HIGHLIGHTS: A NEWS TEXT SUMMARIZATION PROJECT

SHIVANI NAIK

WHAT IS IT?

Automatic News Summarization

On Spot Summary Generator

2 Types of Generators

Sentence Ranking

Fluent summary
that captures context

PROBLEM STATEMENT

Given a fresh news article as input, the system will generate a text summary of it. This project will involve exploring extractive and abstractive text summarization techniques.

WHY IS IT INTERESTING?

Spend minimal time in staying up to date with world

Gist of any news article provided by user

Expand to daily app that can be customized to user's news preference

Provides flexibility to user for detailed news reading

TEXT SUMMARIZATION TECHNIQUES

Extractive Text Summarization

- Identify most significant sentences from text and stack them to create summary
- Uses sentence ranking algorithms
- Time and Resource Efficient

Abstractive Text Summarization

- New words and phrases, different from source article
- Understands context and generates summary
- Usually complicated with deep learning techniques
- Time and Resource Intensive

DATA

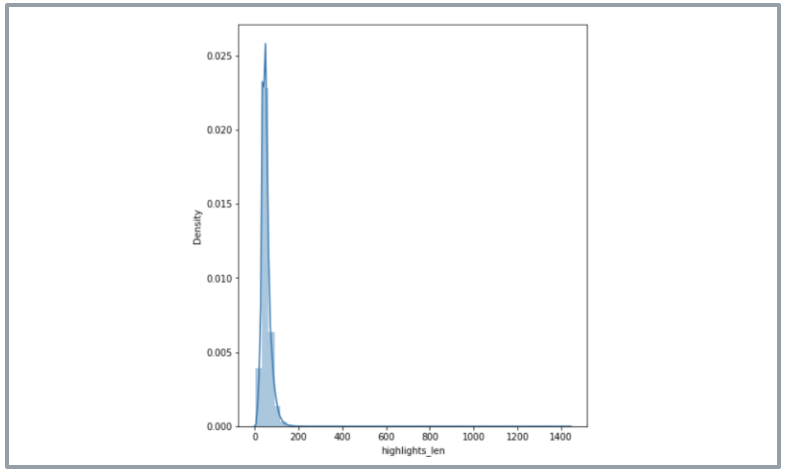
- CNN/DailyMail News Dataset
- Originally collected for question answering system
- Over 300K News Stories
- Each article: Body and Human generated text summary
- CNN: April 2007 to April 2015
- DailyMail: June 2010 and April 2015
- Covers various areas of news

