Shironaam: Bengali News Headline Generation using Auxiliary Information

Abu Ubaida Akash[1*], Mir Tafseer Nayeem[2*], Faisal Tareque Shohan[1], Tanvir Islam[3]

[1]Ahsanullah University of Science and Technology
[2]University of Alberta
[3]University of Hawaii at Manoa

* *[equal contribution]*

News Headline

News Headline

- **Importance**
  - Catching the reader's attention
  - Providing Context
  - Enhancing Search Engine Optimization (SEO)
  - Establishing Credibility

Headline Generation

Headline Generation

- **A special case of abstractive summarization**
  - Does not often maintain grammatical structure
  - More extreme than extreme summarization
  - Highly abstractive

Headline Generation

- **A special case of abstractive summarization**
  - Does not often maintain grammatical structure
  - More extreme than extreme summarization
  - Highly abstractive
- **Involves**
  - Sentence compression
  - Syntactic reorganization
  - Lexical paraphrasing
  - Sentence fusion

Headline Generation

Headline Generation

- **Typically one-to-one mapping (input ← article, output ← headline)**
  - Takase et al. (2016), Zhang et al. (2018), Murao et al. (2019), Colmenares et al. (2019), Song et al. (2020), Li et al. (2021)

Headline Generation

- **Typically one-to-one mapping (input ← article, output ← headline)**
  - Takase et al. (2016), Zhang et al. (2018), Murao et al. (2019), Colmenares et al. (2019), Song et al. (2020), Li et al. (2021)
- **Makes it difficult when the input is necessarily long**
  - Contextualized language models suffer from a limited sequence

Headline Generation

- **Typically one-to-one mapping (input ← article, output ← headline)**
  - Takase et al. (2016), Zhang et al. (2018), Murao et al. (2019), Colmenares et al. (2019), Song et al. (2020), Li et al. (2021)
- **Makes it difficult when the input is necessarily long**
  - Contextualized language models suffer from a limited sequence
- **More challenging for low-resource languages**
  - Unavailability of large-scale human-annotated dataset
  - Limited language models
  - Lack of SOTA models for the downstream task

Our Contributions

Our Contributions

1. **Provided Shironaam, a large-scale news headline generation dataset**
    a. Largest for a low-resource language *i.e.* Bengali
    b. Contains auxiliary information along with article-headline pairs

Our Contributions

1. **Provided Shironaam, a large-scale news headline generation dataset**
    a. Largest for a low-resource language *i.e.* Bengali
    b. Contains auxiliary information along with article-headline pairs
2. **Presented the concept of incorporating auxiliary information in headline generation**
    a. Developed an end-to-end SOTA model for headline generation

Our Contributions

1. **Provided Shironaam, a large-scale news headline generation dataset**
    a. Largest for a low-resource language *i.e.* Bengali
    b. Contains auxiliary information along with article-headline pairs
2. **Presented the concept of incorporating auxiliary information in headline generation**
    a. Developed an end-to-end SOTA model for headline generation
3. **Developed BenSim, a module for measuring semantic similarity among Bengali sentences**
    a. Helps to encode long documents

Our Contributions

1. **Provided Shironaam, a large-scale news headline generation dataset**
   a. Largest for a low-resource language *i.e.* Bengali
   b. Contains auxiliary information along with article-headline pairs
2. **Presented the concept of incorporating auxiliary information in headline generation**
   a. Developed an end-to-end SOTA model for headline generation
3. **Developed BenSim, a module for measuring semantic similarity among Bengali sentences**
   a. Helps to encode long documents
4. **Illustrated the utility and robustness by evaluating the performance with few-shot settings**

Dataset

Dataset

Raw Data Crawling

- ❖ **7** Bengali newspapers
- ❖ **13** different domains

Dataset

Raw Data Crawling

- **7** Bengali newspapers
- **13** different domains

900,000 samples →

Data Preprocessing

- Removed datetime and embedded items
- Preserved only Bengali texts
- Retained English numbers
- Removed repetitive terms from caption
- Discarded samples where len(caption) < 4 words
- Mapped random categories into general terms
    - national ← (national, whole-country, city-news, country, capital, city-roundup, south-city)
- Discarded samples with any missing features (*i.e.* headline, article, caption, or category)

