CONVERSATIONAL AI: NATURAL LANGUAGE PROCESSING(UCS664)

IndicText Summarizer

NLP project

Submitted by :
Hardik Kundal 102216071

Computer Science and Engineering Department
Thapar Institute of Engineering and Technology, Patiala

Introduction

With the surge in digital news content, particularly in regional languages like Hindi, there is a growing need for tools that can distill lengthy articles into concise summaries. Manual summarization is impractical for the vast volume of news generated daily, especially for Hindi-speaking audiences in India. The IndicText Summarizer project is motivated by the goal of automating news summarization, making information more accessible and time-efficient. The model was specifically trained on **12 Indian languages** (Hindi, Bengali, Gujarati, Kannada, Malayalam, Marathi, Oriya (Odia), Punjabi, Tamil, Telugu, Assamese, Urdu.).By fine-tuning the IndicBARTSS model on a Hindi news dataset, the project delivers more accurate summaries, enhancing usability for media, education, and general readers.

Dataset

The hindi_news_dataset.csv dataset, comprising 185,512 Hindi news articles, is used for training and evaluation. Likely sourced from Indian news outlets from kaggle (Inshorts Dataset - Hindi News), (https://www.kaggle.com/datasets/shivamtaneja2304/inshorts-dataset-hindi)
it includes:

- Headline: Reference summary.
- Content: Full article text.
- News Categories: E.g., national, politics.
- Date: Publication date (e.g., 29-06-2023).

The dataset is loaded and inspected using pandas,After cleaning (removing empty rows and normalizing text), the dataset retains 185,512 valid rows. To manage computational resources, a subset of 5,000 samples is used, split into 80% train (4,000 samples) and 20% test (1,000 samples) for fine-tuning.

Results

The IndicText Summarizer project utilized the IndicBARTSS model, a multilingual transformer-based model supporting 12 Indian languages. In this implementation, the model was fine-tuned specifically on **Hindi** summarization data.

The model was trained over **3 epochs**, and both training and validation loss steadily decreased, indicating effective learning and generalization.

Epoch	Training Loss	Validation Loss
1	7.302900	4.266286
2	2.819500	1.063501
3	1.041400	0.637006

The **training loss** dropped from 7.30 to 1.04.
The **validation loss** dropped from 4.26 to 0.63.
The model did not show signs of overfitting within these epochs.
The decreasing validation loss shows the model's effectiveness in generalizing to unseen Hindi text for summarization tasks.

CODE

Hugging Face Authentication

```
from huggingface_hub import login

# Replace with your Hugging Face token
hf_token = "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"  # Paste your token
here
login(token=hf_token)

from huggingface_hub import login
from google.colab import userdata

# Authenticate with Hugging Face
try:
    login(token=userdata.get('HF_TOKEN'))
    print("Authentication successful!")
except Exception as e:
    print("Authentication failed:", e)
```

Preprocessing the data

```
import pandas as pd
from indicnlp.tokenize import indic_tokenize
from datasets import Dataset, DatasetDict
from transformers import AutoTokenizer

# Load dataset
df = pd.read_csv('/content/hindi_news_dataset.csv',
encoding='utf-8',nrows=129934) #add if in dataset error ,nrows=129934

# Rename columns
df = df.rename(columns={'Content': 'article', 'Headline': 'summary'})
# Clean text function
def clean_text(text):
    if isinstance(text, str):
        text = text.strip().replace('\n', ' ').replace('\r',
'').replace('  ', ' ')
        return text
    return ""
```

```
# Apply cleaning
df['article'] = df['article'].apply(clean_text)
df['summary'] = df['summary'].apply(clean_text)



# Remove empty rows
df = df[df['article'].str.len() > 0]
df = df[df['summary'].str.len() > 0]

# Verify cleaned data
print("Cleaned Dataset Size:", len(df))
print("Sample Article:", df['article'].iloc[0])
print("Sample Summary:", df['summary'].iloc[0])

# Use a subset to manage memory
df = df.sample(5000, random_state=42)

# Split into train and test (80% train, 20% test)
train_size = int(0.8 * len(df))
train_df = df[:train_size]
test_df = df[train_size:]

# Convert to Hugging Face Dataset
train_dataset = Dataset.from_pandas(train_df[['article', 'summary']])
test_dataset = Dataset.from_pandas(test_df[['article', 'summary']])
dataset = DatasetDict({'train': train_dataset, 'test': test_dataset})

# Load tokenizer
model_name = "ai4bharat/IndicBARTSS"
tokenizer = AutoTokenizer.from_pretrained(model_name)

# Set the source and target language for mBART
tokenizer.src_lang = "hi_IN"  # Hindi
tokenizer.tgt_lang = "hi_IN"  # Hindi

# Tokenize dataset
def preprocess_function(examples):
    inputs = examples['article']
    targets = examples['summary']
    model_inputs = tokenizer(inputs, max_length=512, truncation=True,
padding="max_length")
    labels = tokenizer(targets, max_length=150, truncation=True,
padding="max_length")
    model_inputs["labels"] = labels["input_ids"]
    return model_inputs

tokenized_dataset = dataset.map(preprocess_function, batched=True)
```