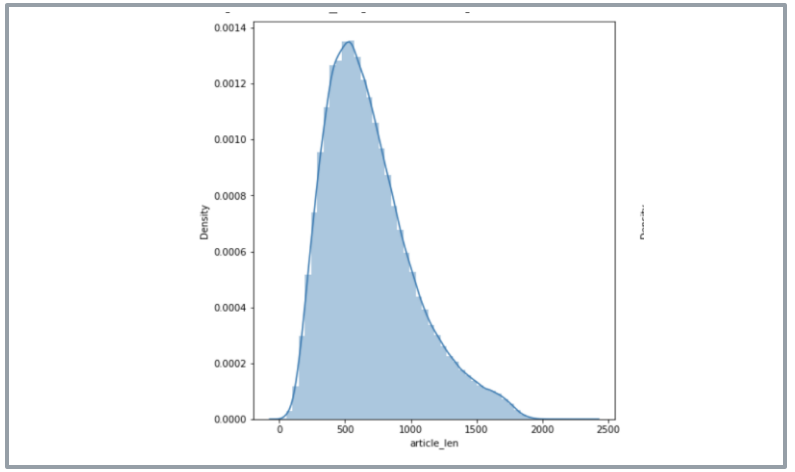
DATA EXPLORATION

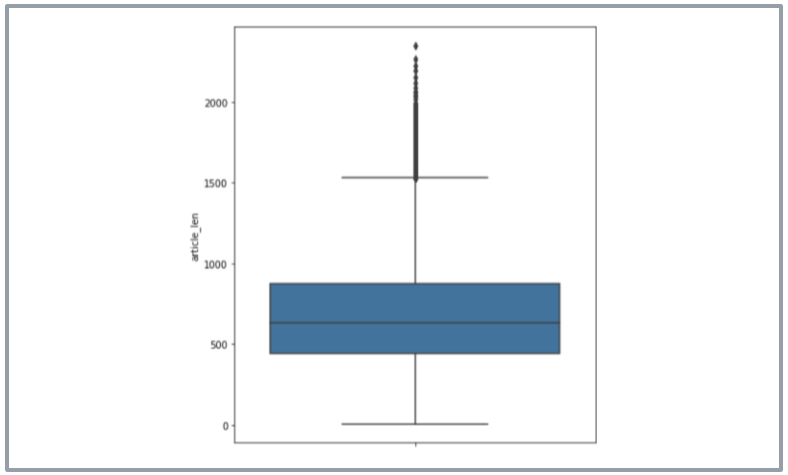
Article Length

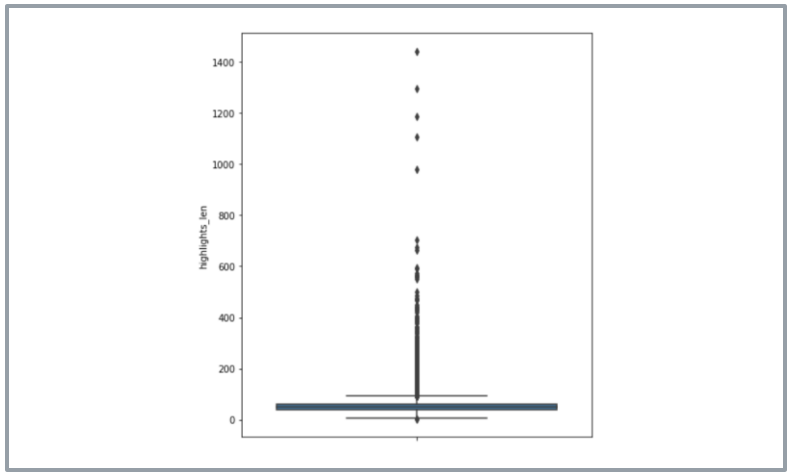
Summary Length

Right Skewed



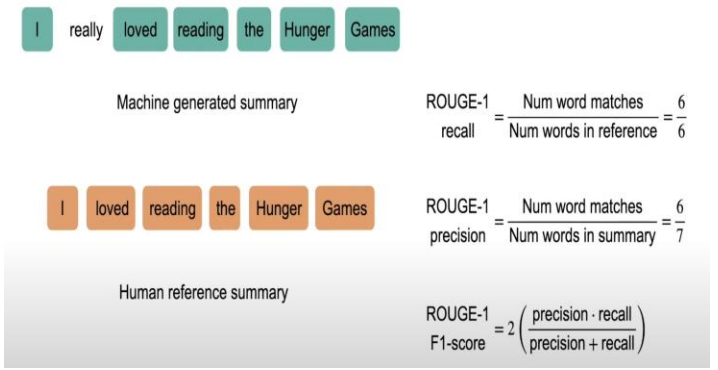


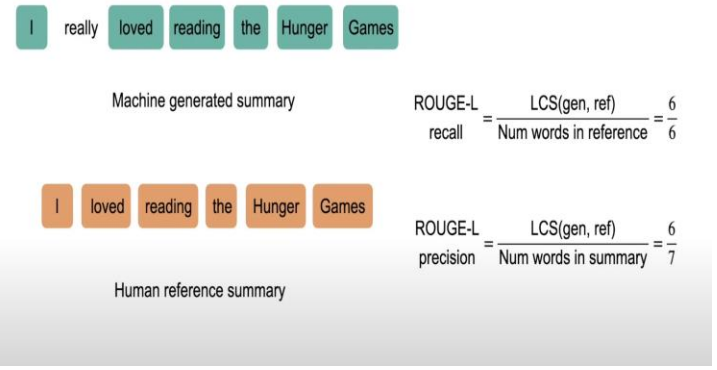




EVALUATION METRIC

- Complex to measure text summarization performance
- ROUGE score (Recall-Oriented Understudy for Gisting Evaluation)
- ROUGE-N computes matching N-grams between original summary and model summary
- Calculates Recall, Precision and F1-Score
- Synonyms are not taken care of, may assign low ROUGE score