Dataset

Raw Data Crawling

- **7** Bengali newspapers
- **13** different domains

900,000 samples →

Data Preprocessing

- Removed datetime and embedded items
- Preserved only Bengali texts
- Retained English numbers
- Removed repetitive terms from caption
- Discarded samples where len(caption) < 4 words
- Mapped random categories into general terms
  - national ← (national, whole-country, city-news, country, capital, city-roundup, south-city)
- Discarded samples with any missing features (*i.e.* headline, article, caption, or category)

← 240,580 samples

Shironaam corpus

- Headline
- Article
- Image caption
- Category
- Topic word

Dataset Statistics

Dataset Statistics

Category	Total	Jaccard (%)	Category	Total	Jaccard (%)
Entertainment	17,565	13.56	Miscellaneous	1,744	11.71
National	128,226	24.60	Opinion	3,819	38.41
Nature	510	23.66	Politics	16,380	23.02
International	33,329	18.09	Edu-Career	4,372	53.58
Sports	19,235	17.82	Science-Tech	1,141	22.95
Economy	7,032	39.37	Religion	294	71.59
Life-Health	6,933	17.83	**Total/Avg.**	**240,580**	**28.94**

Dataset Statistics

- (Train, valid, test): All categories **(92, 2, 6)%**

Category	Total	Jaccard (%)	Category	Total	Jaccard (%)
Entertainment	17,565	13.56	Miscellaneous	1,744	11.71
National	128,226	24.60	Opinion	3,819	38.41
Nature	510	23.66	Politics	16,380	23.02
International	33,329	18.09	Edu-Career	4,372	53.58
Sports	19,235	17.82	Science-Tech	1,141	22.95
Economy	7,032	39.37	Religion	294	71.59
Life-Health	6,933	17.83	**Total/Avg.**	**240,580**	**28.94**

Dataset Statistics

- (Train, valid, test): All categories **(92, 2, 6)%**
- Total (train, valid, test): **(220574, 4994, 15012)**

Category	Total	Jaccard (%)	Category	Total	Jaccard (%)
Entertainment	17,565	13.56	Miscellaneous	1,744	11.71
National	128,226	24.60	Opinion	3,819	38.41
Nature	510	23.66	Politics	16,380	23.02
International	33,329	18.09	Edu-Career	4,372	53.58
Sports	19,235	17.82	Science-Tech	1,141	22.95
Economy	7,032	39.37	Religion	294	71.59
Life-Health	6,933	17.83	**Total/Avg.**	**240,580**	**28.94**

Dataset Statistics

- (Train, valid, test): All categories **(92, 2, 6)%**
- Total (train, valid, test): **(220574, 4994, 15012)**
- Jaccard scores: Similarities (caption ⇆ headline)

Category	Total	Jaccard (%)	Category	Total	Jaccard (%)
Entertainment	17,565	13.56	Miscellaneous	1,744	11.71
National	128,226	24.60	Opinion	3,819	38.41
Nature	510	23.66	Politics	16,380	23.02
International	33,329	18.09	Edu-Career	4,372	53.58
Sports	19,235	17.82	Science-Tech	1,141	22.95
Economy	7,032	39.37	Religion	294	71.59
Life-Health	6,933	17.83	**Total/Avg.**	**240,580**	**28.94**

Dataset Statistics

Features	IndicNLG-BN	Shironaam
Article	Yes	Yes
Headline	Yes	Yes
Image Caption	No	Yes
Category	No	Yes
Topic Word	No	Yes
#Samples	142,731	240,580

Dataset Statistics

Features	IndicNLG-BN	Shironaam
Article	Yes	Yes
Headline	Yes	Yes
Image Caption	No	Yes
Category	No	Yes
Topic Word	No	Yes
#Samples	142,731	240,580