Fine-Tune and train IndicBARTSS

```
from transformers import AutoModelForSeq2SeqLM, Trainer,
TrainingArguments
from google.colab import userdata

# Load model
model_name = "ai4bharat/IndicBARTSS"
hf_token = userdata.get('HF_TOKEN')  # Get token from Colab Secrets
model = AutoModelForSeq2SeqLM.from_pretrained(model_name,
token=hf_token if hf_token else None)

# Define training arguments
training_args = TrainingArguments(
    output_dir="/content/drive/My Drive/indicbartss_finetuned",
    eval_strategy="epoch",
    learning_rate=2e-5,
    per_device_train_batch_size=4,
    per_device_eval_batch_size=4,
    num_train_epochs=3,
    weight_decay=0.01,
    save_strategy="epoch",
    load_best_model_at_end=True,
    push_to_hub=False,
    gradient_accumulation_steps=2,
)

# Initialize Trainer
trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=tokenized_dataset["train"],
    eval_dataset=tokenized_dataset["test"],
)

# Start training
trainer.train()

# Save model to Google Drive
model.save_pretrained("/content/drive/My Drive/indicbartss_finetuned")
tokenizer.save_pretrained("/content/drive/My Drive/
indicbartss_finetuned")
```

Test the Model

```
from transformers import AutoTokenizer, AutoModelForSeq2SeqLM

# Load fine-tuned model and tokenizer
model = AutoModelForSeq2SeqLM.from_pretrained("/content/drive/My
Drive/indicbartss_finetuned")
tokenizer = AutoTokenizer.from_pretrained("/content/drive/My Drive/
indicbartss_finetuned")

# Define summarization function
def summarize_text(article):
    inputs = tokenizer(
        article,
        return_tensors="pt",
        max_length=512,
        truncation=True,
        padding="max_length"
    )
    summary_ids = model.generate(
        inputs["input_ids"],
        max_length=150,
        min_length=50,
        num_beams=4,
        early_stopping=True
    )
    summary = tokenizer.decode(summary_ids[0],
skip_special_tokens=True)
    return summary

# Test on samples
for i in range(2):
    article = test_df['article'].iloc[i]
    reference = test_df['summary'].iloc[i]
    print(f"\nSample {i+1}:")
    print("Article:", article)
    print("Reference Summary:", reference)
    print("Generated Summary:", summarize_text(article))
```

Creating gradio Interface

```
import gradio as gr
from transformers import AutoTokenizer, AutoModelForSeq2SeqLM

# Load model and tokenizer
model = AutoModelForSeq2SeqLM.from_pretrained("/content/drive/My
Drive/indicbartss_finetuned")
tokenizer = AutoTokenizer.from_pretrained("/content/drive/My Drive/
indicbartss_finetuned")

# Define summarization function
def summarize_text(article):
    inputs = tokenizer(
        article,
        return_tensors="pt",
        max_length=1024,
        truncation=True,
        padding="max_length"
    )
    summary_ids = model.generate(
        inputs["input_ids"],
        max_length=150,
        min_length=50,
        num_beams=4,
        early_stopping=True
    )
    summary = tokenizer.decode(summary_ids[0],
skip_special_tokens=True)
    return summary

# Create Gradio interface
interface = gr.Interface(
    fn=summarize_text,
    inputs=gr.Textbox(lines=10, placeholder="Paste news article
here..."),
    outputs="text",
    title="Indian News Summarization",
    description="Enter a news article to get a concise summary."
)

# Launch interface
interface.launch()
```

Some results

Enter a news article to get a concise summary.

