SARVAJANIK COLLEGE OF ENGINEERING AND TECHNOLOGY DEPARTMENT NAME:ARTIFICIAL INTELLIGENCE AND DATA SCIENCE SUBJECT: PROJECT-I CODE: BTAI16701

Intelligent Text Assistant

Presented By:

1. ET21BTAI009 [Manav Desai]
2. ET21BTAI019 [Prayesh Godhani]
3. ET21BTAI036 [Rohit Kunjadiya]
4. ET21BTAI044 [Parthiv Moradiya]
5. ET21BTAI803 [Dhrumil Gabani]

Supervised By:
Prof. Karishma Desai

INTRODUCTION TO INTELLIGENT TEXT ASSISTANT

Overview:

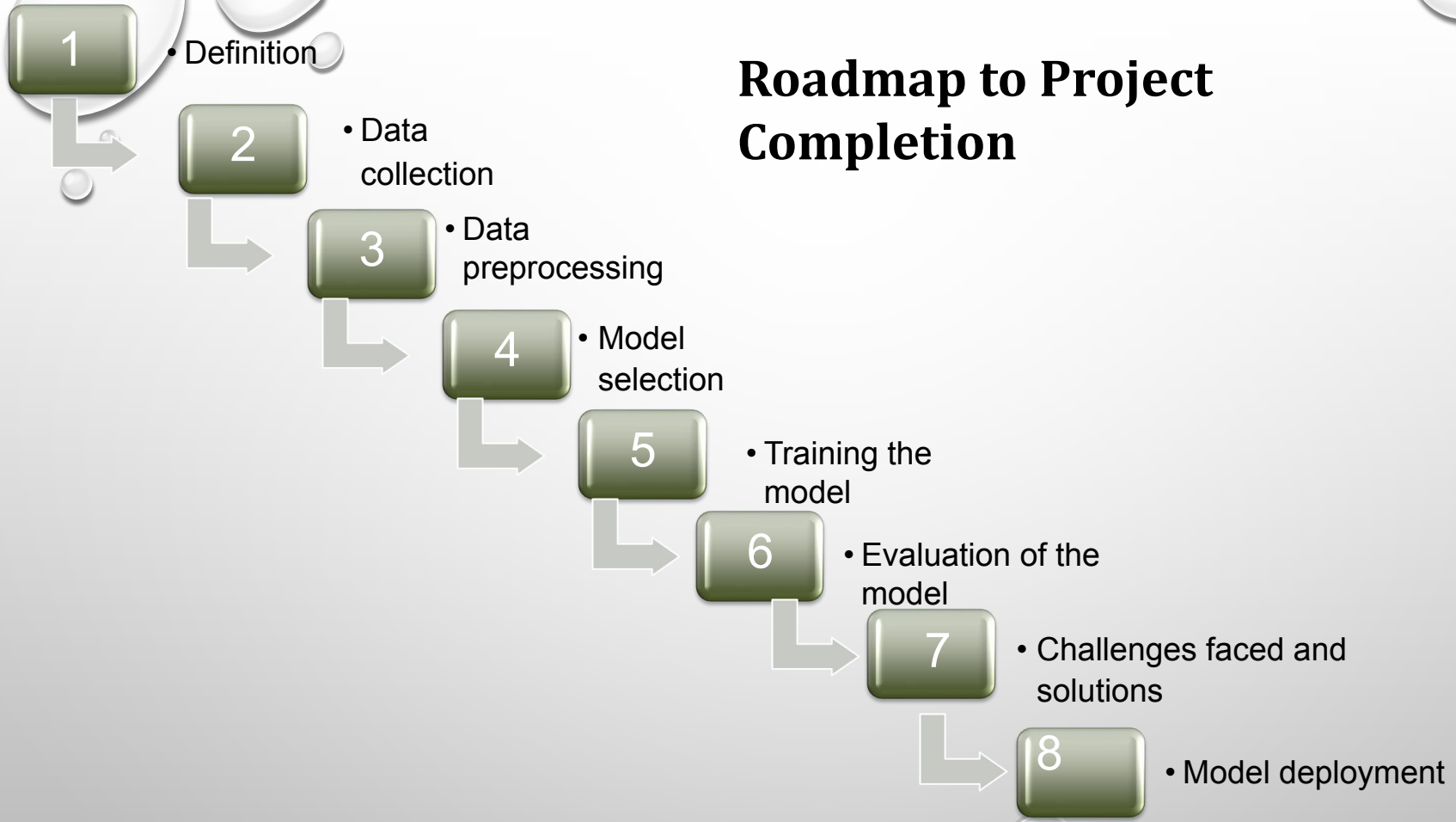
The Intelligent Text Assistant (ITA) project uses natural language processing (NLP) to make text easier to work with. ITA helps users quickly summarize documents and answer questions from text files like :

1. Research papers
2. News articles
3. Blogs
4. PDFs

Project Approach:

1. **Development of a NLP Models:**
   - Focus on fine-tuning pre-trained models for summarization and QA tasks.
   - Provides clear and relevant answers to user questions based on the text content.
2. **User Interface:**
   - Allows users to upload text files or PDFs and receive summarizations or answers to specific questions.
   - Easy process for uploading documents and generating results quickly.
   - Easy-to-use interface for accurate results.