Dataset	% of novel n-gram			
	unigram	bigram	trigram	4-gram
IndicNLG-BN	26.59	66.12	82.71	86.49
Shironaam	46.38	78.92	90.39	94.77

Dataset Statistics

Dataset	Average number of words	IndicNLG BN	Shironaam	Average number of sentences	IndicNLG BN	Shironaam	Vocabulary size	IndicNLG BN	Shironaam
Article		199.83	252.01		15.19	20.05		614,374	605,750
Headline		10.03	6.53		1.19	1.00		65,553	76,732
Image Caption		-	6.80		-	1.04		-	87,644
Topic Words		-	3.21		-	-		-	-

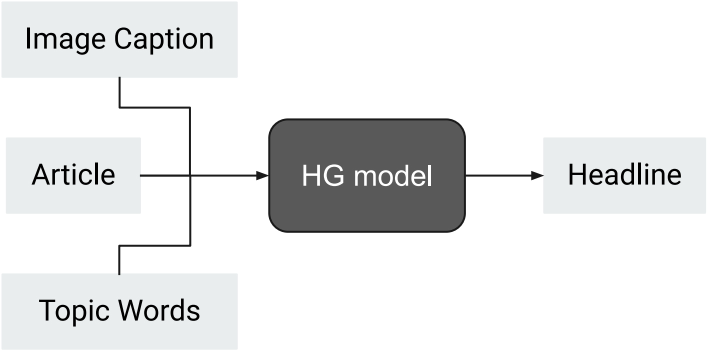
Task

