TEXT SUMMARIZATION

Authors

Aalokik Singh (aalokik21223@iiitd.ac.in) Harsh Yadav (harsh21049@iiitd.ac.in) Pratham Singhal (pratham21347@iiitd.ac.in) Vivek Jain (vivek21218@iiitd.ac.in)

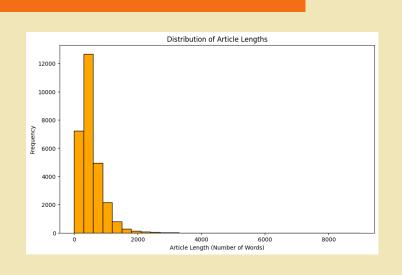
Affiliations



INDRAPRASTHA INSTITUTE of INFORMATION TECHNOLOGY **DELHI**

U P 3 4

DATASET VISUALISATION



PROBLEM STATEMENT

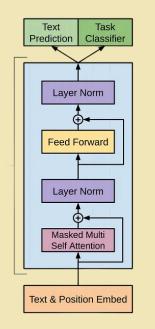
Our digital era's immense volume of text data necessitates efficient text summarization techniques. This project focuses on developing NLP-based methods for summarizing English text, addressing challenges like code-mixing and script diversity, to distill crucial information from extensive content into concise summaries.

OBJECTIVES

- Develop advanced text summarization techniques tailored for English content.
- Tackle challenges in summarizing codemixed English text for improved accuracy.
- Implement culturally adaptive strategies in English text summarization for diverse user needs.
- Enhance user relevance and domain adaptability in our text summarization model for real-world impact.

METHODOLOGY

The methodology in the code covers data preprocessing, train-validation splitting, dataset creation, tokenization using BART tokenizer, model initialization, training with AdamW optimizer, evaluation with ROUGE and BERTScore metrics, model saving, and results reporting including average ROUGE-L F1 score and BERTScore F1 score, as well as saving generated summaries for further analysis in a CSV file.



A

RESULTS

QUANTITATIVE RESULT

The quantitative evaluation of the model's performance was conducted using two key metrics: ROUGE-L F1 score and BERTScore F1 score. The test and validation scores are given below.

Test ROUGE-L (n=2) F1: 0.1872935903289672 Test BERTScore F1: 0.5102450847625732

Validation ROUGE-L (n=2) F1: 0.30781602391507734 Validation BERTScore F1: 0.5951529145240784 • The training loss steadily decreased over three epochs, indicating successful model convergence. The model was run for 3 epoch values as seen below.

epoch 0 running. Epoch 1/3, Train Loss: 4.5805 epoch 1 running. Epoch 2/3, Train Loss: 0.8238

Epoch 3/3, Train Loss: 0.3956

epoch 2 running.

