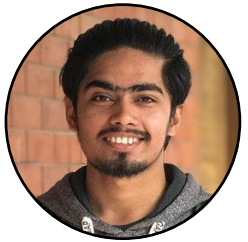
Unsupervised Abstractive Summarization of Bengali Text Documents

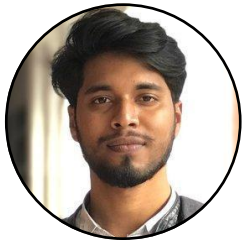
Radia Rayan Chowdhury *

Mir Tafseer Nayeem *

Tahsin Tasnim Mim



Md. Saifur Rahman Chowdhury



Taufiqul Jannat

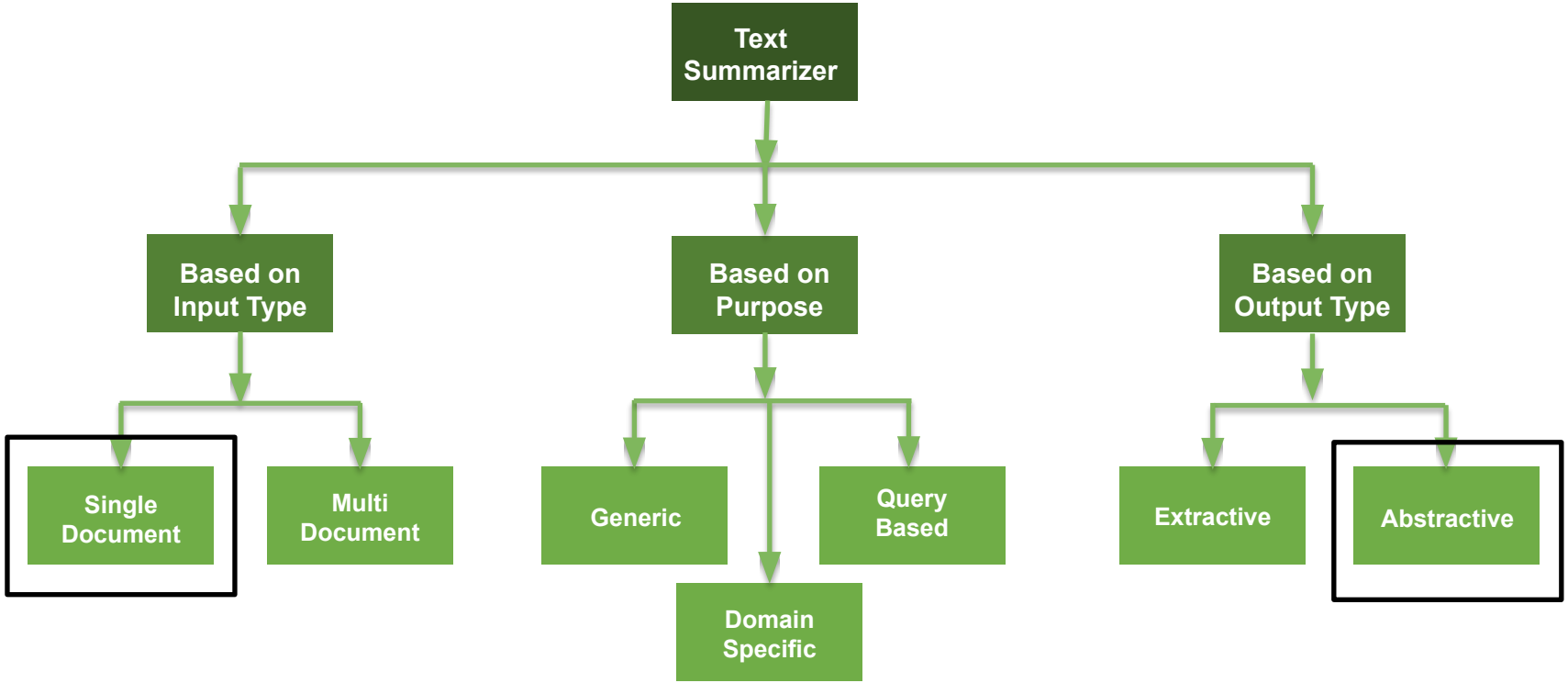
Ahsanullah University of Science and Technology