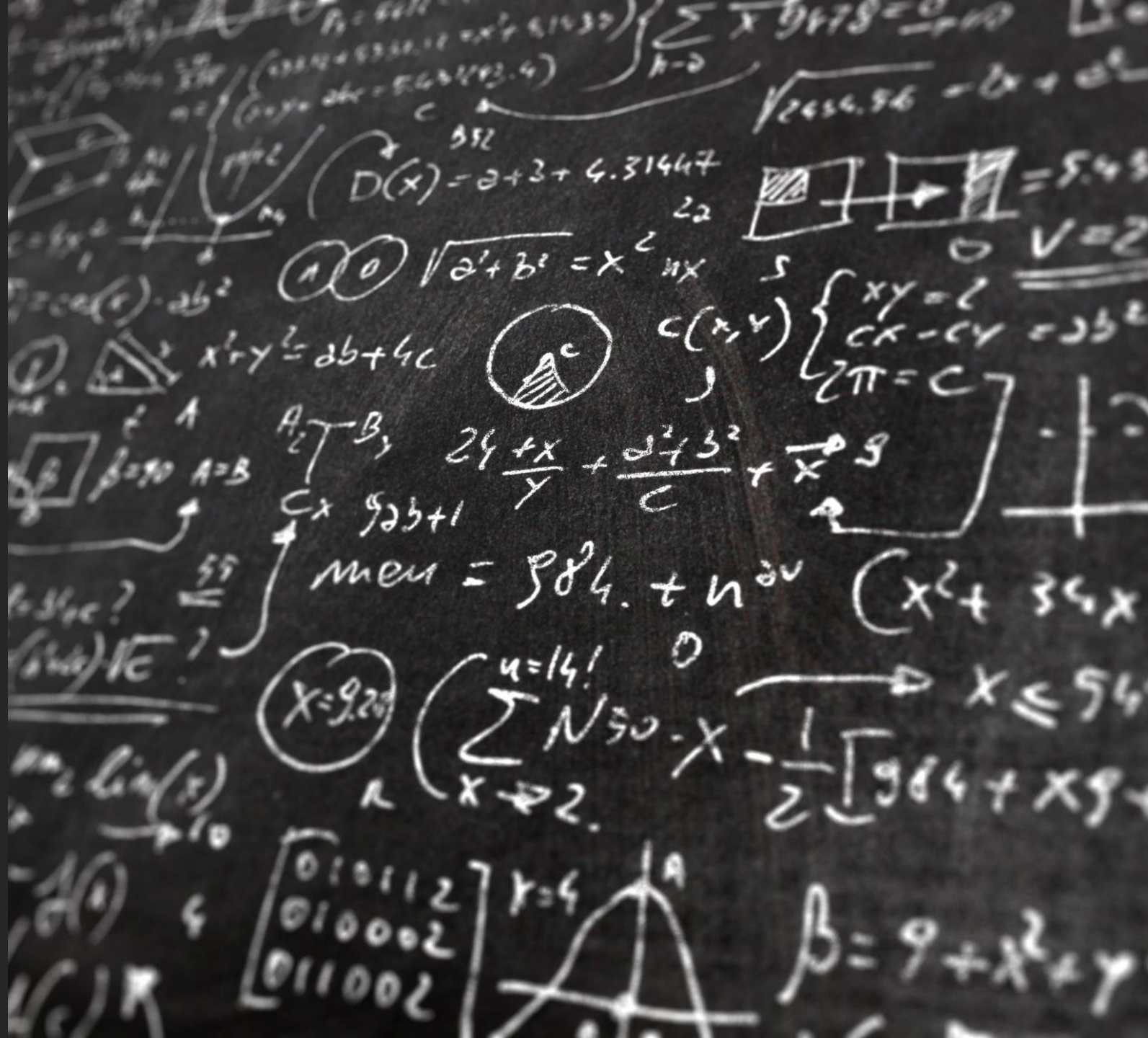




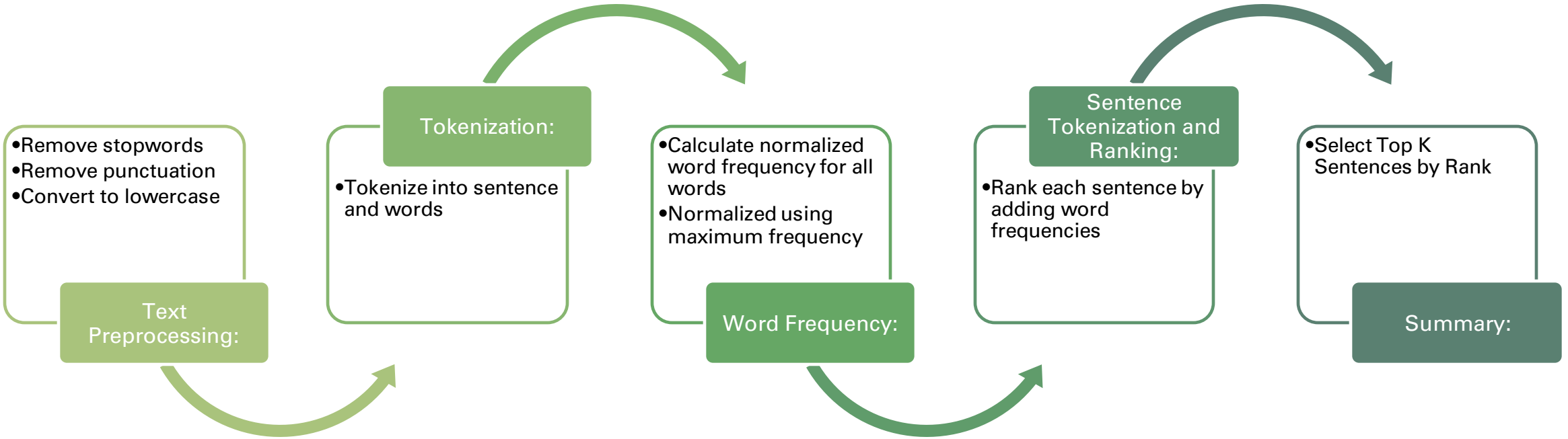
METHODS

Extractive Text Summarization

Abstractive Text Summarization



EXTRACTIVE SUMMARIZATION



EXTRACTIVE SUMMARIZATION EXAMPLE

association between high consumption of ultra-processed foods and cognitive decline, especially memory and executive function. Natalia Goncalves, PhD, with the University of Sao Paulo Medical School, presented the findings.

"High consumption" in the study was classified as more than 20% of daily caloric intake— meaning 400 calories for an active woman, whose recommended daily calorie intake is 2,000, or 500 calories for an active man, whose recommended daily calorie intake is 2,500.
While these findings may not cause a massive sea change in the advice offered to Alzheimer's disease and dementia patients on nutrition, it affirms already-existing knowledge: What's good for the heart is good for the brain.