ਜਲੰਧਰ (ਧਵਨ)–ਵਿਜੀਲੈਂਸ ਬਿਊਰੋ ਵੱਲੋਂ ਜਲੰਧਰ ਨਗਰ ਨਿਗਮ ਦੇ ਏ. ਟੀ. ਪੀ. ਸੁਖਦੇਵ ਵਸ਼ਿਸ਼ਟ ਦੀ ਗ੍ਰਿਫ਼ਤਾਰੀ ਤੋਂ ਬਾਅਦ ਹੁਣ ਜਾਂਚ ਦਾ ਘੇਰਾ ਸੱਤਾਧਾਰੀ ਆਮ ਆਦਮੀ ਪਾਰਟੀ ਦੇ ਕੁਝ ਆਗੂਆਂ ਵੱਲ ਘੁੰਮ ਸਕਦਾ ਹੈ। ਮੁੱਖ ਮੰਤਰੀ ਭਗਵੰਤ ਮਾਨ ਐਲਾਨ ਕਰ ਚੁੱਕੇ ਹਨ ਕਿ ਭ੍ਰਿਸ਼ਟਾਚਾਰ ਵਿਚ ਸ਼ਾਮਲ ਕਿਸੇ ਨੂੰ ਵੀ ਬਖ਼ਸ਼ਿਆ ਨਹੀਂ ਜਾਵੇਗਾ, ਭਾਵੇਂ ਉਹ ਸੱਤਾਧਾਰੀ ਪਾਰਟੀ ਨਾਲ ਹੀ ਸਬੰਧ ਕਿਉਂ ਨਾ ਰੱਖਦਾ ਹੋਵੇ।

ਮੁੱਖ ਮੰਤਰੀ ਨੇ ਬੀਤੇ ਦਿਨੀਂ ਲੁਧਿਆਣਾ ਵਿਚ ਵੀ ਇਕ ਸਮਾਰੋਹ ਦੌਰਾਨ ਭ੍ਰਿਸ਼ਟਾਚਾਰ ਨੂੰ ਲੈ ਕੇ ਸਖ਼ਤ ਸਟੈਂਡ ਲੈਣ ਦਾ ਸੰਕੇਤ ਦਿੱਤਾ ਸੀ। ਉਸ ਤੋਂ ਬਾਅਦ ਵਿਜੀਲੈਂਸ ਬਿਊਰੋ ਇਕਦਮ ਹਰਕਤ ਵਿਚ ਆ ਗਿਆ ਅਤੇ ਉਨ੍ਹਾਂ ਬੀਤੇ ਦਿਨੀਂ ਹੀ ਏ. ਟੀ. ਪੀ. ਸੁਖਦੇਵ ਵਸ਼ਿਸ਼ਟ ਨੂੰ ਭ੍ਰਿਸ਼ਟਾਚਾਰ ਦੇ ਦੋਸ਼ਾਂ ਵਿਚ ਆਪਣੀ ਹਿਰਾਸਤ ਵਿਚ ਲੈ ਲਿਆ ਸੀ। ਵੀਰਵਾਰ ਉਸ ਦੀ ਗ੍ਰਿਫ਼ਤਾਰੀ ਪਾਈ ਗਈ। ਵਿਜੀਲੈਂਸ ਬਿਊਰੋ ਦੇ ਅਧਿਕਾਰੀਆਂ ਦਾ ਮੰਨਣਾ ਹੈ ਕਿ ਜਾਂਚ ਦੌਰਾਨ ਵਿਜੀਲੈਂਸ ਅਧਿਕਾਰੀਆਂ ਵੱਲੋਂ ਇਹ ਵੇਖਿਆ ਜਾਵੇਗਾ ਕਿ ਏ. ਟੀ. ਪੀ. ਸੁਖਦੇਵ ਵਸ਼ਿਸ਼ਟ ਨੇ ਆਪਣੇ ਕਾਰਜਕਾਲ ਦੌਰਾਨ ਸ਼ਹਿਰ ਵਿਚ ਕਿਹੜੀਆਂ-ਕਿਹੜੀਆਂ ਇਮਾਰਤਾਂ ਦੇ ਨਕਸ਼ਿਆਂ ਨੂੰ ਮਨਜ਼ੂਰੀ ਦਿੱਤੀ ਸੀ ਅਤੇ ਉਸ ਦੇ ਪਿੱਛੇ ਕੀ ਗੋਲਮਾਲ ਹੋਇਆ ਹੈ।