∗ Equal Contribution, listed by alphabetical order

Text Summarization

- Compression of large document
- Represents the most important or relevant information within the original content
- Based on natural language processing

Types of Summarisation



Selecting Bengali Language for our research

Native language of **Bangladesh**
Also used in **India** (e.g., **West Bengal, Tripura, Asam**)

7th most spoken language in the world

250 million native speakers [1]

[1] https://en.wikipedia.org/wiki/Bengali_language

Selecting Bengali Language for our research

New field of research

Lack of fundamental resources for **Natural Language Processing (NLP)**

OUR MODEL

BenSumm Model (Abstractive):

This is the First **Ben**gali Unsupervised Text **Summ**arizer for single document setting!!!

Why Unsupervised

- Effective
- Domain Independent
- No Need to Train Data
- Low Resource

Process of BenSumm Model

Working Flow

Figure: **BenSumm Model**

Text Preprocessing

Tokenization

- Sentence Tokenize
- Word Tokenize

Removing Punctuation

Punctuation do not contribute anything to the meaning of the sentence

Removing Stopwords

Stopwords can be destructing and non-informative and are additional memory overhead

POS Tagging

POS tagging is the process of marking up a word in a corpus based on its context and definition

Document Clustering

- Large document into small meaningful cluster
- Differentiate documents into different number of groups

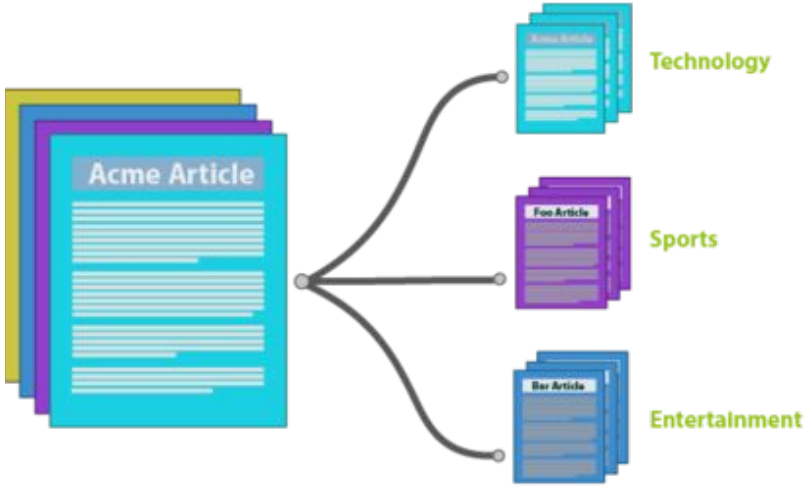
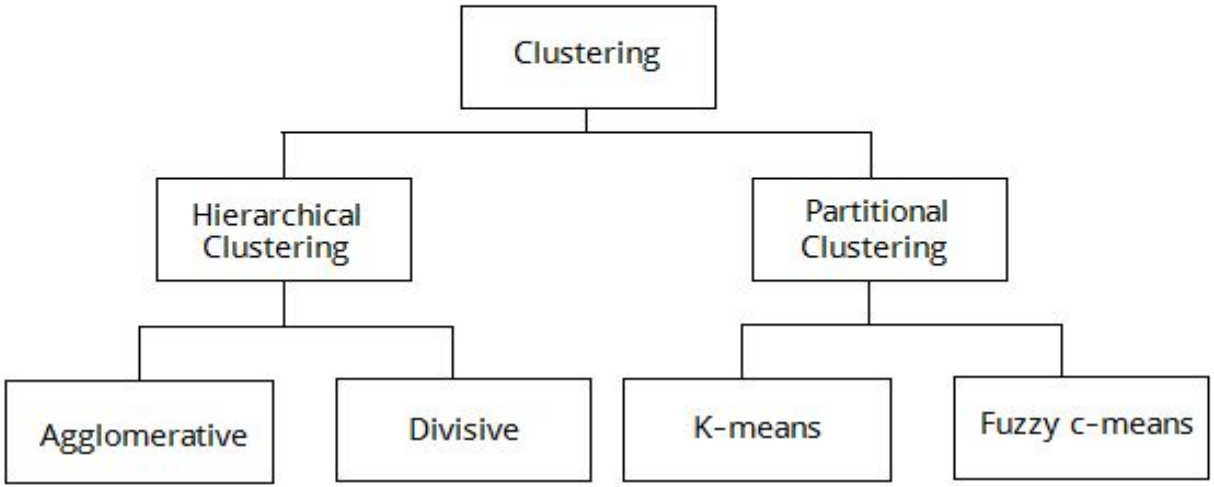


