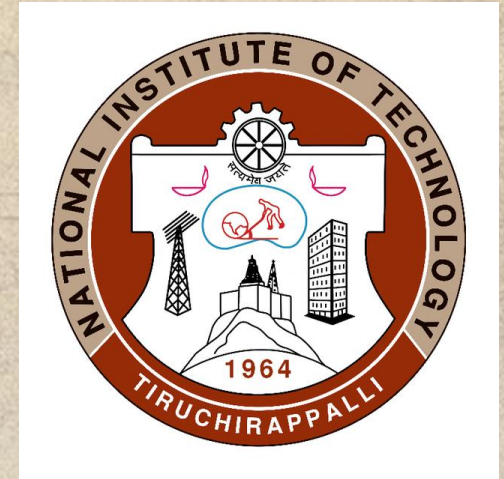


# Automatic Hindi News Headline Generation Using Transformer Based Language Models

---



PRESENTED BY- GAURAV THAKUR

GUIDED BY- DR.VISHNU PRIYA R



# CONTENTS

---



- PROBLEM STATEMENT
- OBJECTIVE
- INTRODUCTION
- DATASET
- DATASET PREPROCESSING
- MODELS
- MODEL TRAINING
- MODEL EVALUATION
- RESULT
- USER INTERFACE
- CONCLUSION
- FUTURE WORK



# PROBLEM STATEMENT

---

- The size of electronic data from several sources such as publications, government data, e-news, and social networks is growing extensively across the web. With so much information spread across various platforms, it becomes challenging to maintain a cohesive understanding of a topic, leading to fragmented knowledge.
- Increased data generation and collection raise concerns about data privacy and security, making it challenging to protect sensitive information.
- The availability of summarization tools is mainly limited to English, hindering the accessibility of NLP applications for other languages such as Hindi, Punjabi etc. Existing summarization techniques often struggle with Hindi documents due to:
  - **Unique linguistic features**
  - **Scarcity of resources**
  - **Limited performance**



# OBJECTIVE

---

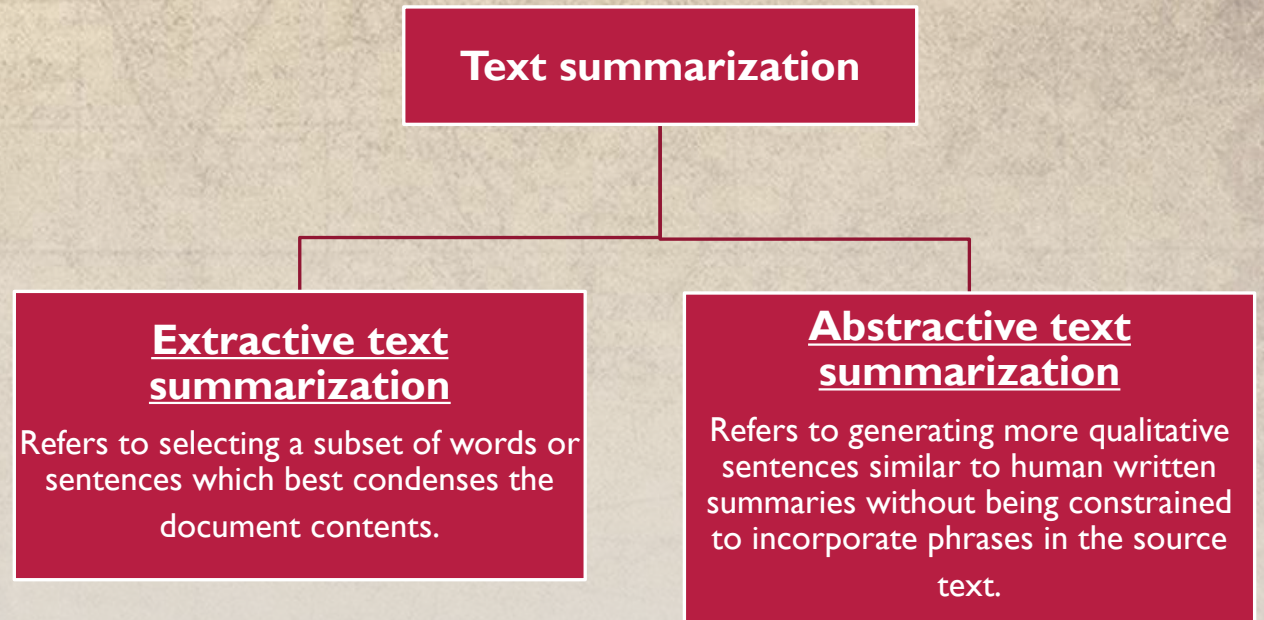
- Develop a robust summarization system tailored for the Hindi language.
- For text summarization, fine-tune and assess how well the pre-trained Transformer based language models perform on the Hindi dataset.
- Evaluate the performance of the summarization system using metrics such as Rogue score and Bleu score to assess its effectiveness in handling Hindi inputs and providing satisfactory summaries.
- Contributing to the development of NLP tools for Hindi: The initiative may aid in the creation of NLP instruments that are applicable to a range of Hindi-language applications.



# INTRODUCTION

---

- Text Summarization is the task of generating a concise summary that captures the principal ideas of the source text. In which news headline generation is a subset of abstractive text summarization.
- Although abstractive text summarization is researched widely for languages like English, French and Chinese. The Hindi language is very different from English or French language due to having tough grammar and matras in letters.
- Furthermore, recent techniques for Hindi corpus summarization predominantly include extractive methods.





# DATASET

---

- **Hindi Text Short Summarization Corpus**

- Hindi Text Short Summarization Corpus is a collection of ~330k articles with their headlines collected from Hindi News Websites.
- It is collected from Kaggle
  - <https://www.kaggle.com/datasets/disisbig/hindi-text-short-summarization-corpus>
- The training corpus consists of approx. 266k and test corpus consists of approx. 66k Hindi news articles along with their headlines(summary).

```
[5] train.shape
```

```
(266607, 2)
```

✓  
9s

```
test = pd.read_csv(default_path+"test.csv")  
test.shape
```

```
(66653, 2)
```



# SNAPSHOT OF DATASET

train = pd.read\_csv(default\_path+"train.csv")  
train.head()