ਇਹ ਵੀ ਪੜ੍ਹੋ: ਖੇਤਾਂ 'ਚ ਲੱਗੀ ਅੱਗ ਨੇ ਤਬਾਹ ਕਰ ਦਿੱਤੇ ਗ਼ਰੀਬਾਂ ਦੇ ਆਸ਼ਿਆਨੇ, ਸਾਮਾਨ ਸੜ ਕੇ ਹੋਇਆ ਸੁਆਹ

ਦੱਸਿਆ ਜਾਂਦਾ ਹੈ ਕਿ ਸੁਖਦੇਵ ਇਕ 'ਆਪ' ਆਗੂ ਦੇ ਕਹਿਣ ’ਤੇ ਫਾਈਲਾਂ ਨੂੰ ਕਲੀਅਰ ਕਰਦਾ ਸੀ। ਇਸ ਆਗੂ ਦੇ ਕਹਿਣ ’ਤੇ ਹੀ ਇਮਾਰਤਾਂ ਨੂੰ ਸੀਲਾਂ ਲਾਈਆਂ ਜਾਂਦੀਆਂ ਸਨ ਅਤੇ ਇਮਾਰਤਾਂ ਦੇ ਨਕਸ਼ਿਆਂ ਨੂੰ ਪਾਸ ਕੀਤਾ ਜਾਂਦਾ ਸੀ। ਉਕਤ ਆਗੂ ਦੀ ਬਿਲਡਿੰਗ ਬਰਾਂਚ ਵਿਚ ਇੰਨੀ ਜ਼ਿਆਦਾ ਦਖ਼ਲਅੰਦਾਜ਼ੀ ਸੀ ਕਿ ਉਸ ਦੇ ਕਹੇ ਬਿਨਾਂ ਕੋਈ ਵੀ ਕੰਮ ਨਹੀਂ ਹੁੰਦਾ ਸੀ।

ਵਿਜੀਲੈਂਸ ਅਧਿਕਾਰੀਆਂ ਨੂੰ ਇਸ ਦੀ ਭਿਣਕ ਲੱਗ ਚੁੱਕੀ ਹੈ ਅਤੇ ਉਹ ਜਾਂਚ ਕਰ ਰਹੇ ਹਨ ਕਿ ਉਕਤ ਆਗੂ ਅਤੇ ਇਸ ਅਧਿਕਾਰੀ ਨੇ ਮਿਲ ਕੇ ਪਿਛਲੇ ਸਮੇਂ ਵਿਚ ਕਿੰਨਾ ਪੈਸਾ ਲੋਕਾਂ ਤੋਂ ਭ੍ਰਿਸ਼ਟਾਚਾਰ ਜ਼ਰੀਏ ਲਿਆ ਹੈ।

