PROJECT FOR

CONVERSATIONAL AI:
NATURAL LANGUAGE PROCESSING (UCS664)

Punjabi Question Answering System Using Transformers

Submitted by:

Rakshit Singh 102203496

Gitika Goyal 102383012

Shruti Dixit 102203532

Submitted to:

Dr. Jasmeet Singh

Computer Science and Engineering Department

Thapar Institute of Engineering and Technology, Patiala - 147004

June -2025

INTRODUCTION

In recent years, advances in Natural Language Processing (NLP) have made it possible to build intelligent systems capable of understanding and answering questions from textual data. However, most Question Answering (QA) systems are developed in English or other high-resource languages, leaving regional languages like Punjabi underrepresented.

This project aims to bridge that gap by implementing a Punjabi Question Answering System that leverages the multilingual capabilities of the XLM-RoBERTa transformer model. The system allows users to input a paragraph and a question in Punjabi and receive an extracted answer with a confidence score, along with the highlighted context for better interpretability.

The user interface is built using Gradio, enabling real-time interaction in the browser with no backend setup required. The model is run using PyTorch and the Hugging Face Transformers library, demonstrating the flexibility and power of zero-shot and multilingual NLP.

DATASET USED

This project does not use a fixed training dataset but builds on top of a pretrained transformer model. Here's how the dataset context is structured:

Pretrained Dataset Used:

SQuAD2.0 (Stanford Question Answering Dataset v2)

The backbone model (deepset/xlm-roberta-large-squad2) is fine-tuned on the SQuAD2.0 dataset. Contains over 100,000 question-context-answer pairs and also includes unanswerable questions to improve model robustness.

Language: English

Format: JSON with fields like context, question, answers, and is_impossible.

Dataset Used in Practice (During Execution):

User-provided Punjabi text is used and no static Punjabi dataset is created. Users input context paragraphs and questions in Punjabi directly into the interface.

The model performs zero-shot inference, leveraging multilingual training to process and answer in Punjabi.

Language: Punjabi (Gurmukhi script)

Format: Text string input (context and question) via Gradio

MODEL USED

Model Name:

deepset/xlm-roberta-large-squad2

Architecture:

This model is based on XLM-RoBERTa Large, a transformer-based multilingual masked language model (MLM). It has been pre-trained on a large corpus of text from 100+ languages, including Punjabi, making it well-suited for multilingual natural language processing (NLP) tasks.

Fine-tuning objective:

The model is specifically fine-tuned for the task of extractive question answering on the SQuAD2.0 dataset. It is capable of predicting the start and end token positions of the answer within a given context passage.

Key Features:

- Supports context–question pair inputs of variable lengths.
- Handles no-answer questions, which are common in real-world scenarios.
- Offers robust multilingual understanding, enabling QA tasks in non-English languages, including Punjabi.
- Can be integrated into interactive applications such as Gradio-based QA interfaces.

RESULTS

ਪੰਜਾਬੀ ਵਿੱਚ ਸਵਾਲ ਪੁੱਛੋ ਅਤੇ ਉੱਤਰ ਪ੍ਰਾਪਤ ਕਰੋ!

ਪੰਜਾਬੀ ਪੈਰਾ (Context)

ਪੰਜਾਬ ਭਾਰਤ ਦੇ ਉੱਤਰ-ਪੱਛਮੀ ਹਿੱਸੇ ਵਿੱਚ ਸਥਿਤ ਇੱਕ ਰਾਜ ਹੈ। ਇਹ ਰਾਜ ਖੇਤੀਬਾੜੀ, ਸੱਭਿਆਚਾਰ, ਅਤੇ ਭੰਗੜੇ ਲਈ ਪ੍ਰਸਿੱਧ ਹੈ। ਪੰਜਾਬ ਦੀ ਰਾਜਧਾਨੀ ਚੰਡੀਗੜ੍ਹ ਹੈ। ਸਤਲੁਜ, ਬਿਆਸ, ਅਤੇ ਰਾਵੀ ਇੱਥੇ ਦੀਆਂ ਪ੍ਰਮੁੱਖ ਦਰਿਆਵਾਂ ਹਨ। ਇੱਥੇ ਦੀ ਭਾਸ਼ਾ ਪੰਜਾਬੀ ਹੈ, ਜੋ ਗੁਰਮੁਖੀ ਲਿਪੀ ਵਿੱਚ ਲਿਖੀ ਜਾਂਦੀ ਹੈ।

ਪ੍ਰਸ਼ਨ (Question)

ਪੰਜਾਬ ਭਾਰਤ ਵਿੱਚ ਕਿੱਥੇ ਸਥਿਤ ਹੈ?