1 to 5 of 5 entries 

Filter

| index                  | headline   | article  |
|------------------------|--|--|
| 0                      | EXCLUSIVE: दिल्ली में डीज़ल टैंक्सियों पर बैन से मुश्किल में पड़ा चुनाव आयोग | दिल्ली में सुप्रीम कोर्ट के डीज़ल टैंक्सियों को बंद करने के फैसले के बाद हजारों टैंकसी ड्राइवरों की रोज़ी रोटी पर तो असर पड़ा ही है, लेकिन अब दिल्ली पर एक और नई मुसीबत आ गई है. चुनाव आयोग राजधानी के 13 वार्ड में उपचुनाव करवा रहा है, लेकिन चुनावों से दो हफ्ते पहले चुनाव आयोग में कामकाज ठप्प हो गया है. कमीशन ने किराए पर ली थी डीज़ल गाड़ियां दरअसल कमीशन ने लगभग सौ गाड़ियां चुनाव के कामकाज को करने के लिए किराए पर लीं, जिनमें सभी डीज़ल से चलने वाली टैंकसी थीं. इन्हीं टैंक्सियों से चुनाव अधिकारी से लेकर चुनावों का जिम्मा संभालने वाले बाकी कर्मचारी भी एक जगह से दूसरी जगह आते जाते थे. अचानक चुनावों से ठीक पहले आई इस परेशानी ने दिल्ली चुनाव आयोग का कामकाज ही ठप्प कर दिया है. रियायत के लिए की जा सकती है मांग दिल्ली के राज्य चुनाव अधिकारी राकेश मेहता ने इस मुश्किल का रास्ता निकालने के लिए मंगलवार को दिल्ली के पुलिस कमिश्नर और ट्रांसपोर्ट कमिश्नर की बैठक बुलाई है. इस बैठक में राज्य चुनाव आयुक्त 15 मई को होने वाले चुनावों को लेकर गाड़ियों की उपलब्धता को लेकर पुलिस और सरकार से समाधान निकालने के लिए भी कहेंगे. चुनाव आयोग इन दोनों एजेंसियों को कह सकता है कि चूंकि चुनाव में अब दो हफ्ते का भी वक्त नहीं बचा है, ऐसे में इन गाड़ियों को बैन से रियायत दी जाए.   |
| 1                      | जॉर्डन: राष्ट्रपति मुखर्जी ने 86 करोड़ डॉलर के संयंत्र का उद्घाटन किया       | जॉर्डन के ऐतिहासिक दौरे पर पहुंचे राष्ट्रपति प्रणब मुखर्जी ने 86 करोड़ डॉलर की लागत से निर्मित भारत-जॉर्डन उर्वरक संयंत्र का जॉर्डन के शाह अब्दुल्ला (द्वितीय) इब्र अल हुसैन के साथ उद्घाटन किया. यह संयंत्र एक साल से कम समय में बनकर तैयार हुआ है. राष्ट्रपति मुखर्जी के यहां एयर इंडिया की उड़ान से दोपहर पहुंचने के कुछ देर बाद ही शाह के महल से इस संयंत्र का रिमोट के जरिए उद्घाटन किया गया. अधिकारियों ने बताया कि भारतीय उर्वरक कंपनी इफको और जॉर्डन के फास्फेट्स माइन कंपनी ने इस संयंत्र के लिए 2008 में एक संयुक्त उद्यम कंपनी जॉर्डन इंडिया फर्टिलाइजर कंपनी बनाया. संयुक्त उद्यम में इफको की हिस्सेदारी 52 प्रतिशत है. इस संयंत्र से प्रति वर्ष 450 करोड़ टन सल्फ्यूरिक एसिड और 150 करोड़ टन फास्फेरिक एसिड के उत्पादन का अनुमान है. राष्ट्रपति का इससे पहले यहां पारंपरिक स्वागत किया गया और राष्ट्रपति भवन के सामने उन्हें 21 तोपों की सलामी दी गई. इसके बाद वह शाह अब्दुल्ला (द्वितीय) इब्र अल हुसैन के साथ वार्ता में व्यस्त हो गए. वार्ता के बाद दोनों नेताओं ने इंडो-जर्मन उर्वरक संयंत्र का संयुक्त रूप से उद्घाटन किया. इस संयंत्र से कच्चे माल का उत्पादन किया जाएगा. इसमें फास्फोरिक एसिड और सल्फ्यूरिक एसिड प्रमुख हैं. यहां पहुंचने से पहले राष्ट्रपति ने कहा, 'दोनों देशों के क्षेत्रीय एवं अंतर्राष्ट्रीय मुद्दे मिलते-जुलते हैं और दोनों सीरिया के साथ ही मध्य पूर्व में शांति प्रक्रिया का समर्थन करते हैं.' उन्होंने कहा कि दोनों देश उग्रवाद और आतंकवाद के सभी रूपों की निंदा करते हैं और धार्मिक सौहार्द्र में भरोसा करते हैं. राष्ट्रपति के इस दौरे के दौरान व्यापार एवं निवेश पर भी जोर है. उन्होंने कहा कि दोनों देश द्विपक्षीय व्यापार को पांच अरब डॉलर करना चाहते हैं. अभी दोनों देशों के बीच व्यापार दो अरब डॉलर है. प्रणब मुखर्जी ने जिस संयंत्र का उद्घाटन किया, उससे भारत 30 करोड़ टन फास्फोरिक एसिड का आयात करेगी. भारत बड़ी मात्रा में पोटैश एवं फास्फेट जॉर्डन से हासिल करता है. भारत और जॉर्डन ने 1947 में सामंजस्य के लिए द्विपक्षीय समझौते पर हस्ताक्षर किया था, हालांकि इसे औपचारिक रूप 1950 में दिया गया जब पूर्ण कूटनीतिक संबंध दोनों देश के बीच बने. शाह अब्दुल्ला और बेगम रानिया ने अक्टूबर 2012 में भारत का दौरा किया था. राष्ट्रपति के इस दौरे से पूर्व करीब 30 साल पहले तत्कालीन प्रधानमंत्री राजीव गांधी ने इस देश का दौरा किया था. इनपुट-IANS |
| 2                      | UN में पाकिस्तान की राजदूत मलीहा लोधी ने कराई फजीहत, मांगनी पड़ी माफी        | पाकिस्तानी नेताओं को विवादित और हास्यास्पद बयान आए दिन सुर्खियां बंटोते रहते हैं और कई बार तो पाकिस्तान के मंत्री तक भड़काऊ बयान देने से बाज नहीं आते. इस कड़ी में संयुक्त राष्ट्र में पाकिस्तान की प्रतिनिधि का नाम भी जुड़ गया है जिन्होंने ब्रिटेन के प्रधानमंत्री बोरिस जॉनसन को ब्रिटिश विदेश मंत्री बता दिया. मलीहा लोधी ने एक फोटो ट्वीट की जिसमें पाकिस्तान प्रधानमंत्री इमरान खान और ब्रिटिश प्रधानमंत्री बोरिस जॉनसन की मुलाकात हो रही है. इस फोटो का कैप्शन लिखते वक्त लोधी से चूक हो गई और वह जॉनसन को ब्रिटेन का विदेश मंत्री लिख बैठीं. इसके कुछ देर बाद ही उन्हें अपनी गलती का अहसास हुआ और उन्होंने अपना ट्वीट डिलीट कर दिया. मलीहा लोधी का ट्वीट पाकिस्तान की राजदूत ने फिर से ट्वीट करते हुए अपने पुराने ट्वीट पर माफी मांगी और कहा कि टाइप करने में गलती हो गई. प्रधानमंत्री इमरान ने ब्रिटिश पीएम से मुलाकात की है. Sorry typo in previous tweet. Prime Minister Imran Khan Met British PM this morning. pic.twitter.com/Ufp9vz5Ent — Maleeha Lodhi (@LodhiMaleeha) September 23, 2019 बता दें कि पाकिस्तान के प्रधानमंत्री इमरान खान संयुक्त राष्ट्र महासभा के 74वें सत्र में हिस्सा लेने अमेरिका गए हुए हैं. यहां इमरान खान ने कई नेताओं से मुलाकात की और कश्मीर को लेकर बातचीत है. पाकिस्तान के प्रधानमंत्री सोमवार रात को राष्ट्रपति डोनाल्ड ट्रंप से मुलाकात करेंगे जहां वह ट्रंप से कश्मीर के मसले पर मध्यस्थता करने की गुहार फिर से लगाने वाले हैं.   |
| 38 देशों में पीएम मोदी | 38 देशों में पीएम मोदी   | पीएम नरेंद्र मोदी बायोपिक में विवेक ओबेरॉय ने प्रधानमंत्री नरेंद्र मोदी का किरदार निभाया है. जानकारी के अनुसार फिल्म भारत के अलावा कई देशों में रिलीज हो सकती है. प्रोड्यूसर और डिस्ट्रीब्यूटर आनंद पंडित ने बताया कि पीएम नरेंद्र मोदी बायोपिक की टीम फिल्म को 38 देशों में रिलीज करने का प्लान बना रही है. इसमें यूएस, यूके, कनाडा, ऑस्ट्रेलिया और यूएई जैसे देश शामिल हैं. पिंकविला के अनुसार प्रोड्यूसर आनंद ने बताया कि पीएम नरेंद्र मोदी की लाइफ को लेकर देश ही नहीं विदेशों के ऑडियंस में ही गजब की रुचि है. इसलिए इसे इंडिया के अलावा 38 देशों में रिलीज करनी योजना है. उन्होंने कहा, भारत में यह फिल्म 1700 स्क्रीन पर रिलीज होगी. इसके अलावा इसे विदेशों में लगभग 600 स्क्रीन पर रिलीज करने की तैयारी की जा रही है. मोदी बायोपिक को बन करने की मांग पर आनंद पंडित ने कहा, "फिल्म को लेकर कोई शक नहीं होना चाहिए. सभी जानते हैं कि ये फिल्म क्या है. यह बायोपिक एक सिनेमेटिक प्रोडक्ट है. जो लोग इस फिल्म को लेकर सवाल उठा रहे हैं और इस पर बैन लगाने की मांग कर रहे हैं. दरअसल, वे लोग अभिव्यक्ति की स्वतंत्रता को दबाना चाहते हैं. क्या यह पाखंड नहीं है कि जो लोग दूसरी फिल्मों को बैन करने की आलोचना करते हैं वे ही आज मोदी बायोपिक को बैन करने की मांग कर रहे हैं.   |