ਇਹ ਵੀ ਦੱਸਿਆ ਜਾ ਰਿਹਾ ਹੈ ਕਿ ਉਕਤ ਅਧਿਕਾਰੀ ਨਗਰ ਨਿਗਮ ਵਿਚ ਕੰਮ ਕਰਦੇ ਹੋਏ ਆਪਣੇ ਸਹਿਯੋਗੀ ਅਧਿਕਾਰੀਆਂ ਦੀ ਗੱਲ ਵੀ ਨਹੀਂ ਸੁਣਦਾ ਸੀ ਕਿਉਂਕਿ ਉਸ ਨੂੰ ਉਕਤ 'ਆਪ' ਆਗੂ ਦਾ ਆਸ਼ੀਰਵਾਦ ਹਾਸਲ ਸੀ। ਸ਼ਹਿਰ ਵਿਚ ਪ੍ਰਾਪਰਟੀ ਨਾਲ ਸਬੰਧਤ ਜਿੰਨੇ ਵੀ ਸੌਦੇ ਹੋਏ ਹਨ, ਉਨ੍ਹਾਂ ਵਿਚ ਇਨ੍ਹਾਂ ਦੋਵਾਂ ਦੀ ਸ਼ਮੂਲੀਅਤ ਦੱਸੀ ਜਾ ਰਹੀ ਹੈ। ਉਂਝ ਆਮ ਆਦਮੀ ਪਾਰਟੀ ਦੇ ਵਧੇਰੇ ਆਗੂ ਖ਼ੁਸ਼ ਹਨ ਕਿਉਂਕਿ ਹੁਣ ਨਗਰ ਨਿਗਮ ਵਿਚ ਮੈਰਿਟ ਦੇ ਆਧਾਰ ’ਤੇ ਕੰਮ ਹੋਵੇਗਾ। ਉਕਤ ਆਗੂ ਦੀ ਦਖ਼ਲਅੰਦਾਜ਼ੀ ਹੁਣ ਬੰਦ ਹੋ ਜਾਵੇਗੀ ਅਤੇ ਅਧਿਕਾਰੀ ਵੀ ਹੁਣ ਬਿਨਾਂ ਕਿਸੇ

ਜਲੰਧਰ ਨਗਰ ਨਿਗਮ ਦੇ ਏ. ਟੀ. ਪੀ. ਸੁਖਦੇਵ ਵਸ਼ਿਸ਼ਟ ਦੀ ਗ੍ਰਿਫ਼ਤਾਰੀ ਤੋਂ ਬਾਅਦ ਹੁਣ ਜਾਂਚ ਦਾ ਘੇਰਾ ਸੱਤਾਧਾਰੀ ਆਮ ਆਦਮੀ ਪਾਰਟੀ ਦੇ ਕੁਝ ਆਗੂਆਂ ਵੱਲ ਘੁੰਮ ਸਕਦਾ ਹੈ। ਮੁੱਖ ਮੰਤਰੀ ਭਗਵੰਤ ਮਾਨ ਐਲਾਨ ਕਰ ਚੁੱਕੇ ਹਨ ਕਿ ਭ੍ਰਿਸ਼ਟਾਚਾਰ ਵਿਚ ਸ਼ਾਮਲ ਕਿਸੇ ਨੂੰ ਵੀ ਬਖ਼ਸ਼ਿਆ ਨਹੀਂ ਜਾਵੇਗਾ, ਭਾਵੇਂ ਉਹ ਸੱਤਾਧਾਰੀ ਪਾਰਟੀ ਨਾਲ ਹੀ ਸਬੰਧ ਕਿਉਂ ਨਾ ਰੱਖਦਾ ਹੋਵੇ। ਮੁੱਖ ਮੰਤਰੀ ਨੇ ਬੀਤੇ ਦਿਨੀਂ ਲੁਧਿਆਣਾ ਵਿਚ ਵੀ ਇਕ ਸਮਾਰੋਹ ਦੌਰਾਨ ਭ੍ਰਿਸ਼ਟਾਚਾਰ ਨੂੰ ਲੈ ਕੇ ਸਖ਼ਤ ਸਟੈਂਡ ਲੈਣ ਦਾ ਸੰਕੇਤ ਦਿੱਤਾ ਸੀ। ਉਸ ਤੋਂ ਬਾਅਦ ਵਿਜੀਲੈਂਸ ਬਿਊਰੋ ਇਕਦਮ ਹਰਕਤ ਵਿਚ ਆ ਗਿਆ ਅਤੇ ਉਨ੍ਹਾਂ ਬੀਤੇ ਦਿਨੀਂ ਹੀ ਏ. ਟੀ. ਪੀ. ਸੁਖਦੇਵ ਵਸ਼ਿਸ਼ਟ ਨੂੰ ਦੋਸ਼ੀ ਕਰਾਰ ਦਿੱਤਾ ਸੀ। ਉਨ੍ਹਾਂ ਦੱਸਿਆ ਕਿ ਉਹ ਸੱਤਾਧਾਰੀ ਪਾਰਟੀ ਨਾਲ ਹੀ ਸਬੰਧ ਕਿਉਂ ਨਾ ਰੱਖਦਾ ਹੋਵੇ। ਉਸ ਦੇ ਕਹਿਣ 'ਤੇ ਉਸ ਦੇ

Enter a news article to get a concise summary.