Task

Previously: One-to-One

Task

Previously: One-to-One



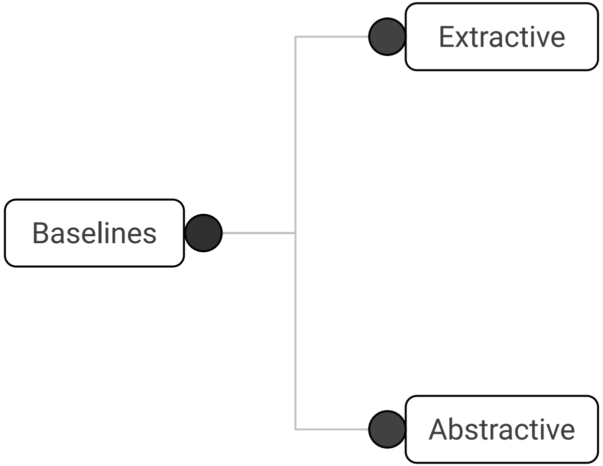
Our task: Three-to-One

Approach

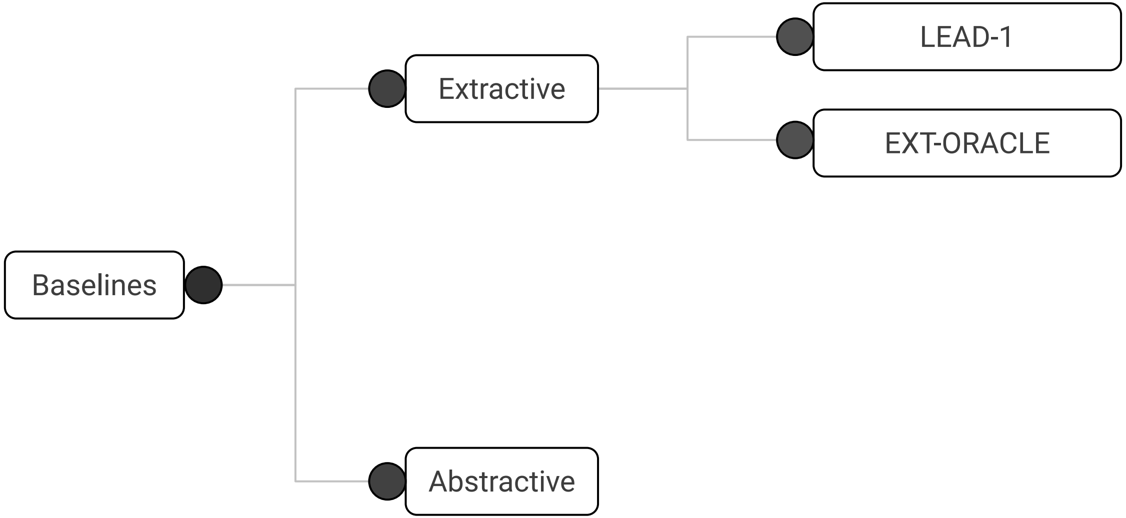
Approach

Baselines

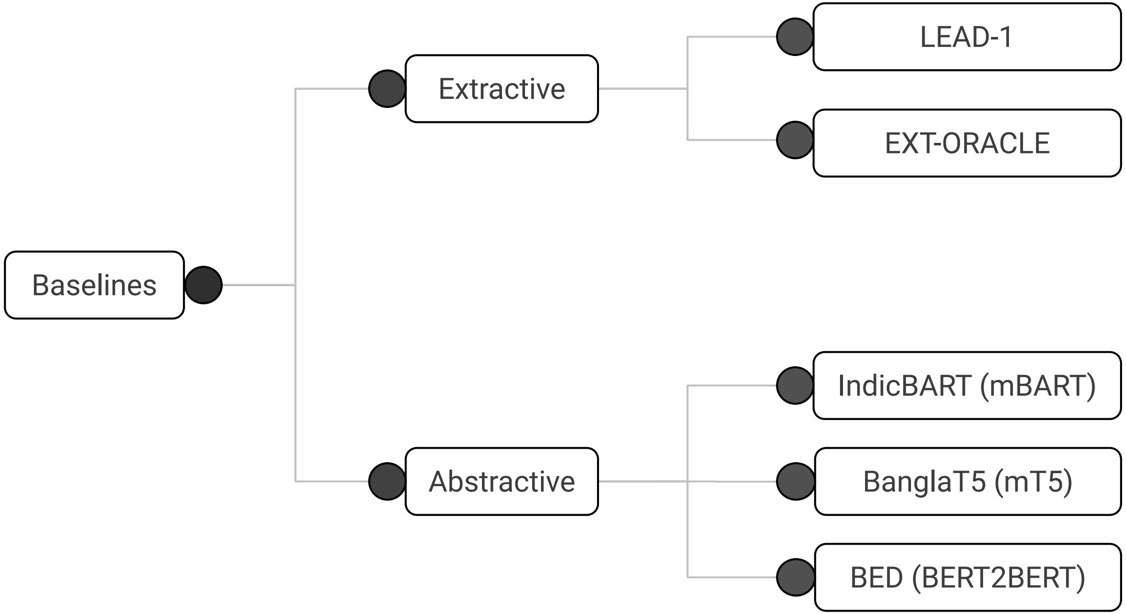
Approach



Approach

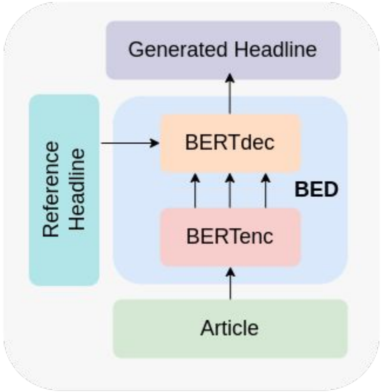


Approach


Baselines
Extractive
LEAD-1
EXT-ORACLE
Abstractive
IndicBART (mBART)
BanglaT5 (mT5)
BED (BERT2BERT)

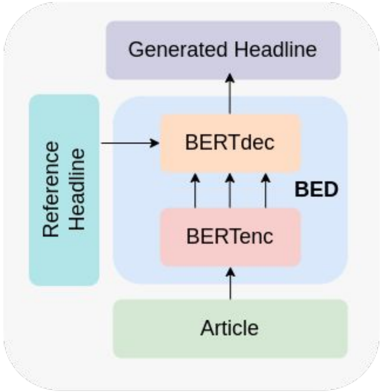
Proposed Model

Proposed Model



(a) BED(base)