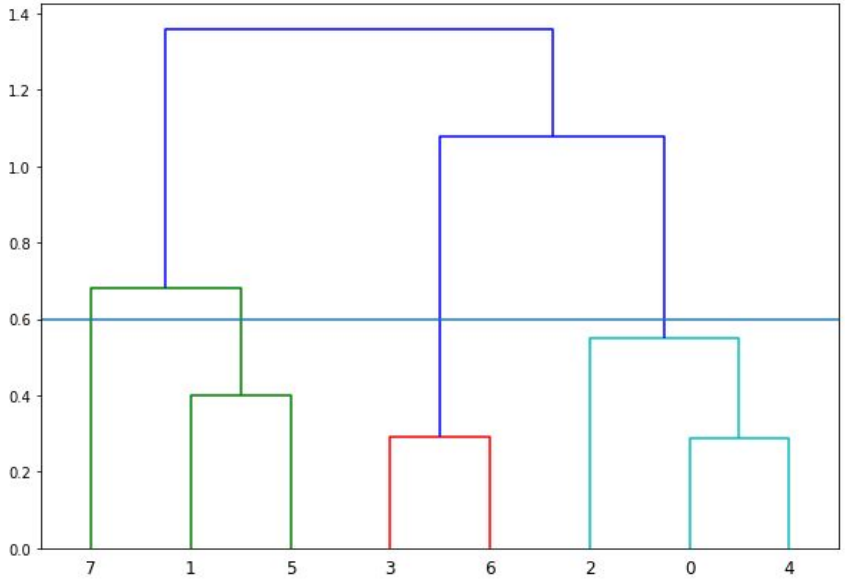
Figure: **Document Clustering**

Types of Clustering



Clustering in our System

- Hierarchical Agglomerative Clustering
- Bottom Up Approach



Hierarchical Agglomerative Clustering

Clustering in our System

- Calculating Cosine Similarity using **ULMFit pre-trained language model** (Aggarwal and Zhai, 2012)
- Measure the number of clusters for a given document using the **silhouette value** [2]

[2] https://shorturl.at/efgpF

Why Document Clustering

- An intermediate step of summarization
- Avoid incoherent summary
- To ensure good coverage
- Avoid redundancy

Example

Document Clustering

Input:

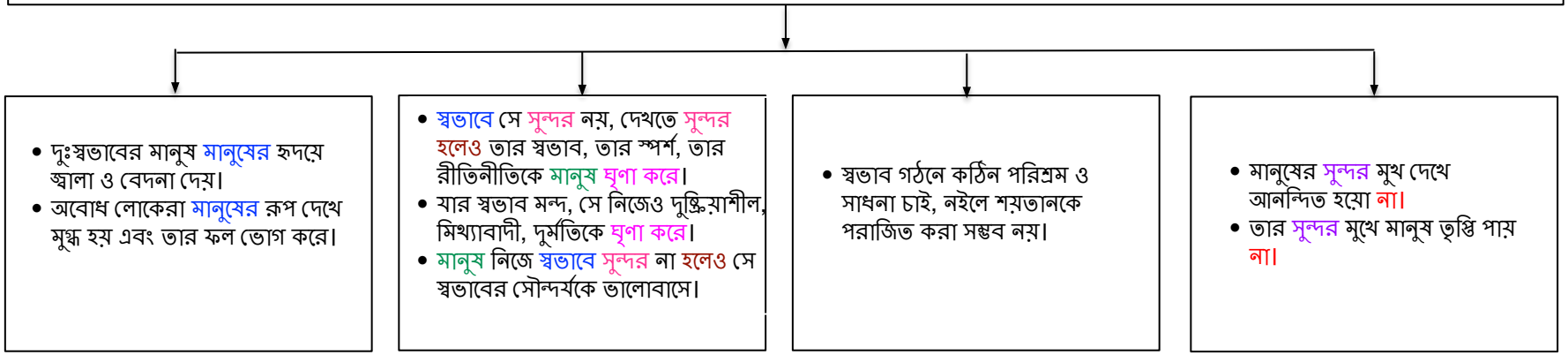
মানুষের সুন্দর মুখ দেখে আনন্দিত হয়ো না। স্বভাবে সে সুন্দর নয়, দেখতে সুন্দর হলেও তার স্বভাব, তার স্পর্শ, তার রীতিনীতিকে মানুষ ঘৃণা করে। দুঃস্বভাবের মানুষ মানুষের হৃদয়ে জ্বালা ও বেদনা দেয়। তার সুন্দর মুখে মানুষ তৃপ্তি পায় না। অবোধ লোকেরা মানুষের রূপ দেখে মুগ্ধ হয় এবং তার ফল ভোগ করে। যার স্বভাব মন্দ, সে নিজেও দুষ্ক্রিয়াশীল, মিথ্যাবাদী, দুর্মতিকে ঘৃণা করে। মানুষ নিজে স্বভাবে সুন্দর না হলেও সে স্বভাবের সৌন্দর্যকে ভালোবাসে। স্বভাব গঠনে কঠিন পরিশ্রম ও সাধনা চাই, নইলে শয়তানকে পরাজিত করা সম্ভব নয়।