"We know that a healthy diet, a heart-healthy diet full of fruits and vegetables, we know that it is

Extractive Summary

"High consumption" in the study was classified as more than 20% of daily caloric intake— meaning 400 calories for an active woman, whose recommended daily calorie intake is 2,000, or 500 calories for an active man, whose recommended daily calorie intake is 2,500. The not-yet-peer-reviewed study, which looked at 10,775 people in Brazil over 8 years, found an association between high consumption of ultra-processed foods and cognitive decline, especially memory and executive function. While these findings may not cause a massive sea change in the advice offered to Alzheimer's disease and dementia patients on nutrition, it affirms already-existing knowledge: What's good for the heart is good for the brain.

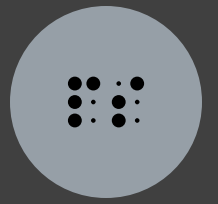
EXTRACTIVE SUMMARIZATION ROUGE

Metric	Value
ROUGE-1	28.59
ROUGE-2	11.93
ROUGE-L	17.54

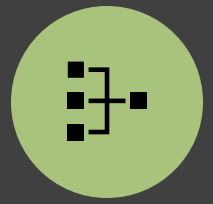
ABSTRACTIVE SUMMARIZATION

Understands context and rewrites summary

Closer to human summary generation process



Complex, difficult to generate fluent sentences



Transformer sequence-to-sequence model

Time consuming

T5 TRANSFORMER

Text-To-Text Transfer Transformer

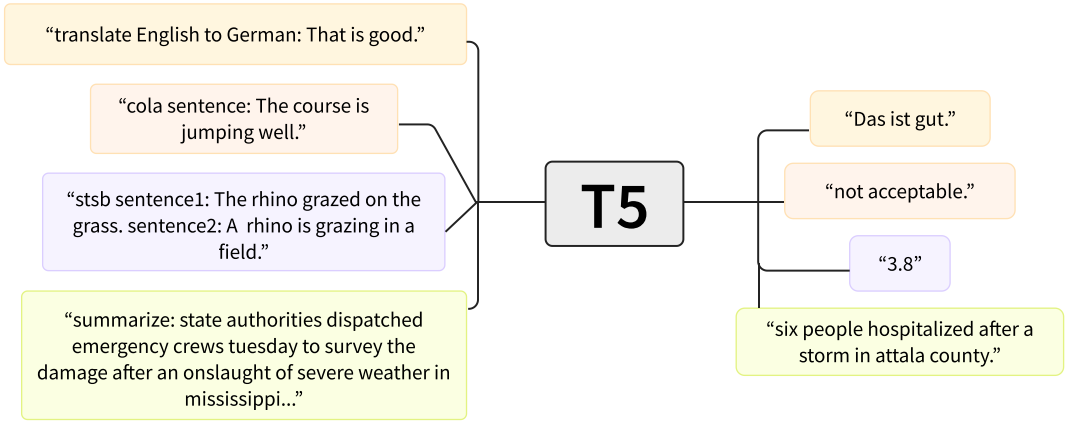
Use same model for different tasks like translation, summarization etc

Model can recognize which task: Task specific text prefix to input

Trained on C4 – Colossal Clean Crawled Corpus dataset

Training strategies

- Language Modeling
  - Next word prediction
- Deshuffling
  - Shuffle and predict original text
- Corrupting Spans
  - Mask words and predict them