✓

Connected to Python 3 Google Compute Engine backend (TPU)

✕



# DATASET PREPROCESSING

- **Dataset Cleaning**

- Removed emoticons, symbols & pictographs, Chinese characters, transport symbols, HTML tags, URLs.
- Removed English alphabetic characters.
- Removed punctuations and special characters from the sentences.

```
[ ] train_data_trg[1]
```

```
'जॉर्डन: राष्ट्रपति मुखर्जी ने 86 करोड़ डॉलर के संयंत्र का उद्घाटन किया'
```

```
▶ tokenized_corpus_trg[1]
```

```
🔗 'जॉर्डन राष्ट्रपति मुखर्जी ने 86 करोड़ डॉलर के संयंत्र का उद्घाटन किया'
```

```
[ ] train_data_src[1]
```

```
'जॉर्डन के ऐतिहासिक दौर पर पहुंचे राष्ट्रपति प्रणब मुखर्जी ने 86 करोड़ डॉलर की लागत से निर्मित भारत-जॉर्डन उर्वरक संयंत्र का जॉर्डन के शाह अब्दुल्ला (द्वितीय) इब्र अल हुसैन के साथ उद्घाटन किया. यह संयंत्र एक साल से कम समय में बनकर तैयार हुआ है.\n\nराष्ट्रपति मुखर्जी के यहां एयर इंडिया की उड़ान से दोपहर पहुंचने के कुछ देर बाद ही शाह के महल से इस संयंत्र का रिमोट के जरिए उद्घाटन किया गया. अधिकारियों ने बताया कि भारतीय उर्वरक कंपनी इफको और जॉर्डन के फास्फेट्स माइन कंपनी ने इस संयंत्र के लिए 2008 में एक संयुक्त उद्यम कंपनी जॉर्डन इंडिया फर्टिलाइजर कंपनी बनाया. संयुक्त उद्यम में इफको की हिस्सेदारी 52 प्रतिशत है.\n\nइस संयंत्र से प्रति वर्ष 450 करोड़ टन सल्फ्यूरिक एसिड और 150 करोड़ टन फास्फेरिक एसिड के उत्पादन का अनुमान है. राष्ट्रपति का इससे पहले यहां पारंपरिक स्वागत किया गया और राष्ट्रपति भवन के सामने उन्हें 21 तोपों की सलामी दी गई. इसके बाद वह शाह अब्दुल्ला (द्वितीय) इब्र अल हुसैन के साथ वार्ता में व्यस्त हो गए. वार्ता के बाद दोनों नेताओं ने इंडो-जर्मन उर्वरक संयंत्र का संयुक्त रूप से...
```

