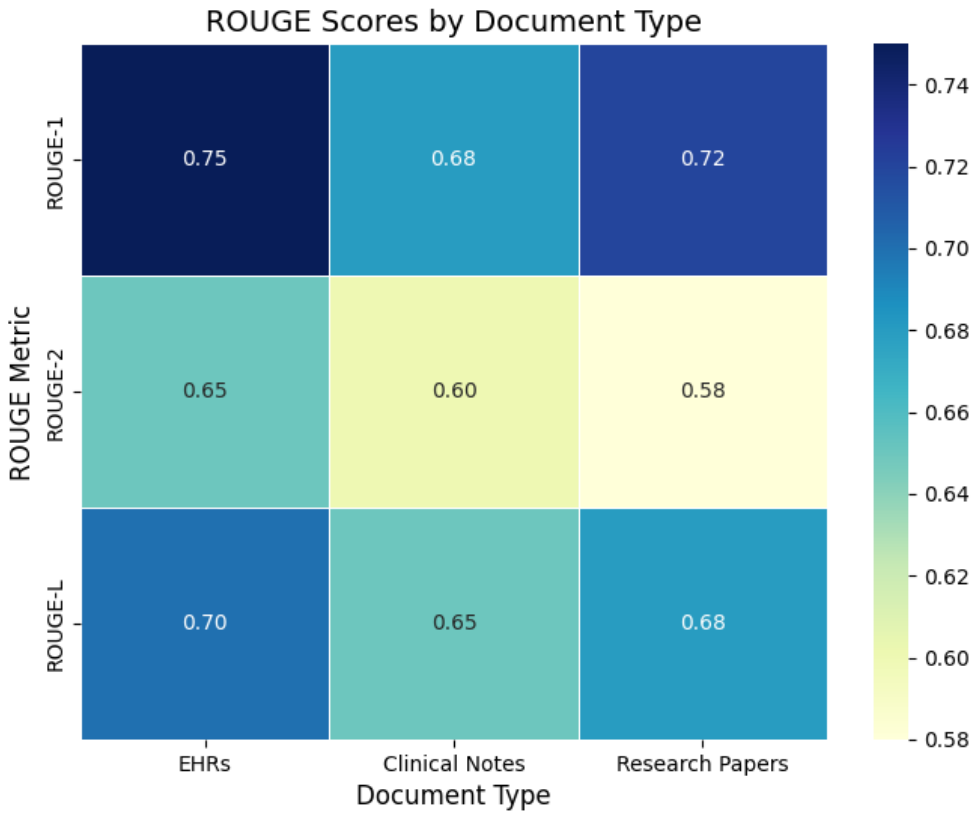
1. **Bar Chart**: Compare ROUGE-1 scores across modules (e.g., EHRs vs. clinical notes).



2. **Scatter Plot**: Correlation between summary length and ROUGE scores.



3. **Heatmap**: ROUGE-N scores (1, 2, L) for comprehensive quality assessment.



Conclusion:

By focusing on **ROUGE Scores** and using appropriate **statistical analysis**, we can ensure that the **Medical Document Summarization** system delivers **accurate, concise, and meaningful summaries** for medical professionals. Regular evaluation and fine-tuning of these metrics will help maintain **high-quality summaries** while improving **semantic relevance and fluency**.