HUGGINGFACE

PROVIDES ACCESS TO THOUSANDS OF PRETRAINED MODULES FOR FINE-TUNING

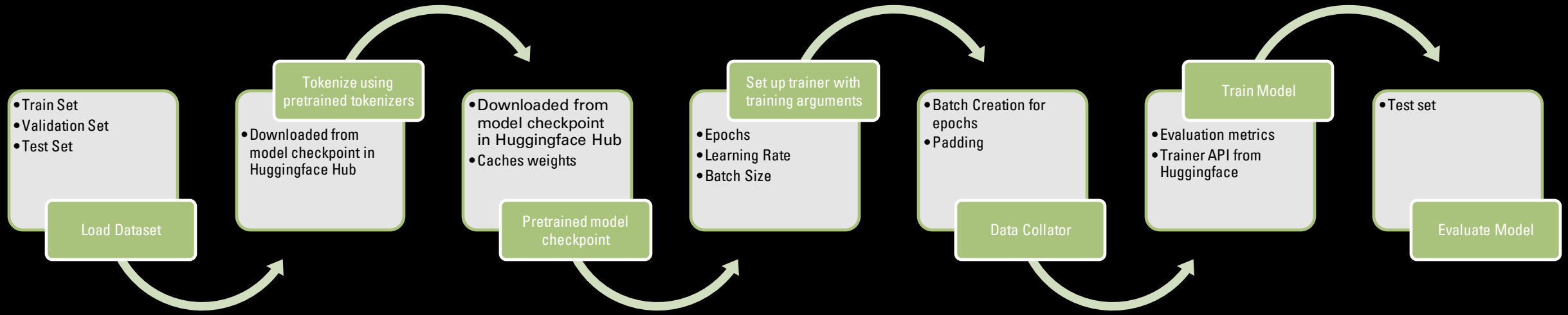
HUGGINGFACE HUB WITH MODELS TRAINED BY COMMUNITY ON MANY DATASETS

PUBLISH MODELS TO HUGGINGFACE HUB

TRAINER API FOR EFFICIENT MODEL TRAINING

REDUCE TRAINING TIME AND RESOURCES

HUGGINGFACE FINETUNING



ABSTRACTIVE SUMMARIZATION EXAMPLE

association between high consumption of ultra-processed foods and cognitive decline, especially memory and executive function. Natalia Goncalves, PhD, with the University of Sao Paulo Medical School, presented the findings.

"High consumption" in the study was classified as more than 20% of daily caloric intake— meaning 400 calories for an active woman, whose recommended daily calorie intake is 2,000, or 500 calories for an active man, whose recommended daily calorie intake is 2,500.
While these findings may not cause a massive sea change in the advice offered to Alzheimer's disease and dementia patients on nutrition, it affirms already-existing knowledge: What's good for the heart is good for the brain.

"We know that a healthy diet, a heart-healthy diet full of fruits and vegetables, we know that it is

Abstractive Summary

Study: An association between high consumption of ultra-processed foods and cognitive decline . Study looked at 10,775 people in Brazil over 8 years . "High consumption" in the study was classified as more than 20% of daily caloric intake . The study affirms already-existing knowledge: What's good for the heart is good for brain . 'We know that a healthy diet, a heart-healthy diet full of fruits and vegetables, we know that it is protective in many ways," says Alzheimer's expert .