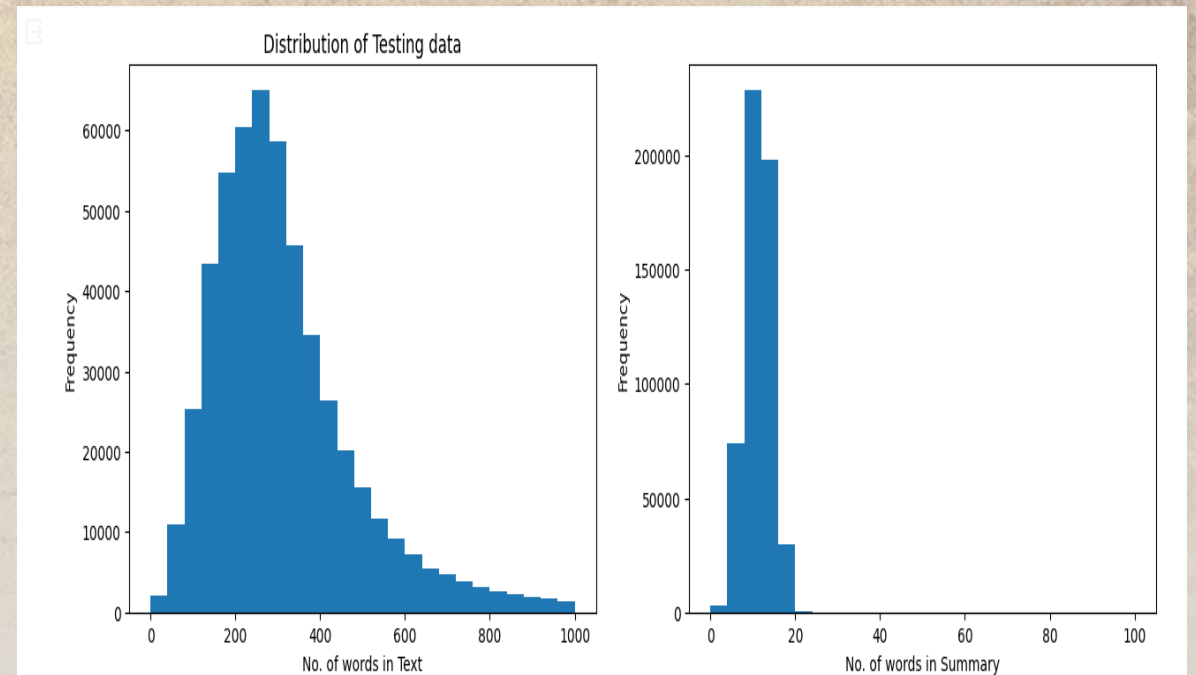
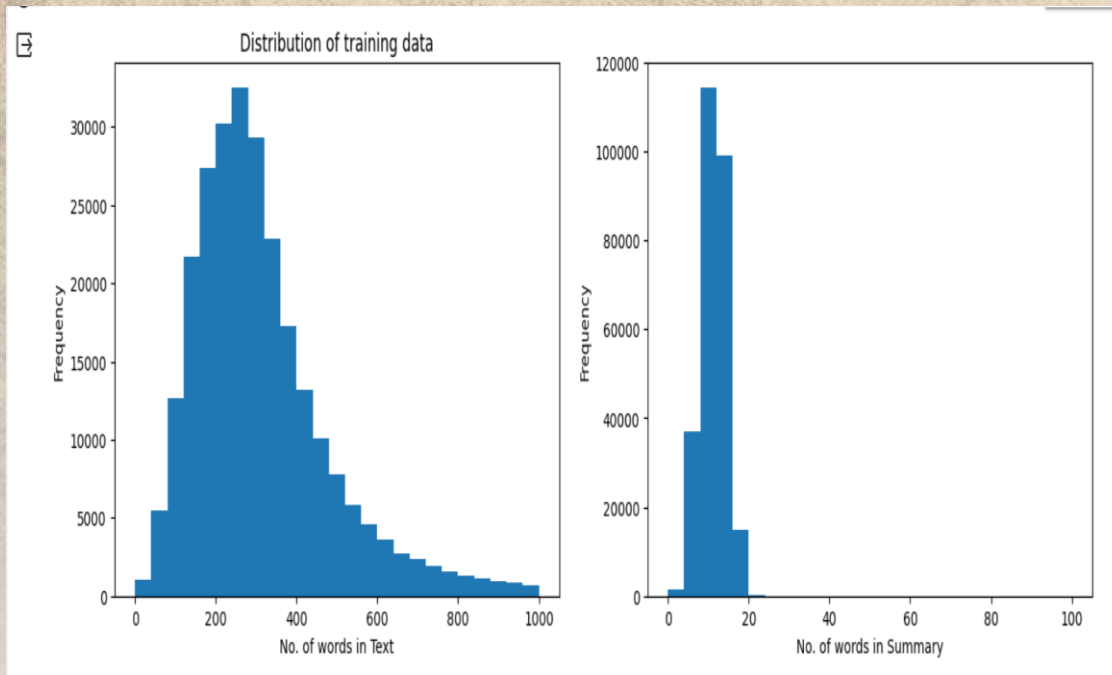
```
[ ] tokenized_corpus_src[1] #cleaned sentences
```

```
'जॉर्डन के ऐतिहासिक दौर पर पहुंचे राष्ट्रपति प्रणब मुखर्जी ने 86 करोड़ डॉलर की लागत से निर्मित भारत-जॉर्डन उर्वरक संयंत्र का जॉर्डन के शाह अब्दुल्ला द्वितीय इब्र अल हुसैन के साथ उद्घाटन किया यह संयंत्र एक साल से कम समय में बन कर तैयार हुआ है राष्ट्रपति मुखर्जी के यहां एयर इंडिया की उड़ान से दोपहर पहुंचने के कुछ देर बाद ही शाह के महल से इस संयंत्र का रिमोट के जरिए उद्घाटन किया गया अधिकारियों ने बताया कि भारतीय उर्वरक कंपनी इफको और जॉर्डन के फास्फेट्स माइन कंपनी ने इस संयंत्र के लिए 2008 में एक संयुक्त उद्यम कंपनी जॉर्डन इंडिया फर्टिलाइजर कंपनी बनाया संयुक्त उद्यम में इफको की हिस्सेदारी 52 प्रतिशत है इस संयंत्र से प्रति वर्ष 450 करोड़ टन सल्फ्यूरिक एसिड और 150 करोड़ टन फास्फेरिक एसिड के उत्पादन का अनुमान है राष्ट्रपति का इससे पहले यहां पारंपरिक स्वागत किया गया और राष्ट्रपति भवन के सामने उन्हें 21 तोपों की सलामी दी गई इसके बाद वह शाह अब्दुल्ला द्वितीय इब्र अल हुसैन के साथ वार्ता में व्यस्त हो गए वार्ता के बाद दोनों नेताओं ने इंडो-जर्मन उर्वरक संयंत्र का संयुक्त रूप से उद्घाटन किया इस संयंत्र...'
```



# CONTD..

- **Word Counts**





# CONTD..

---

- **Padding**

- Padding is done in order to make all sequences in a batch fit a given standard length.