Clear

Submit

ਉੱਤਰ & ਵਿਸ਼ਵਾਸ ਪੱਧਰ

- ਉੱਤਰ: ਉੱਤਰ-ਪੱਛਮੀ ਹਿੱਸੇ
- ਵਿਸ਼ਵਾਸ-ਪੱਧਰ: 0.74

ਪੰਜਾਬ ਭਾਰਤ ਵਿੱਚ ਕਿੱਥੇ ਸਥਿਤ ਹੈ? ਪੰਜਾਬ ਭਾਰਤ ਦੇ **ਉੱਤਰ-ਪੱਛਮੀ ਹਿੱਸੇ** ਵਿੱਚ ਸਥਿਤ ਇੱਕ ਰਾਜ ਹੈ। ਇਹ ਰਾਜ ਖੇਤੀਬਾੜੀ, ਸੱਭਿਆਚਾਰ, ਅਤੇ ਭੰਗੜੇ ਲਈ ਪ੍ਰਸਿੱਧ ਹੈ। ਪੰਜਾਬ ਦੀ ਰਾਜਧਾਨੀ ਚੰਡੀਗੜ੍ਹ ਹੈ। ਸਤਲੁਜ, ਬਿਆਸ, ਅਤੇ ਰਾਵੀ ਇੱਥੇ ਦੀਆਂ ਪ੍ਰਮੁੱਖ ਦਰਿਆਵਾਂ ਹਨ। ਇੱਥੇ ਦੀ ਭਾਸ਼ਾ ਪੰਜਾਬੀ ਹੈ, ਜੋ ਗੁਰਮੁਖੀ ਲਿਪੀ ਵਿੱਚ ਲਿਖੀ ਜਾਂਦੀ ਹੈ।

Flag

ਪੰਜਾਬੀ ਵਿੱਚ ਸਵਾਲ ਪੁੱਛੋ ਅਤੇ ਉੱਤਰ ਪ੍ਰਾਪਤ ਕਰੋ!

ਪੰਜਾਬੀ ਪੈਰਾ (Context)

ਕ੍ਰਿਤ੍ਰਿਮ ਬੁੱਧੀ ਜਾਂ ਆਰਟੀਫੀਸ਼ਲ ਇੰਟੈਲੀਜੈਂਸ (AI) ਇੱਕ ਤਕਨਾਲੋਜੀ ਹੈ ਜੋ ਕੰਪਿਊਟਰਾਂ ਨੂੰ ਮਨੁੱਖੀ ਬੁੱਧੀ ਵਰਗਾ ਸੋਚਣ ਅਤੇ ਫੈਸਲੇ ਲੈਣ ਯੋਗ ਬਣਾਉਂਦੀ ਹੈ। ਇਸ ਵਿਚ ਮਸ਼ੀਨ ਲਰਨਿੰਗ, ਡੀਪ ਲਰਨਿੰਗ, ਨੈਚਰਲ ਲੈਂਗਵਿਜ ਪ੍ਰੋਸੈਸਿੰਗ ਅਤੇ ਕੰਪਿਊਟਰ ਵਿਜ਼ਨ ਵਰਗੀਆਂ ਵਿਧੀਆਂ ਸ਼ਾਮਲ ਹਨ। AI ਦਾ ਉਦੇਸ਼ ਐਸੀਆਂ ਮਸ਼ੀਨਾਂ ਤਿਆਰ ਕਰਨਾ ਹੈ ਜੋ ਸਵੈ-ਸੀਖਣ, ਸਵੈ-ਨਿਰਣੈ ਅਤੇ ਮਨੁੱਖੀ ਤਰ੍ਹਾਂ ਕੰਮ ਕਰਨ ਵਿੱਚ ਸਮਰਥ ਹੋਣ। AI ਦੀ ਵਰਤੋਂ ਅੱਜਕੱਲ੍ਹ ਹੈਲਥਕੇਅਰ, ਆਟੋਮੋਬਾਈਲ, ਫਾਈਨੈਂਸ, ਖੇਤੀਬਾੜੀ ਅਤੇ ਸਿੱਖਿਆ ਵਿਚ ਹੋ ਰਹੀ ਹੈ।

ਪ੍ਰਸ਼ਨ (Question)

ਕ੍ਰਿਤ੍ਰਿਮ ਬੁੱਧੀ ਦੇ ਮੁੱਖ ਹਿੱਸੇ ਕਿਹੜੇ ਹਨ?

Clear | Submit

ਉੱਤਰ & ਵਿਸ਼ਵਾਸ ਪੱਧਰ

ਉੱਤਰ: ਮਸ਼ੀਨ ਲਰਨਿੰਗ, ਡੀਪ ਲਰਨਿੰਗ, ਨੈਚਰਲ ਲੈਂਗਵਿਜ ਪ੍ਰੋਸੈਸਿੰਗ ਅਤੇ ਕੰਪਿਊਟਰ ਵਿਜ਼ਨ
ਵਿਸ਼ਵਾਸ-ਪੱਧਰ: 0.80