Proposed Model

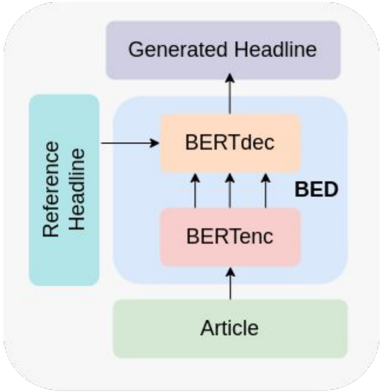


(a) BED(base)

BERT based Encoder Decoder (BED)

- Both encoder and decoder weights initialization with pre-trained BERT checkpoint (*e.g.* BanglaBERT)
- Cross attention weights randomly initialized
- Hugging Face encoder-decoder paradigm

Proposed Model



(a) BED(base)

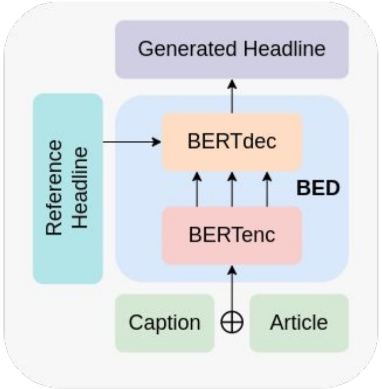
BERT based Encoder Decoder (BED)

a) *Article Only:*

- Input: Article; Output: Headline
- First SOTA baseline in Bengali language

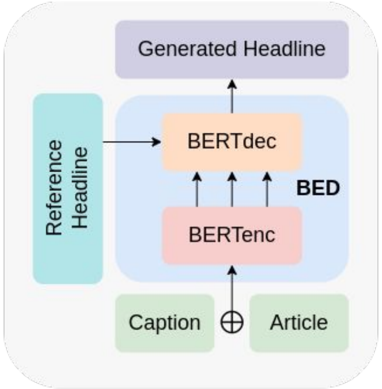
Proposed Model

Proposed Model



(b) BED(w/ Article + Caption)

Proposed Model



(b) BED(w/ Article + Caption)

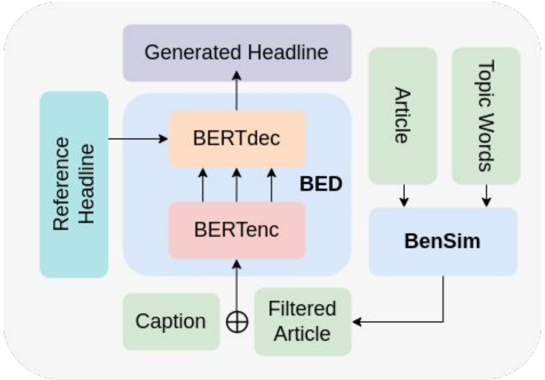
BERT based Encoder Decoder (BED)

b) *Article and Image Caption:*

- Input: Article, Image caption; Output: Headline
- Parallel fusion mechanism
- Separated by a special token

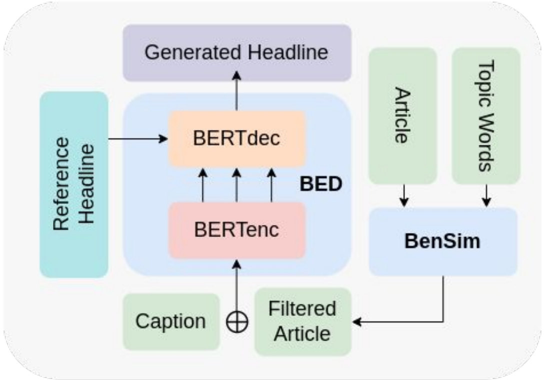
Proposed Model

Proposed Model



(c) BED(w/ FilteredArticle + Caption)