- **Embedding**

- Text data has been converted to word embedding using the same embedding used in pre-trained models. This ensures that data will be compatible with the pre-trained model.



# MODELS

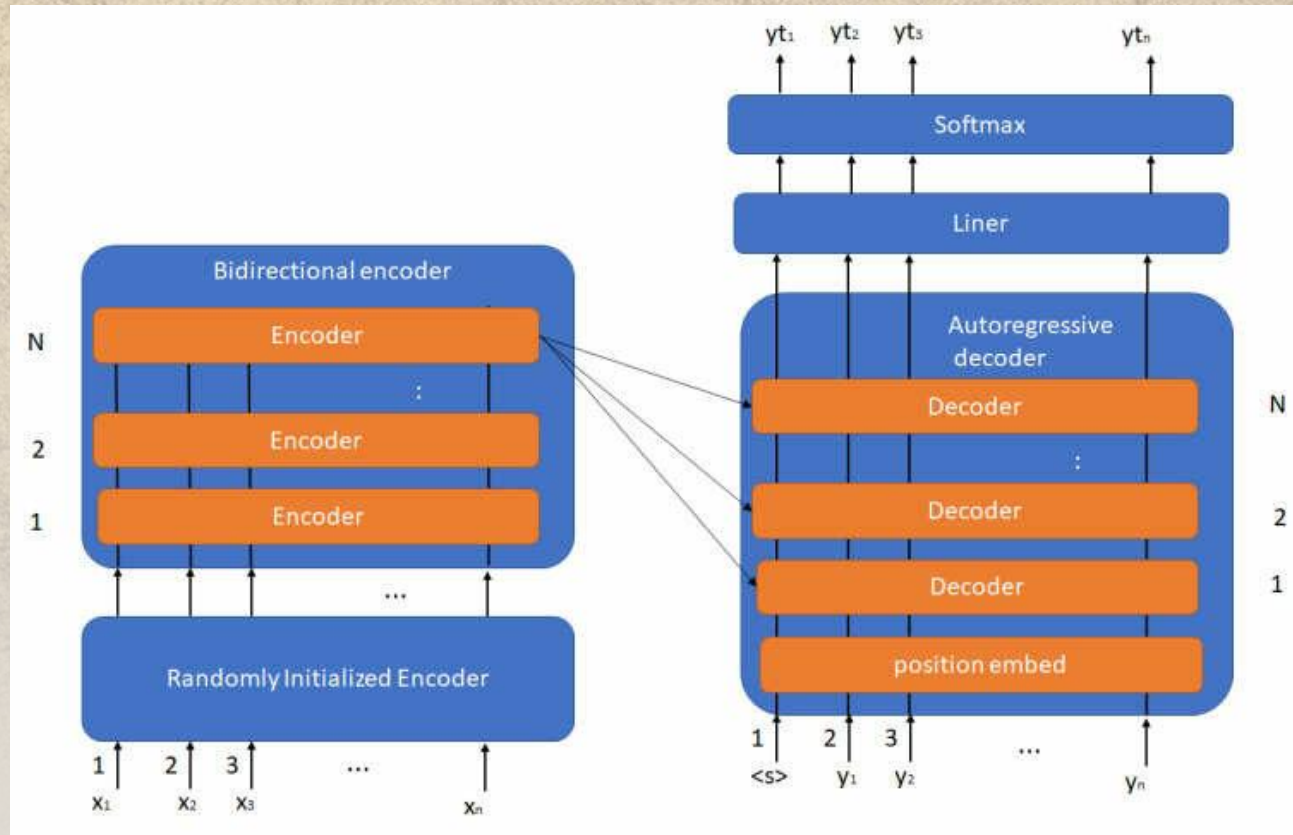
---

- **Multilingual based BART Model (IndicBART)**
  - It is a multilingual, sequence-to-sequence pre-trained transformer model focusing on 11 Indic languages( e.g. Bangla, Hindi, Marathi etc.) and English.
  - It combines bidirectional and auto-regressive learning, enabling it to understand both past and future context in a given text sequence, enhancing its ability to generate coherent and contextually appropriate responses.



CONTD..