Roadmap to Project Completion



INTRODUCTION TO NLP MODELS

THE INTELLIGENT TEXT ASSISTANT (ITA) USES ADVANCED NLP MODELS TO IMPROVE TEXT-RELATED TASKS. IT HAS TWO MAIN FEATURES:

1. **TEXT SUMMARIZATION:**
   - PULLS OUT THE KEY POINTS FROM PDFS AND TEXT FILES.
   - HELPS USERS UNDERSTAND THE MAIN IDEAS QUICKLY WITHOUT HAVING TO READ LONG DOCUMENTS.
2. **QUESTION-ANSWERING (QA) SYSTEM**:
   - ANSWERS USER QUESTIONS BY ANALYZING THE CONTENT OF UPLOADED PDFS OR TEXT.
   - PROVIDES QUICK, RELEVANT RESPONSES TO SAVE TIME AND IMPROVE EFFICIENCY.

TEXT SUMMARIZATION

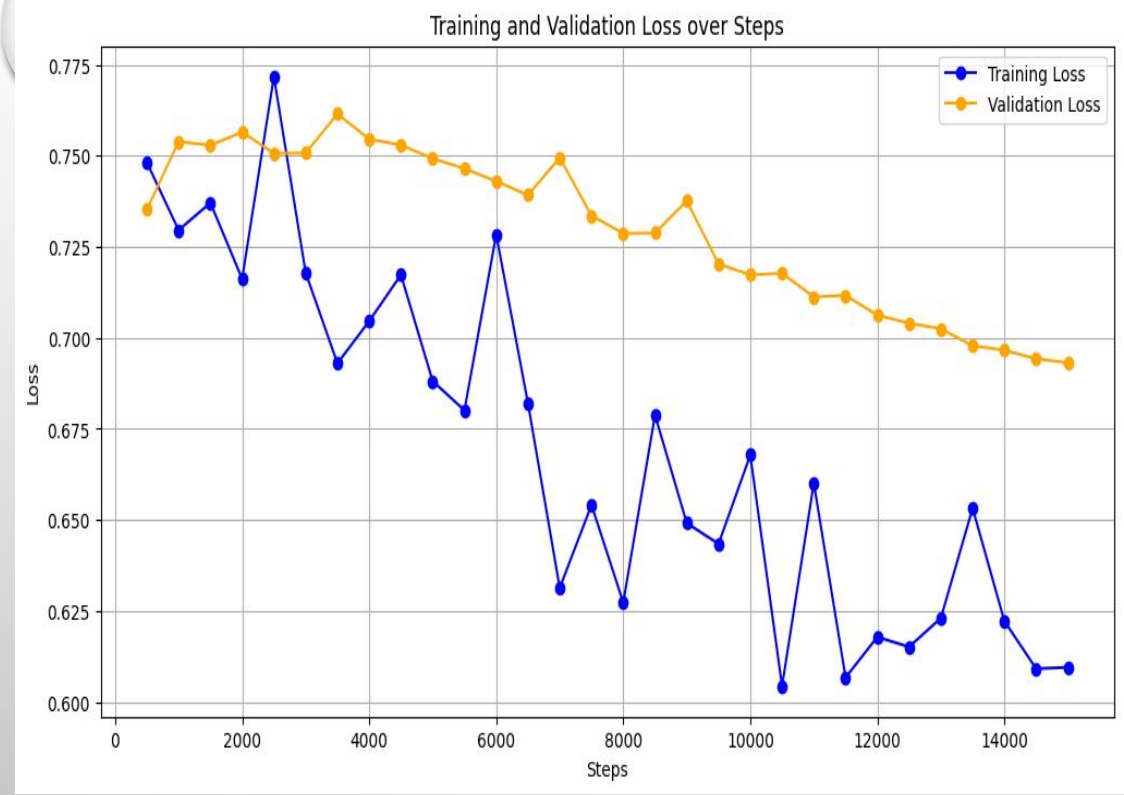
- Text summarization is an essential task in natural language processing that helps users quickly understand large volumes of information.
- It involves condensing text from documents, articles, or reports into shorter, meaningful summaries.
- We gather various text sources, including research papers, blogs, and news articles.
- Each text is analyzed to extract key points, enabling users to grasp the main ideas without reading the entire content. This forms the basis for our summarization model.