[Do not be happy to see the beautiful faces of people. He/She is not beautiful by nature, although he/she is beautiful to look at, people hate his/her nature, touch, and manners. People with bad temper irritate and hurt people's hearts. People are not satisfied with the beautiful face. Ignorant people are fascinated by the human form and suffer in the long run. The one whose nature is evil, he is mischievous, a liar, and evil. Man himself is not beautiful by nature, but he loves the beauty of people's nature. We need hard work and pursuit to form nature; otherwise, it is impossible to defeat the devil.]

Document Clustering

Input:

মানুষের সুন্দর মুখ দেখে আনন্দিত হয়ো না। স্বভাবে সে সুন্দর নয়, দেখতে সুন্দর হলেও তার স্বভাব, তার স্পর্শ, তার রীতিনীতিকে মানুষ ঘৃণা করে। দুঃস্বভাবের মানুষ মানুষের হৃদয়ে জ্বালা ও বেদনা দেয়। তার সুন্দর মুখে মানুষ তৃপ্তি পায় না। অবোধ লোকেরা মানুষের রূপ দেখে মুগ্ধ হয় এবং তার ফল ভোগ করে। যার স্বভাব মন্দ, সে নিজেও দুষ্ক্রিয়াশীল, মিথ্যাবাদী, দুর্মতিকে ঘৃণা করে। মানুষ নিজে স্বভাবে সুন্দর না হলেও সে স্বভাবের সৌন্দর্যকে ভালোবাসে। স্বভাব গঠনে কঠিন পরিশ্রম ও সাধনা চাই, নইলে শয়তানকে পরাজিত করা সম্ভব নয়।



Cluster 1 **Cluster 2** **Cluster 3** **Cluster 4**

Word Graph- For Cluster 4

Cluster 4

Sentence 1 : মানুষের সুন্দর মুখ দেখে আনন্দিত হয়ো না।
Sentence 2 : তার সুন্দর মুখে মানুষ তৃপ্তি পায় না।

[Do not be happy to see the beautiful faces of people. People are not satisfied with the beautiful face.]

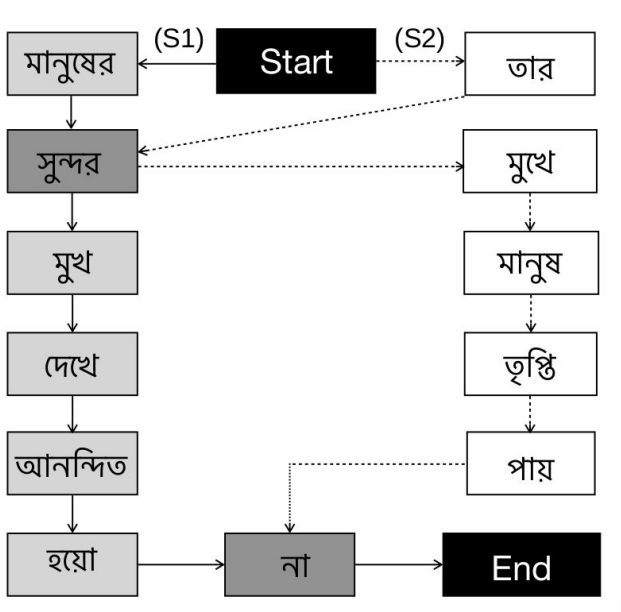


Figure: Word Graph

Sentence Fusion, Ranking & Selection

Fused & Ranked Sentences:

0.783 মানুষের সুন্দর মুখে মানুষ তৃপ্তি পায় না।
[People are not satisfied with the beautiful faces of people]

0.812 তার সুন্দর মুখে মানুষ তৃপ্তি পায় না।
[People are not satisfied with the beautiful face]

1.512 মানুষের সুন্দর মুখ দেখে আনন্দিত হয়ো না।
[Don't be happy to see the beautiful faces of people]

1.625 তার সুন্দর মুখ দেখে আনন্দিত হয়ো না।
[Don't be happy to see the beautiful faces]

Selected Sentence:

তার সুন্দর মুখ দেখে আনন্দিত হয়ো না।

Sentence Selection & Merge

Cluster 1	Cluster 2	Cluster 3	Cluster 4
• দুঃস্বভাবের মানুষ মানুষের হৃদয়ে জ্বালা ও বেদনা দেয়। • অবোধ লোকেরা মানুষের রূপ দেখে মুগ্ধ হয় এবং তার ফল ভোগ করে।	• স্বভাবে সে সুন্দর নয়, দেখতে সুন্দর হলেও তার স্বভাব, তার স্পর্শ, তার রীতিনীতিকে মানুষ ঘৃণা করে। • যার স্বভাব মন্দ, সে নিজেও দুষ্ক্রিয়াশীল, মিথ্যাবাদী, দুর্মতিকে ঘৃণা করে। • মানুষ নিজে স্বভাবে সুন্দর না হলেও সে স্বভাবের সৌন্দর্যকে ভালোবাসে।	• স্বভাব গঠনে কঠিন পরিশ্রম ও সাধনা চাই, নইলে শয়তানকে পরাজিত করা সম্ভব নয়।	• মানুষের সুন্দর মুখ দেখে আনন্দিত হয়ো না। • তার সুন্দর মুখে মানুষ তৃপ্তি পায় না।
দুঃস্বভাবের মানুষ মানুষের রূপ দেখে মুগ্ধ হয় এবং তার ফল ভোগ করে।	যার স্বভাব, তার স্পর্শ, তার রীতিনীতিকে মানুষ ঘৃণা করে।	• স্বভাব গঠনে কঠিন পরিশ্রম ও সাধনা চাই, নইলে শয়তানকে পরাজিত করা সম্ভব নয়।	তার সুন্দর মুখ দেখে আনন্দিত হয়ো না।

Summary:

দুঃস্বভাবের মানুষ মানুষের রূপ দেখে মুগ্ধ হয় এবং তার ফল ভোগ করে। যার স্বভাব, তার স্পর্শ, তার রীতিনীতিকে মানুষ ঘৃণা করে। স্বভাব গঠনে কঠিন পরিশ্রম ও সাধনা চাই, নইলে শয়তানকে পরাজিত করা সম্ভব নয়। তার সুন্দর মুখ দেখে আনন্দিত হয়ো না।

Final Result

Input

মানুষের সুন্দর মুখ দেখে আনন্দিত হয়ো না। স্বভাবে সে সুন্দর নয়, দেখতে সুন্দর হলেও তার স্বভাব, তার স্পর্শ, তার রীতিনীতিকে মানুষ ঘৃণা করে। দুঃস্বভাবের মানুষ মানুষের হৃদয়ে জ্বালা ও বেদনা দেয়। তার সুন্দর মুখে মানুষ তৃপ্তি পায় না। অবোধ লোকেরা মানুষের রূপ দেখে মুগ্ধ হয় এবং তার ফল ভোগ করে। যার স্বভাব মন্দ, সে নিজেও দুষ্ক্রিয়াশীল, মিথ্যাবাদী, দুর্মতিকে ঘৃণা করে। মানুষ নিজে স্বভাবে সুন্দর না হলেও সে স্বভাবের সৌন্দর্যকে ভালোবাসে। স্বভাব গঠনে কঠিন পরিশ্রম ও সাধনা চাই, নইলে শয়তানকে পরাজিত করা সম্ভব নয়।

[Do not be happy to see the beautiful faces of people. He/She is not beautiful by nature, although he/she is beautiful to look at, people hate his/her nature, touch, and manners. People with bad temper irritate and hurt people's hearts. People are not satisfied with the beautiful face. Ignorant people are fascinated by the human form and suffer in the long run. The one whose nature is evil, he is mischievous, a liar, and evil. Man himself is not beautiful by nature, but he loves the beauty of people's nature. We need hard work and pursuit to form nature; otherwise, it is impossible to defeat the devil.]