ਕ੍ਰਿਤ੍ਰਿਮ ਬੁੱਧੀ ਦੇ ਮੁੱਖ ਹਿੱਸੇ ਕਿਹੜੇ ਹਨ? ਕ੍ਰਿਤ੍ਰਿਮ ਬੁੱਧੀ ਜਾਂ ਆਰਟੀਫੀਸ਼ਲ ਇੰਟੈਲੀਜੈਂਸ (AI) ਇੱਕ ਤਕਨਾਲੋਜੀ ਹੈ ਜੋ ਕੰਪਿਊਟਰਾਂ ਨੂੰ ਮਨੁੱਖੀ ਬੁੱਧੀ ਵਰਗਾ ਸੋਚਣ ਅਤੇ ਫੈਸਲੇ ਲੈਣ ਯੋਗ ਬਣਾਉਂਦੀ ਹੈ। ਇਸ ਵਿਚ **ਮਸ਼ੀਨ ਲਰਨਿੰਗ, ਡੀਪ ਲਰਨਿੰਗ, ਨੈਚਰਲ ਲੈਂਗਵਿਜ ਪ੍ਰੋਸੈਸਿੰਗ ਅਤੇ ਕੰਪਿਊਟਰ ਵਿਜ਼ਨ** ਵਰਗੀਆਂ ਵਿਧੀਆਂ ਸ਼ਾਮਲ ਹਨ। AI ਦਾ ਉਦੇਸ਼ ਐਸੀਆਂ ਮਸ਼ੀਨਾਂ ਤਿਆਰ ਕਰਨਾ ਹੈ ਜੋ ਸਵੈ-ਸੀਖਣ, ਸਵੈ-ਨਿਰਣੈ ਅਤੇ ਮਨੁੱਖੀ ਤਰ੍ਹਾਂ ਕੰਮ ਕਰਨ ਵਿੱਚ ਸਮਰਥ ਹੋਣ। AI ਦੀ ਵਰਤੋਂ ਅੱਜਕੱਲ੍ਹ ਹੈਲਥਕੇਅਰ, ਆਟੋਮੋਬਾਈਲ, ਫਾਈਨੈਂਸ, ਖੇਤੀਬਾੜੀ ਅਤੇ ਸਿੱਖਿਆ ਵਿਚ ਹੋ ਰਹੀ ਹੈ।

Flag

CODE

```
!pip install -q transformers gradio torch

from transformers import AutoTokenizer, AutoModelForQuestionAnswering

import torch

model_name = "deepset/xlm-roberta-large-squad2"

tokenizer = AutoTokenizer.from_pretrained(model_name)

model = AutoModelForQuestionAnswering.from_pretrained(model_name)


def answer_question(context, question):

    inputs = tokenizer(question, context, return_tensors="pt", truncation=True)


    with torch.no_grad():

        outputs = model(**inputs)


    start_logits = outputs.start_logits

    end_logits = outputs.end_logits


    answer_start = torch.argmax(start_logits)

    answer_end = torch.argmax(end_logits) + 1


    answer_ids = inputs["input_ids"][0][answer_start:answer_end]

    answer = tokenizer.decode(answer_ids, skip_special_tokens=True,
clean_up_tokenization_spaces=True)
```

```
    start_prob = torch.softmax(start_logits, dim=1)[0][answer_start]
    end_prob = torch.softmax(end_logits, dim=1)[0][answer_end - 1]
    confidence = (start_prob * end_prob).item()

    tokens = tokenizer.convert_ids_to_tokens(inputs["input_ids"][0])
    decoded_context = tokenizer.decode(inputs["input_ids"][0], skip_special_tokens=True)

    if answer in decoded_context:
        highlighted_context = decoded_context.replace(answer,
f"<mark><b>{answer}</b></mark>")
    else:
        highlighted_context = decoded_context

    return f" ਉੱਤਰ: {answer}\n ਵਿਸ਼ਵਾਸ-ਪੱਧਰ: {confidence:.2f}", highlighted_context

import gradio as gr

interface = gr.Interface(
    fn=answer_question,
    inputs=[
        gr.Textbox(label="ਪੰਜਾਬੀ ਪੈਰਾ (Context)", lines=7, placeholder="ਇੱਥੇ ਪੈਰਾ ਦਿਓ..."),
        gr.Textbox(label="ਪ੍ਰਸ਼ਨ (Question)", placeholder="ਇੱਥੇ ਪ੍ਰਸ਼ਨ ਲਿਖੋ...")
    ],
```

```
    outputs=[
        gr.Textbox(label="ਉੱਤਰ & ਵਿਸ਼ਵਾਸ ਪੱਧਰ"),
        gr.HTML(label="ਪੈਰਾ (ਉੱਤਰ ਹਾਈਲਾਈਟ ਹੋਇਆ)")
    ],
    title="ਪੰਜਾਬੀ ਪ੍ਰਸ਼ਨ ਉੱਤਰਣ ਸਿਸਟਮ",
    description="ਪੰਜਾਬੀ ਵਿੱਚ ਸਵਾਲ ਪੁੱਛੋ ਅਤੇ ਉੱਤਰ ਪ੍ਰਾਪਤ ਕਰੋ!"
)

interface.launch(debug=True)
```