## BART Architecture





# CONTD..

---

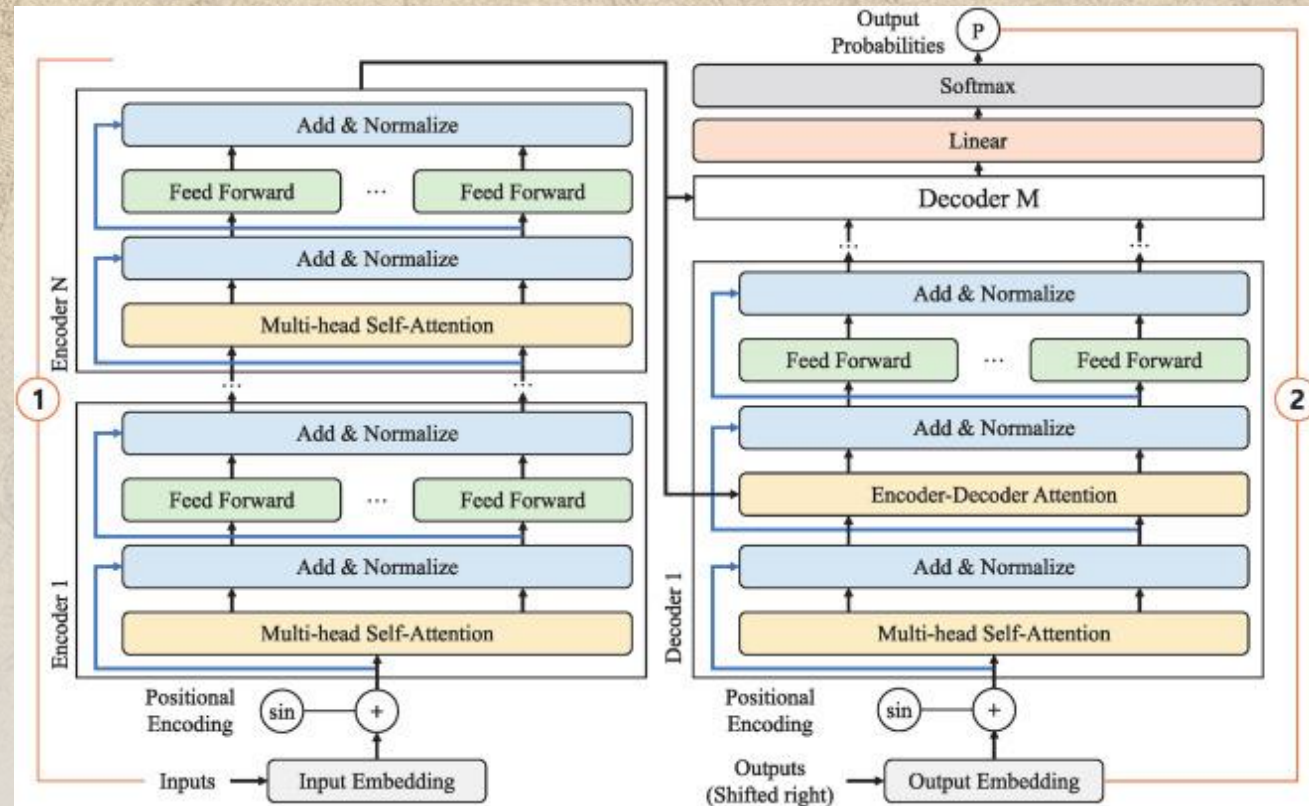
- **Multilingual based T5 Model (mT5)**

- It is a multilingual variant of the “Text-to-Text Transfer Transformer” or T5.
- It operates on the text-to-text principle, meaning it frames all NLP tasks as a text-to-text mapping problem.
- mT5 is designed to be language agnostic, meaning it can handle languages with varying grammar, syntax, and vocabulary without requiring language-specific modifications. This makes it highly adaptable to diverse linguistic environments



CONTD..

## T5 Architecture





# MODEL TRAINING

---

```
# Create a custom optimizer using torch.optim.AdamW
custom_optimizer = torch.optim.AdamW(
    model.parameters(),
    lr=1e-3,
    eps=1e-8,
    weight_decay=0.01,
)
```

```
# Define the TrainingArguments for fine-tuning
training_args = TrainingArguments(
    output_dir='/kaggle/working/',
    num_train_epochs=5,
    per_device_train_batch_size=5,
    gradient_accumulation_steps=8,
    evaluation_strategy="epoch",
    save_total_limit=1,
    save_steps=5000,
    learning_rate=1e-3,
    do_train=True,
    do_eval=True,
    remove_unused_columns=False,
    push_to_hub=False,
    report_to="none",
    load_best_model_at_end=False,
    lr_scheduler_type="cosine_with_restarts",
    warmup_steps=100,
    weight_decay=0.01,
    logging_dir='/kaggle/working/',
    logging_steps=200,
```



# CONTD..

---

## BART TRAINING

| Epoch | Training Loss | Validation Loss |
|-------|---------------|-----------------|
| 1     | 0.280600      | 0.260148        |
| 2     | 0.246400      | 0.250798        |
| 3     | 0.216000      | 0.247301        |
| 4     | 0.196200      | 0.247048        |
| 5     | 0.185300      | 0.248916        |

```
[17]: TrainOutput(global_step=4375, training_loss=0.3344861833844866, metrics={'train_runtime': 19744.5985, 'train_samples_per_second': 17.726, 'train_steps_per_second': 0.222, 'total_flos': 4.74070646784e+16, 'train_loss': 0.3344861833844866, 'epoch': 5.0})
```



# CONTD..

---

## T5 TRAINING

| Epoch | Training Loss | Validation Loss |
|-------|---------------|-----------------|
| 0     | 0.574500      | 0.404492        |
| 2     | 0.399700      | 0.362477        |
| 4     | 0.374200      | 0.355247        |

```
: TrainOutput(global_step=2185, training_loss=0.4426255005720824, metrics={'train_runtime': 17415.8632, 'train_samples_per_second': 20.097, 'train_steps_per_second': 0.125, 'total_flos': 4.6212722982912e+16, 'train_loss': 0.4426255005720824, 'epoch': 4.99})
```



# MODEL EVALUATION

---

- **ROUGE(Recall-Oriented Understudy for Gisting Evaluation) SCORE**

- The Summarization model is evaluated in terms of the ROUGE Score . The score matches the generated text to the reference. ROUGE-N stands for n-gram co-occurrences between the nominee and testimonial summaries, where n-gram is the total number of words to be evaluated.

$$ROUGE - N = \frac{\sum S_e S_h \sum_{g_n \in S} C_m(g_n)}{\sum S_e S_h \sum_{g_n \in S} C(g_n)}$$

- where  $S_h$  denotes a collection of nominee summaries,  $S_e$  denotes a single nominee summary,  $g_n$  denotes an N-gram, and  $C(g_n)$  is total times  $g_n$  comes to sight in both the nominee and automatic summary.



# CONTD..

---

- **BLEU (Bilingual Evaluation Understudy) SCORE**

- BLEU score measures the similarity between the machine-generated text and the reference using n-grams, which are contiguous sequences of n words. The most common n-grams used are unigrams (single words), bigrams (two-word sequences), trigrams (three-word sequences), and so on.

$$\text{BLEU} = \text{BP} * \exp(\sum p_n)$$

- Where, BP (Brevity Penalty) is a penalty term that adjusts the score for translations that are shorter than the reference translations. It is calculated as