மே 7 ஆம் தேதி காலை இந்தியாவுக்கும் பாகிஸ்தானுக்கும் இடையில் மிகவும் பதற்றமான சூழல் நிலவியது. மே 6-ம் தேதி இரவு, இந்தியா பாகிஸ்தானுக்குள் ராணுவ நடவடிக்கையை மேற்கொண்டது, காலையில் இந்த செய்தி காட்டுத்தீ போல் பரவியது.

பாகிஸ்தானும் பதிலடி நடவடிக்கையைத் தொடங்கியது, அடுத்த நான்கு நாட்களுக்கு இரு தரப்பிலும் தாக்குதல்கள் தொடர்ந்தன.

இத்தகைய சூழ்நிலையில், உலகின் சக்திவாய்ந்த நாடுகள் எவ்வாறு எதிர்வினையாற்றின என்பதைப் பார்ப்பது சுவாரஸ்யமாக இருந்தது. பாகிஸ்தானின் இறையாண்மையை பாதுகாக்க உறுதி பூண்டுள்ளதாக சீனா வெளிப்படையாக கூறியுள்ளது. துருக்கியும் பாகிஸ்தானுடன் இருந்தது.

மறுபுறம், இந்தியாவுக்காக எந்த நாடும் அவ்வாறு கூறவில்லை.

இந்தியாவுக்கு தற்காப்பு உரிமை உண்டு என்று இஸ்ரேல் கூறியது. ஆனால் இஸ்ரேலுடன் ஒப்பிடுகையில், சீனா மிகப் பெரிய நாடு மற்றும் உலகின் இரண்டாவது பெரிய பொருளாதாரத்தைக் கொண்டுள்ளது.

பாகிஸ்தானுக்குள் ராணுவ நடவடிக்கையை மேற்கொண்டது இந்தியா; பாகிஸ்தானுக்கு எதிராக இந்தியா பதிலடி நடவடிக்கையை மேற்கொண்டது; இந்தியாவிற்கும் பாகிஸ்தானுக்கும் இடையில் கடும் பதற்றமான சூழல் நிலவியது. காலையில் பாகிஸ்தானுடன் ஒப்பிடுகையில் இந்தியா பாகிஸ்தானுக்குள் ராணுவ நடவடிக்கையை மேற்கொண்டது; காலையில் பாகிஸ்தானுடன் ஒப்பிடுகையில் இந்தியாவிற்கும் பாகிஸ்தானுடன் ஒப்பிடுகையில் இந்தியாவிற்கும் பாகிஸ்தானுடன் ஒப்பிடுகையில் இந்தியாவிற்கும் பாகிஸ்தானுடன் ஒப்பிடுகையில் இந்தியாவிற்கும் பாகிஸ்தானுடன் ஒப்பிடுகையில் இந்தியாவிற்கும் பாகிஸ்தானுடன் ஒப்பிடுகையில் இந்தியாவிற்கும் பாகிஸ்தானுடன் ஒப்பிடுகையில் இந்தியாவிற்கும் பாகிஸ்தானுடன் ஒப்பிடுகையில் இந்தியாவிற்கும் பாகிஸ்தானுடன் ஒப்பிடுகையில் இந்தியாவிற்கும் பாகிஸ்தானுடன் ஒப்பிடுகையில் இந்தியாவிற்கும் பாகிஸ்தானுடன் ஒப்பிடுகையில் இந்தியாவிற்கும் பாகிஸ்தானுடன் ஒப்பிடுகையில் இந்தியாவிற்கும் பாகிஸ்தானுடன் ஒப்பிடுகையில் இந்தியாவிற்கும் பாகிஸ்தான