CONTINUOUS…

Train Dataset Size	287113
Validation Dataset Size	13368
Test Dataset Size	11490

- We tested various deep learning models for text summarization, and **DistilBERT** emerged as the best, samller and faster.
- We trained the **DistilBERT** model for **15000 steps**, after which it reached its optimal performance, proving to be the best model for breast cancer classification.

Metric	Accuracy
ROUGE-1	42%
ROUGE-2	20%
ROUGE-L	29%
BLUE SCORE	13%



Model	RG-1	RG-2	RG-L
Lead-1	16.30	1.60	11.95
BERTSUM	38.81	16.50	31.27
MATCHSUM	24.86	4.66	18.41
PGNet	29.70	9.21	23.24
PGNet+Cov	28.10	8.02	21.72
BART	45.14	22.27	37.25

```
Average ROUGE-1 Score: 0.4200
Average ROUGE-2 Score: 0.1914
Average ROUGE-L Score: 0.2933
Average BLEU Score: 12.7138
```

QUESTION ANSWERING SYSTEM

- A **question-answering (QA)** system is designed to provide precise answers to user inquiries based on a given text or document.
- We use the DistilBERT pre-trained model for our question-answering (QA) system, leveraging its capabilities to understand and respond to user inquiries effectively. Additionally, we save user questions in a database to continuously improve the model's performance and enhance the user experience.
- This approach enhances our system's ability to provide relevant answers based on the context of the text, ensuring accurate and meaningful responses.

CONTINUOUS…

- We used the **BERT** model for a Question Answering (**QA**) system.
- The model leverages transformers, which use self-attention mechanisms to capture relationships between words in a sentence. This allows for a better understanding of context compared to previous sequential models.
- **Pre-trained** on a large corpus of **text**, making it effective for various NLP tasks.
- The model can effectively **handle text extracted from PDFs** and other structured QA texts.
- The system processes the input and generates answers based on the **contextual information**.

Research Paper : BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	**Average** -
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

System	Dev	Test
ESIM+GloVe	51.9	52.7
ESIM+ELMo	59.1	59.2
OpenAI GPT	-	78.0
$\text{BERT}_{\text{BASE}}$	81.6	-
$\text{BERT}_{\text{LARGE}}$	**86.6**	**86.3**
Human (expert)†	-	85.0
Human (5 annotations)†	-	88.0