$$\min(1, (\text{reference\_length} / \text{translated\_length}))$$

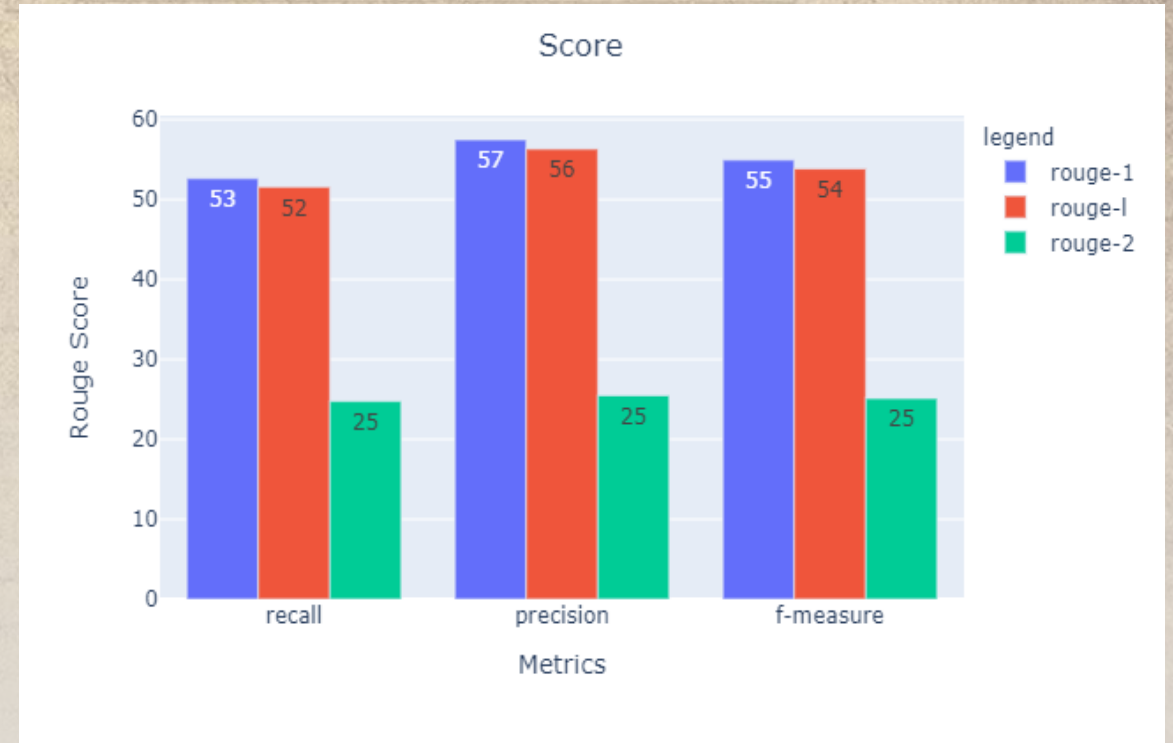
- $p_n$  is the precision of n-grams, which is calculated as the number of n-grams that appear in both the machine-generated translation and the reference divided by the total number of n-grams in the machine-generated translation.



# RESULT (BART)

|   | Summary  | BARTSummary                                       |
|---|--|---|
| 0 | मिशेल के जन्मदिन पर ओबामा ने दी शानदार दावत        | मिशेल ओबामा के जन्मदिन पर ओबामा ने की शानदार दावत |
| 1 | पतंजलि नमक का उड़ाया मजाक तो ट्रोल हो गए जावेद...  | रामदेव की कंपनी पर चुटकुले के बाद जावेद जाफरी ... |
| 2 | शिवराज के साले को चुनाव लड़ना पड़ा भारी मिले स...  | मध्य प्रदेश चुनाव वारासिवनी सीट पर काउंटिंग जारी  |
| 3 | रानी मुखर्जी को बेटी कहना काफी मुश्किल था शाहर...  | मेरे लिए फैन का मतलब है किसी को बेइंतेहा प्यार... |
| 4 | उत्तराखंड में जारी रहेगा राष्ट्रपति शासन केंद्र... | उत्तराखंड में राष्ट्रपति शासन पर सुप्रीम कोर्ट... |
| 5 | कैसे पूरा होगा मोदी का भारत जोड़ो सपना कर्नाटक...  | कर्नाटक के हालात पीएम मोदी के मंसूबों पर सवाल...  |

|                  | rouge-1  | rouge-2  | rouge-l  |
|------------------|----------|----------|----------|
| <b>recall</b>    | 0.526316 | 0.247619 | 0.515789 |
| <b>precision</b> | 0.574713 | 0.254902 | 0.563218 |
| <b>f-measure</b> | 0.549451 | 0.251208 | 0.538462 |





# CONTD..

---

## BLEU (Bilingual Evaluation Understudy) SCORE

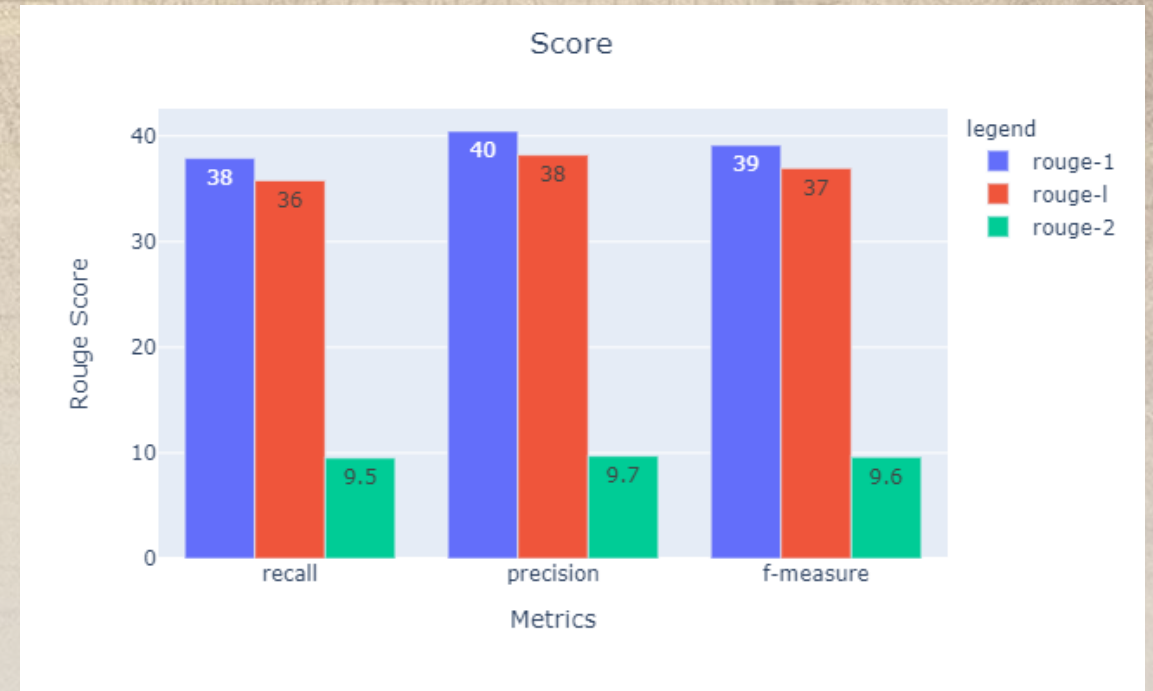
|   | Metric             | Value         |
|---|--------------------|---------------|
| 0 | BLEU Score         | 0.133607      |
| 1 | 1-gram Precision   | 0.404000      |
| 2 | 2-gram Precision   | 0.192700      |
| 3 | 3-gram Precision   | 0.102958      |
| 4 | 4-gram Precision   | 0.059781      |
| 5 | Brevity Penalty    | 0.903037      |
| 6 | Length Ratio       | 0.907447      |
| 7 | Translation Length | 101292.000000 |
| 8 | Reference Length   | 111623.000000 |