Proposed Model



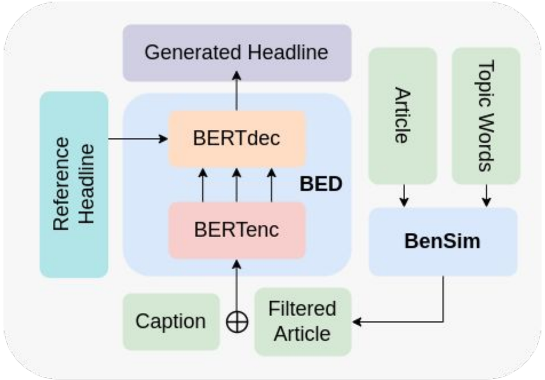
(c) BED(w/ FilteredArticle + Caption)

BERT based Encoder Decoder (BED)

c) *Filtered Article and Image Caption:*

- Input: Article, Image caption, Topic words; Output: Headline
- Parallel fusion mechanism
- Separated by a special token
- Additionally BenSim

Proposed Model



(c) BED(w/ FilteredArticle + Caption)

BERT based Encoder Decoder (BED)

- *BenSim Module:*
  - Input: Article, Topic words; Output: Filtered article
  - Measures semantic similarity between Bengali sentences utilizing bangla-bert-base embeddings
  - Picks most relevant sentences from long articles (we consider top 40)
  - Mean pool operation followed by Cosine similarity

Experiments

Experiments

RQ #1

Can we use auxiliary information (e.g., image caption and topic words) to improve the performance of the headline generation?

Experiments

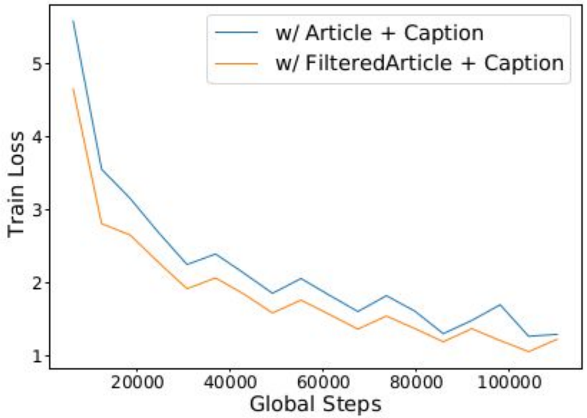
RQ #1 Can we use auxiliary information (e.g., image caption and topic words) to improve the performance of the headline generation?

RQ #2 Which domain(s) benefit from the auxiliary information in few-shot and non few-shot settings?

Experiments

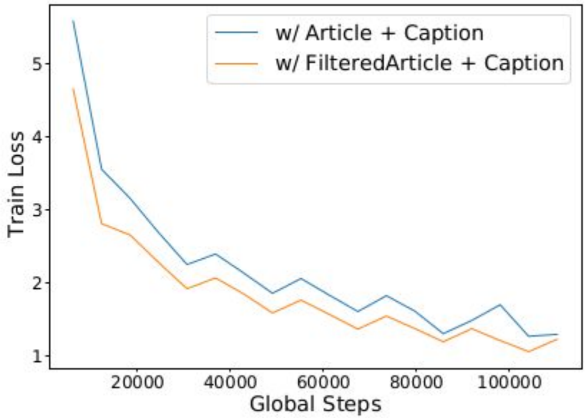
Models		Rouge			BLEU			BERT Score	METEOR Score
		R-1	R-2	R-L	BLEU Score	Brevity Penalty	Length Ratio		
Baselines	LEAD-1	30.50	13.86	28.00	5.65	97.71	2.48	74.63	29.90
	EXT-ORACLE	39.92	22.89	37.28	9.17	97.16	2.30	77.16	39.65
	IndicBART	28.76	12.65	27.11	15.03	99.91	1.14	74.95	20.39
	BanglaT5	44.13	23.03	42.12	13.05	91.33	1.15	80.13	34.65
Our Ablations	BED Base	44.22	24.18	42.28	22.06	94.47	0.94	80.53	34.16
	BED (Article+Caption)	51.62	33.62	49.94	31.39	96.02	0.96	82.93	42.57
	BED (FilteredArticle+Caption)	**52.19**	**34.27**	**50.31**	**31.80**	**98.57**	**0.99**	**83.10**	**43.52**