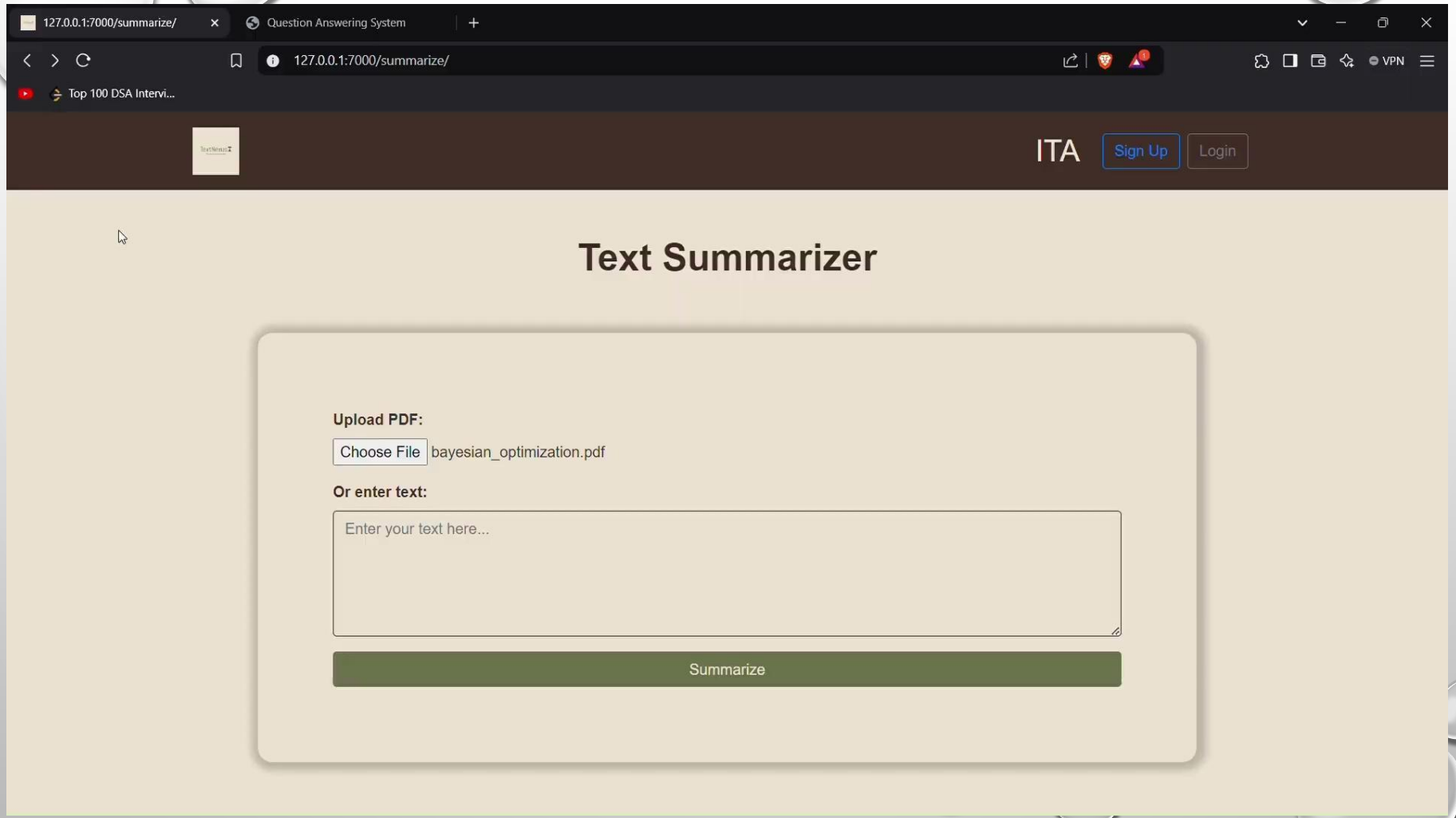
FLASK APP

1. Flask is a **lightweight web framework** for Python that enables developers to **build web applications** quickly and efficiently. It is designed to be **simple and flexible**, allowing for easy integration of various components and extensions.

2. In this app, we have implemented two main functionalities: text summarization and question-answering. Each functionality is organized into separate modules, using Flask Blueprints to ensure modularity and maintainability.

3. The app allows users to **upload PDF** and **text**, from which it generates concise **summaries and provides answers to specific questions** based on the content.

4. Users can interact with the application through a user-friendly interface, making it easy to access and utilize the summarization and QA features effectively.


127.0.0.1:7000/summarize/
Question Answering System
127.0.0.1:7000/summarize/
VPN
Top 100 DSA Intervi...
ITA
Sign Up
Login
Text Summarizer
Upload PDF:
Choose File
bayesian_optimization.pdf
Or enter text:
Enter your text here...
Summarize

Question Answering System

Enter Text Here:

Paste your text here...

Or Upload PDF (optional):

Choose File No file chosen

Enter your question:

Type your question here...

Add Another Question | Submit Questions

Reset Session

Answers

Q: what is ESACL?
A: Enhanced Seq2Seq Autoencoder via Contrastive Learning

UPDATED TIMELINE

Month	Task/Activity
August	- Data collection and preprocessing
	- Build and refine custom NLP model architecture
	- Initial training of the model
September	- Continue model training and tuning (hyperparameter tuning)

CONTINUOUS…

Month	Task/Activity
	- Validate models with early results
	-Selecting pretrained models like gpt2, DistilBERT, google-Pegasus-large, BERT, T5
	- Begin integrating the model into the web application
October	- Robust testing and validation of CNN models

CONTINUOUS…

Month	Task/Activity
	- Final model adjustments based on feedback
	- Prepare final deployment of the web application
	- Create final documentation and presentation materials
	- Collect feedback from test users

FUTURE WORK

- **Enhanced Summarization**: We will improve our models to generate even clearer and more concise summaries from various text sources.
- **User Feedback Integration**: We plan to implement a feedback mechanism to continuously enhance model performance based on user input.
- **Model Optimization:** We will strive to develop the most accurate and fastest models possible.
- **Mobile App Development**: We will explore creating a mobile application to make our services more accessible to users on the go.
- **Improved QA Capabilities**: We will refine the question-answering system to provide more accurate and context-aware responses.

CONCLUSION

We have successfully created the Intelligent Text Assistant (ITA), which employs advanced NLP models for effective text summarization and question-answering. Our models are designed for clarity and relevance, ensuring users can quickly access key information. The application features an intuitive interface, making it easy for users to interact with the system. Looking ahead, we plan to enhance summarization techniques, expand document support, and refine QA capabilities. Overall, our project lays a strong groundwork for future developments in text processing technology.

Thank You