# RESULT (T5)

|   | Summary  | T5Summary   |
|---|--|---|
| 0 | मिशेल के जन्मदिन पर ओबामा ने दी शानदार दावत        | अमेरिका की प्रथम महिला मिशेल ओबामा ने किया शान... |
| 1 | पतंजलि नमक का उड़ाया मजाक तो ट्रोल हो गए जावेद...  | नमक के दाने समय के साथ जावेद जाफरी ने किया ट्रोल  |
| 2 | शिवराज के साले को चुनाव लड़ना पड़ा भारी मिले स...  | वारासिवनी विधानसभा सीट पर शिवराज सिंह के साले ... |
| 3 | रानी मुखर्जी को बेटी कहना काफी मुश्किल था शाहर...  | ऋषि कपूर की फिल्म फैन रिलीज जानें क्या है वजह     |
| 4 | उत्तराखंड में जारी रहेगा राष्ट्रपति शासन केंद्र... | नैनीताल हाईकोर्ट के फैसले पर बोले अटॉर्नी जनरल... |
| 5 | कैसे पूरा होगा मोदी का भारत जोड़ो सपना कर्नाटक...  | कर्नाटक में कन्नड़ फॉन्ट के इस्तेमाल करने की त... |

|                  | rouge-1  | rouge-2  | rouge-l  |
|------------------|----------|----------|----------|
| <b>recall</b>    | 0.378947 | 0.095238 | 0.357895 |
| <b>precision</b> | 0.404494 | 0.097087 | 0.382022 |
| <b>f-measure</b> | 0.391304 | 0.096154 | 0.369565 |





# CONTD..

---

## BLEU (Bilingual Evaluation Understudy) SCORE

|   | Metric             | Value         |
|---|--------------------|---------------|
| 0 | BLEU Score         | 0.089154      |
| 1 | 1-gram Precision   | 0.313743      |
| 2 | 2-gram Precision   | 0.131226      |
| 3 | 3-gram Precision   | 0.061919      |
| 4 | 4-gram Precision   | 0.032408      |
| 5 | Brevity Penalty    | 0.935139      |
| 6 | Length Ratio       | 0.937155      |
| 7 | Translation Length | 104608.000000 |
| 8 | Reference Length   | 111623.000000 |



# USER INTERFACE

---

## Hindi News Headline Generation

Enter Hindi News Article for Headline Generation:

चीनी मीडिया ने लिखा, "विश्लेषकों ने कहा कि यदि मोदी जीतते हैं, जो उन्हें देश के पहले प्रधानमंत्री जवाहरलाल नेहरू के बाद तीसरे कार्यकाल के लिए सत्ता में बने रहने वाला दूसरा भारतीय नेता बना देगा, तो मोदी की समग्र घरेलू और विदेश नीतियों में मजबूत निरंतरता बनी रहने की उम्मीद है, जिसमें महत्वपूर्ण बदलाव होने की संभावना कम है। सिंधुआ विश्वविद्यालय में राष्ट्रीय रणनीति संस्थान में अनुसंधान विभाग के निदेशक कियान फेंग ने रविवार को ग्लोबल टाइम्स को बताया कि मोदी भारत के लिए निर्धारित घरेलू और विदेश नीति के उद्देश्यों को आगे बढ़ाना जारी रखेंगे, जिसमें कुछ वर्षों के भीतर अमेरिका और चीन के बाद दुनिया की तीसरी सबसे बड़ी अर्थव्यवस्था बनने की दिशा में देश को आगे बढ़ाने पर जोर दिया जाएगा।

Summarize

### IndicBART Summary:

चीनी मीडिया का दावा मोदी जीतते हैं तो मोदी होंगे तीसरी सबसे बड़ी अर्थव्यवस्था

### mT5 Summary:

चीनी मीडिया ने बताया- मोदी जीतते हैं और विदेश नीति के उद्देश्यों को आगे बढ़ाना



# CONCLUSION

---

- The performance of transformer based pre-trained models INDIC-BART and mT5 on our dataset shows ROUGE-1 scores of 54.9 and 39.13 respectively on text summarization task during evaluation. BART has outperformed the T5 model in the result by 15.77%.
- The BLEU-score for INDIC-BART and mT5 on our dataset is 0.13 and 0.08 respectively.



# FUTURE WORK

---

- This project mainly focuses on the mono-lingual(hindi) for news summarization. The model can adapt to summarize news articles in multiple languages, thereby making it more versatile and globally applicable.
- Experimentation with more advanced transformer architectures like Longformer, Reformer, or BigBird to handle longer sequences more effectively and capture richer contextual information.
- Develop techniques to enable real-time summarization of live news feeds or streaming content, ensuring timely and relevant summaries for users.



# REFERENCES

---

- Rao, R., Sharma, S. & Malik, N. "Automatic text summarization using transformer-based language models". <https://link.springer.com/article/10.1007/s13198-024-02280-4>
- Padhma Muniraj, K.R. Sabarmathi, R. Leelavathi, Saravana Balaji B, "HNTSumm: Hybrid text summarization of transliterated news articles". <https://www.sciencedirect.com/science/article/pii/S2666603023000027>
- Fahad, A.R., Al Nahian, N., Islam, M.A. et al. "Answer Agnostic Question Generation in Bangla Language". <https://link.springer.com/article/10.1007/s44227-023-00018-5>
- Kumar, A., Parida, S., Pratap, A. et al. "Machine translation by projecting text into the same phonetic-orthographic space using a common encoding". <https://link.springer.com/article/10.1007/s12046-023-02275-0>
- Jannatul Ferdous Ruma, Tasmiah Tahsin Mayeesha, Rashedur M. Rahman, "Transformer based Answer-Aware Bengali Question Generation". <https://www.sciencedirect.com/science/article/pii/S2666307423000311#bib0010>
- Supriyono, Aji Prasetya Wibawa, Suyono, Fachrul Kurniawan, "A survey of text summarization: Techniques, evaluation and challenges". <https://www.sciencedirect.com/science/article/pii/S2949719124000189>





THANKYOU

---