Experiments



Experiments

- Few lengthier articles in Shironaam
- Slightly better performance
- Learns faster with the filtered articles
- Score difference will increase with the number of longer articles
- Following RQ#1, auxiliary information aids headline generation



Domain Specific Analysis

Domain Specific Analysis

Category	R-1			R-2			R-L		
	BED (base)	**BNT5**	**BED (FA+C)**	**BED (base)**	**BNT5**	**BED (FA+C)**	**BED (base)**	**BNT5**	**BED (FA+C)**
Non-Few-Shot Domains									
National	48.03	47.33	55.84	27.29	25.83	37.88	46.06	45.37	53.95
International	44.44	46.04	50.47	22.92	23.08	29.96	42.02	43.49	48.13
Sports	30.14	33.46	39.20	11.57	13.43	20.40	28.75	31.59	37.33
Entertainment	33.05	32.99	35.14	15.07	14.32	16.64	31.26	31.33	33.44
Politics	49.28	49.66	**57.16**	28.80	27.32	**39.73**	47.53	47.68	**55.73**
Few-Shot Domains									
Economy	38.95	40.03	60.32	18.81	19.74	45.85	36.44	37.62	58.53
Life-Health	35.87	39.20	44.97	17.61	19.78	27.21	33.90	37.38	43.08
Edu-Career	50.57	51.12	71.55	31.92	30.82	59.54	48.05	48.82	70.48
Opinion	16.11	15.82	44.53	4.69	5.24	36.63	15.82	15.44	44.25
Miscellaneous	33.64	34.92	35.29	16.16	17.98	17.41	30.48	32.82	31.87
Science-Tech	41.82	44.14	51.03	19.54	22.61	31.20	39.30	41.82	48.49
Nature	36.07	37.89	46.54	15.78	16.65	30.07	34.84	35.79	45.53
Religion	27.29	35.48	**72.10**	12.28	19.63	**62.05**	26.96	34.42	**72.14**

Domain Specific Analysis

Category	R-1 BED (base)	R-1 BNT5	R-1 BED (FA+C)	R-2 BED (base)	R-2 BNT5	R-2 BED (FA+C)	R-L BED (base)	R-L BNT5	R-L BED (FA+C)
Non-Few-Shot Domains									
National	48.03	47.33	55.84	27.29	25.83	37.88	46.06	45.37	53.95
International	44.44	46.04	50.47	22.92	23.08	29.96	42.02	43.49	48.13
Sports	30.14	33.46	39.20	11.57	13.43	20.40	28.75	31.59	37.33
Entertainment	33.05	32.99	35.14	15.07	14.32	16.64	31.26	31.33	33.44
Politics	49.28	49.66	**57.16**	28.80	27.32	**39.73**	47.53	47.68	**55.73**
Few-Shot Domains									
Economy	38.95	40.03	60.32	18.81	19.74	45.85	36.44	37.62	58.53
Life-Health	35.87	39.20	44.97	17.61	19.78	27.21	33.90	37.38	43.08
Edu-Career	50.57	51.12	71.55	31.92	30.82	59.54	48.05	48.82	70.48
Opinion	16.11	15.82	44.53	4.69	5.24	36.63	15.82	15.44	44.25
Miscellaneous	33.64	34.92	35.29	16.16	17.98	17.41	30.48	32.82	31.87
Science-Tech	41.82	44.14	51.03	19.54	22.61	31.20	39.30	41.82	48.49
Nature	36.07	37.89	46.54	15.78	16.65	30.07	34.84	35.79	45.53
Religion	27.29	35.48	**72.10**	12.28	19.63	**62.05**	26.96	34.42	**72.14**

- Two baselines: BED (base), BanglaT5 (BNT5)
- Few shot domain less than 6500 samples
- Entertainment: Casual, click-bait style, no identical nature
- Miscellaneous: Randomness of various domains

Future Works

Future Works

- Utilization of multimodal information
- Human evaluation on generated samples
- Language agnostic model

Thank You!