Our System Made Summary

দুঃস্বভাবের মানুষ মানুষের রূপ দেখে মুগ্ধ হয় এবং তার ফল ভোগ করে। যার স্বভাব, তার স্পর্শ, তার রীতিনীতিকে মানুষ ঘৃণা করে। স্বভাব গঠনে কঠিন পরিশ্রম ও সাধনা চাই, নইলে শয়তানকে পরাজিত করা সম্ভব নয়। তার সুন্দর মুখ দেখে আনন্দিত হয়ো না।

[Evil people are fascinated by human form and enjoy its fruits. People hate his nature, his touch, his customs. We need hard work and pursuit to form the nature, otherwise it is not possible to defeat the devil. Don't be happy to see the beautiful faces.]

Human Made Summary

বাহ্যিক সৌন্দর্য নয়, স্বভাবের সৌন্দর্যই মানুষকে বিচারের মাপকাঠি। খারাপ স্বভাবের মানুষও বাহ্যিক সৌন্দর্যের অধিকারী হতে পারে। আর যারা খারাপ স্বভাবের তারাও সুন্দর স্বভাবের মানুষকে পছন্দ করে। তাই কঠোর পরিশ্রম ও সাধনার মাধ্যমে সুন্দর স্বভাবের অধিকারী হতে হবে।

[The beauty of nature, not external beauty, is the measure of human judgment. People with bad tempers can also have external beauty. And those who are bad in nature also like people who are good in nature. So you have to have a beautiful nature through hard work and pursuit.]

Dataset

Dataset

NCTB

Created a set of 139 samples of human-written **abstractive** document-summary pairs written by professional summary writers of the **National Curriculum and Textbook Board** (NCTB) [3]

BNLPC

Experiment with an **Extractive** Dataset
Bangla Natural Language Processing Community [4]

	NCTB	BNLPC
Total #Sample	139	200
Source Length (Avg)	91.33	150.75
Human Reference Length (Avg)	36.23	67.06
Summary Copy Rate	27%	99%

