END-TO-END OCR AND DOCUMENT SUMMARIZER

ABSTRACT

Document summarization condenses text while preserving key information. Translation abstracts summarize content in another language, ensuring clarity and accuracy. Both methods improve accessibility, comprehension, and cross-lingual understanding in various contexts.

DATA SOURCE

IndicNLP Corpus:-
A collection of datasets for multiple Indian languages, including parallel corpora for summarization and translation tasks.

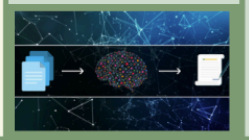
FLOW

OBJECTIVES

- CONDENSE INFORMATION – EXTRACT KEY POINTS WHILE RETAINING MEANING.
- ENHANCE ACCESSIBILITY – ENABLE CROSS-LANGUAGE UNDERSTANDING.
- ENSURE ACCURACY – MAINTAIN FACTUAL CORRECTNESS AND CONTEXT.

METHODS

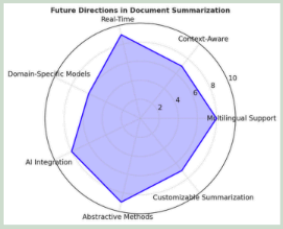
- OCR: TESSERACT, TROCR FOR TEXT EXTRACTION.
- SUMMARIZATION: TEXTRANK, BERTSUM, MBART, INDICBART.
- TRANSLATION: INDICTRANS, MARIANMT, GOOGLE API.



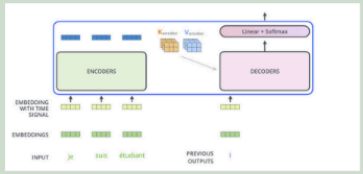
RESULTS

Test	Attribute	MIN	MAX	AVERAGE
Source text	Number of Sentence	11	94	25
	Number of Words	205	1527	463
Manual summary	Number of Sentence	3	31	7
	Number of Words	62	619	161
System summary	Number of Sentence	2	37	8
	Number of Words	61	699	168
Evaluation	Precision	0,38	0,88	0,54
	Recall	0,40	0,92	0,66
	F-measure	0,46	0,80	0,58

FUTURE DIRECTIONS



DATA VISULIZATION



MATHS BEHIND THE MODEL

$$\text{Attention}_i(Q_i, K_i, V_i) = \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i$$

$$PR(p_i; t+1) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j; t)}{L(p_j)}$$

CONCLUSION

- THE TOOL USES OCR (TROCR) AND SUMMARIZATION (MBART/INDICBART) TO EXTRACT AND CONDENSE TEXT FOR ENHANCED DIGITIZATION AND ACCESSIBILITY.

REFERENCES

- HTTPS://WWW.CANVA.COM
- HTTPS://COLAB.RESEARCH.GOOGLE.COM/
- KAGGLE.COM
- HTTPS://SCHOLAR.GOOGLE.COM
- HTTPS://PAPERSWITHCODE.COM/

GROUP MEMBERS

1. HRISHIT MADHAVI 1032220164
2. YUVRAJ KHAMKAR 1032220251
3. JACOB CHERIAN 1032220333
4. PRAJWAL AHER 1032220399
5. VISHWAJEET KAPALE 1032220296