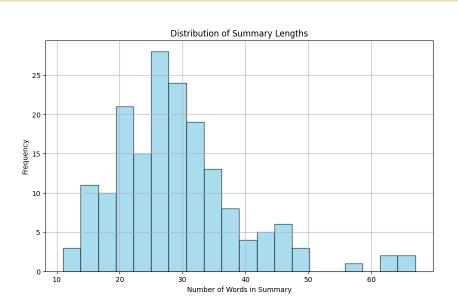
QUALITATIVE RESULT

	A	В	С	D	E	F
1	Article	Generated_Sum	imary			
2	Kriti Sanon is all set to take her entrepreneurial gan	Kriti Sanon is all	set to take her e	ntrepreneurial gar	me a notch highe	r as she launch
3	During a recent interview, actor Suniel Shetty voice	Suniel Shetty ha	s now said that h	is wife, Mana She	etty, prefers buyir	ng vegetables fo
4	Homegrown FMCG major Marico Ltd on Friday rep	Marico reported	a 15.64 per cent	rise in its consoli	dated net profit to	Rs 436 crore fo
5	Neeyat Movie Review: Neeyat revolves around exil	Neeyat Movie R	eview: The film is	a flawless and u	npredictable who	dunit in which ti
6	Women in sports are making huge strides! From pr	Women in sport	s are making hug	strides! From pr	ofessional leagu	es to college an
7	HOROSCOPE TODAY, JULY 30, 2023: Prepare to	Taurus, avoid er	igaging in unnece	ssary discussion	s, while Cancer b	e mindful of you
8	Twitter co-founder Jack Dorsey, who has also lost E	The app offers a	host of new feat	res and was initi	ally launched to t	he iOS users in
9	Mumbai Indians (MI) secured a six-wicket victory as	As the five-time	IPL champions of	itshone the hosts	, fans took to so	cial media to ex
10	Chennai Super Kings have reached the final of IPL	Chennai Super I	Kings have reach	ed the final of IPL	2023 for the rec	ord-extending 1
11	Kylian Mbappe continues to be at the epi-centre of	Kylian Mbappe	continues to be at	the epi-centre of	what promises to	be yet another
12	The principal of a private school in Pune district wa	A parent in her	complaint to police	alleged that the	school installed	a CCTV camera
13	Star Indian shuttlers PV Sindhu and Lakshya Sen n	Double Olympic	medallist Sindhu	seeded third, de	feated Korea's S	ung Shuo Yun 2
14	A Delhi Police Inspector was killed on Sunday morr	The accident to	k place in the ea	ly hours today or	Rohtak Road, n	ear Madipur me
15	Sanjay Dutt, who forayed into Bollywood with Rock	Sanjay Dutt, by	his own admissio	n, claimed to hav	e been with 308	women. Howeve
16	The Karnataka Primary and Secondary Education I	Karnataka Prima	ary and Secondar	y Education Minis	ster S Suresh Ku	mar said no dec
17	Niger's President Mohamed Bazoum has b	A group of soldi	ers, most of them	unknown to the p	oublic, claimed or	national TV to
18	Top-seeded American Taylor Fritz defeated compat	Taylor Fritz defe	ated compatriot J	J. Wolf 6-3, 7-5 t	o book a title cla	sh with Australia
19	At a time when the world needs new solutions, we f	Samsung has kr	nown this for a wh	ile. Samsung sta	rted the Solve fo	Tomorrow com
20	India's ambitious Chandrayaan-3 spacecraft lifted of	The budget for t	he department of	space (DoS), val	uation of space l	aunches and the

- Generated Summaries: The model produced concise and coherent summaries that captured key information from source articles.
- Word Clouds: Visualization of recurring themes and topics the model highlights across summaries.
- Length Comparison:
 Summaries were of
 appropriate length, balancing
 conciseness with
 informativeness.

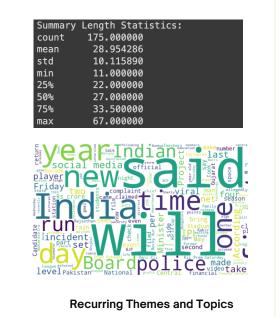
ANALYSIS

The training process showed significant improvement, with the model's loss decreasing from 4.58 to 0.39 over three epochs, indicating effective learning and convergence. Evaluation on validation and test datasets revealed higher scores on the validation set (ROUGE-L 0.30, BERTScore 0.59) compared to the test set (ROUGE-L 0.18, BERTScore 0.51), suggesting potential for further optimization. Qualitative analysis highlighted the model's capacity to distill relevant information, evident in word clouds showcasing recurring themes. A length comparison graph demonstrated the model's ability to generate concise yet informative summaries. These results emphasize the effectiveness of the text summarization techniques in capturing essential content from English text, with future improvements aimed at enhancing generalizability and adaptability to diverse user needs.



The summaries effectively balanced length and conciseness, with a maximum length of 67 words and an average length of 28.95 words, demonstrating their ability to distill key information efficiently while maintaining informativeness.

Training Loss 5 BERTScore Rogue-L Test Validation Epoch 2 Epoch 1 Epoch 3 0.0 0.1 0.6 0.2 0.3 0.5 **Training Loss vs Epoch** Validation and Test Metric Comparison



INTERESTING OBSERVATION

The effective balance achieved by the model in summary length, with a maximum of 67 words and an average of 28.95 words, showing model's capability to distill essential information into concise summaries while maintaining informativeness, even for long articles.

FUTURE WORK

Future work will focus on enhancing the model's adaptability to diverse text genres and languages to broaden its applicability in multilingual and cross-domain settings.

CONCLUSION

In conclusion, this study successfully developed and evaluated NLP-based text summarization techniques for English content, demonstrating significant improvements in model performance and promising evaluation metrics. The qualitative analysis confirmed the model's ability to produce concise and informative summaries, highlighting its effectiveness in distilling key information. Future research directions will focus on enhancing the model's generalizability and adaptability for real-world applications.