Statistics of the datasets

[3] https://w.wiki/ZwJ

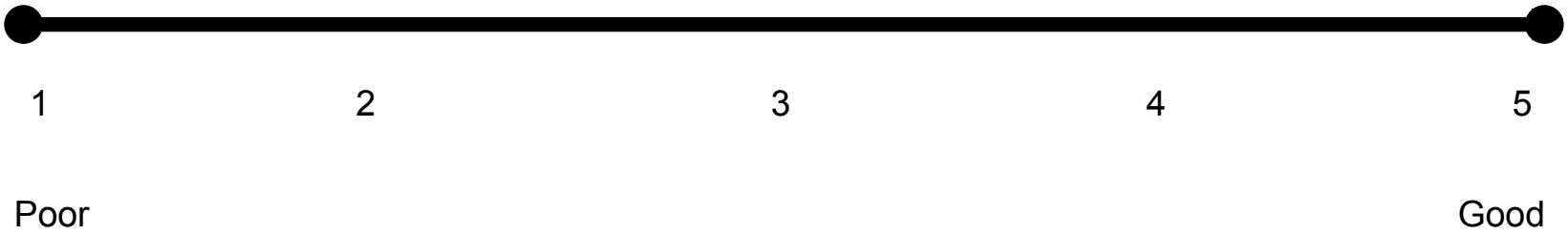
[4] http://www.bnlpc.org/research.php

Evaluation

Human Evaluation

Average Score

Content : 4.41
Readability : 3.95
Overall quality : 4.2



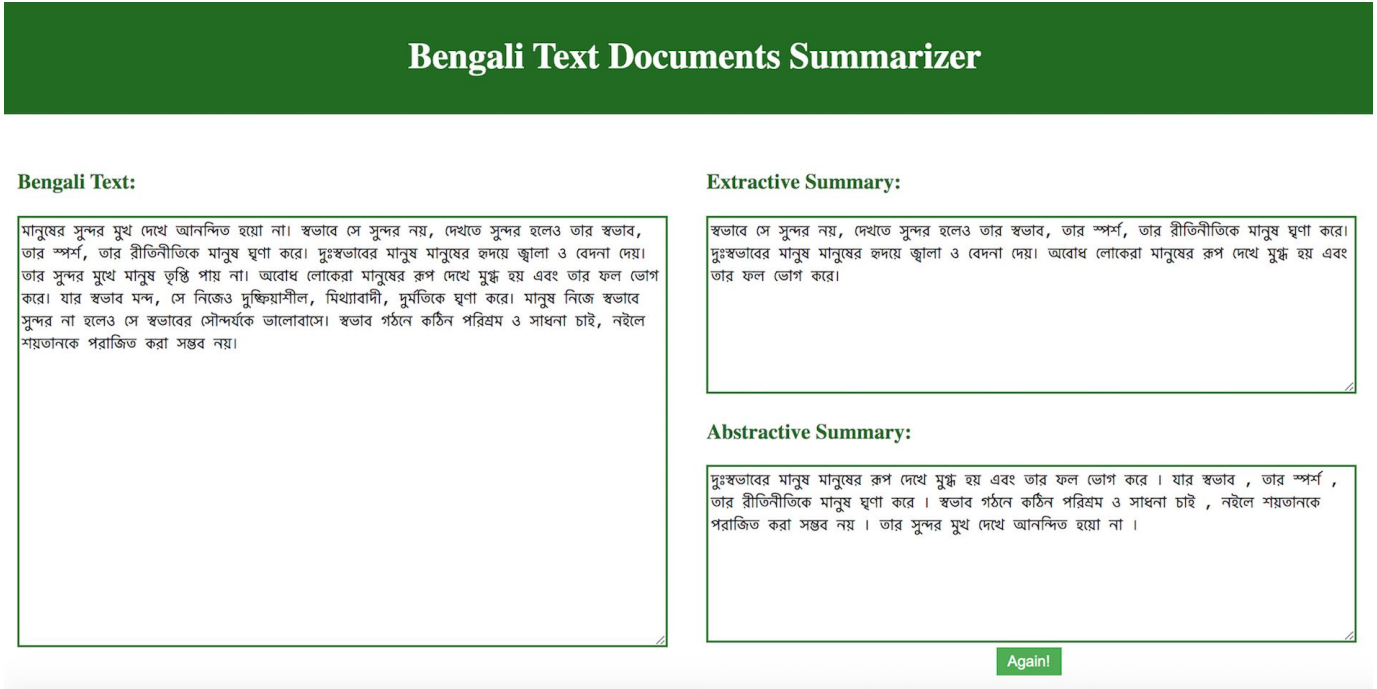
Automatic Evaluation

NCTB (Abstractive)	R-1	R-2	R-L
Random Baseline	9.43	1.45	9.08
GreedyKL	10.01	1.84	9.46
LexRank	10.65	1.78	10.04
TextRank	10.69	1.62	9.98
SumBasic	10.57	1.85	10.09
BenSumm[Abs] (ours)	**12.17**	**1.92**	**11.35**

BNLPC (Extractive)	R-1	R-2	R-L
Random Baseline	35.57	28.56	35.04
GreedyKL	48.85	43.80	48.55
LexRank	45.73	39.37	45.17
TextRank	60.81	**56.46**	60.58
SumBasic	35.51	26.58	34.72
BenSumm[Abs] (ours)	**61.62**	55.97	**61.09**

Results on our NCTB Dataset and BNLPC

Bengali Document Summarization Tool



Demo Video: https://youtu.be/LrnskktiXcg

Future Works

- Increasing document-summary pair dataset
- Implementing multi-sentence compression and paraphrasing
- Experiment with Multi-Document

Our code, data and all other resources:

https://github.com/tafseer-nayeem/BengaliSummarization

THANK YOU!

Questions?

You can also mail us at

radiarayan.rrc@gmail.com || mir.nayeem@alumni.uleth.ca ||

tahsintasnimmim@gmail.com || saif.chowdhury1997@gmail.com ||

taufiquljannat@gmail.com