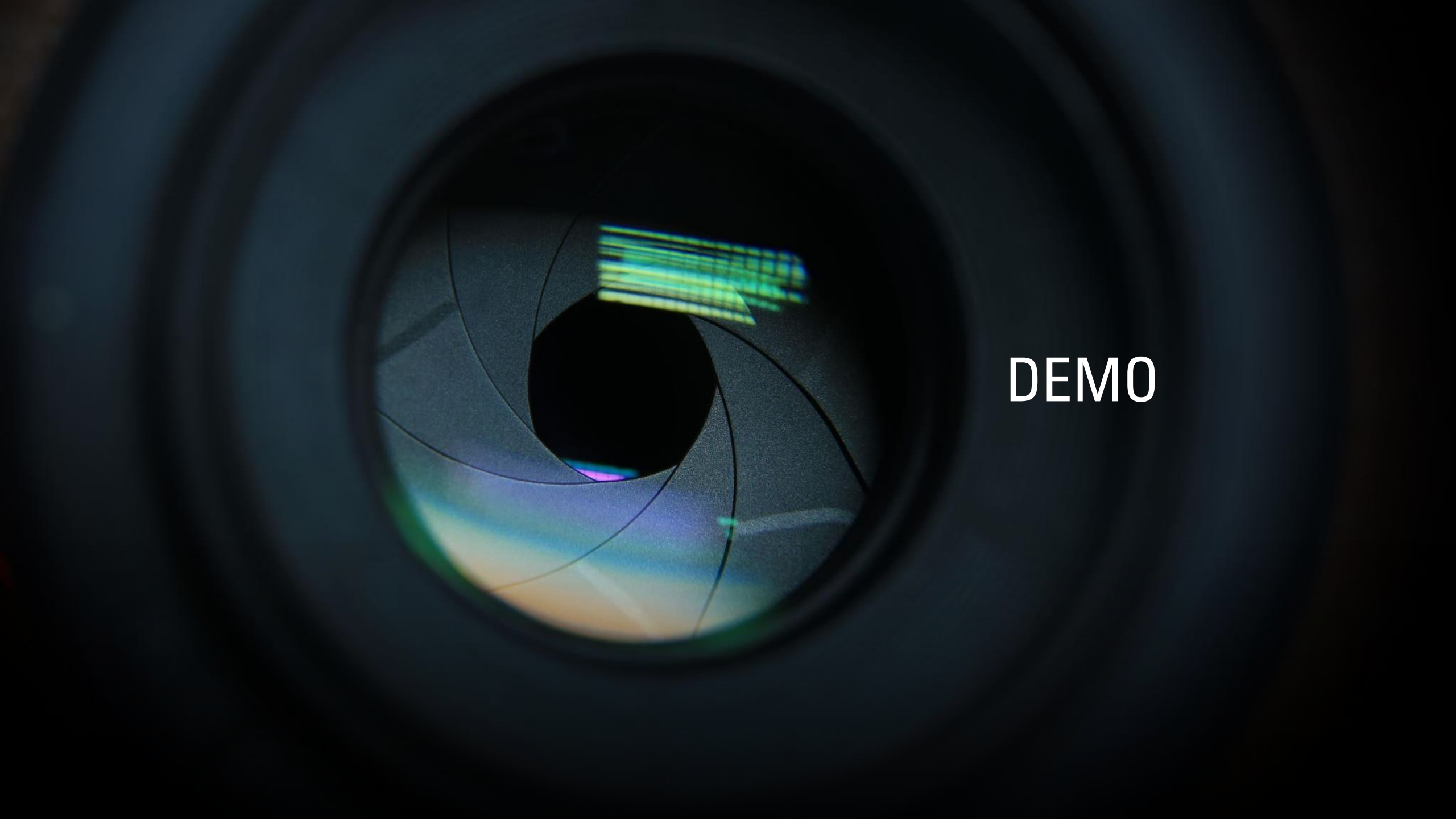
ABSTRACTIVE SUMMARIZATION ROUGE

Metric	Value
ROUGE-1	38.74
ROUGE-2	17.24
ROUGE-L	26.73

COMPARISON AND CONCLUSION

Extractive Summarization	Abstractive Summarization
• ROUGE 1 – 28.59 • ROUGE 2 – 11.93 • Faster • No heavy resources required	• ROUGE 1 – 38.74 • ROUGE 2 – 17.24 • Slower • GPU required

State-of-the-art model ROUGE 1 – 42.3, ROUGE 2 – 17.8


DEMO

FUTURE IMPROVEMENTS

Include more data for training (more resources)

Weekly Newsletter of Latest News Summaries

User's interest based recommended summaries

REFERENCES

- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020, July 28). *Exploring the limits of transfer learning with a unified text-to-text transformer*. arXiv.org. Retrieved August 4, 2022, from https://arxiv.org/abs/1910.10683
- Nallapati, R., Zhou, B., santos, C. N. dos, Gulcehre, C., & Xiang, B. (2016, August 26). *Abstractive text summarization using sequence-to-sequence RNNS and beyond*. arXiv.org. Retrieved August 4, 2022, from https://arxiv.org/abs/1602.06023v5
- Lin, C.-Y. (n.d.). *Rouge: A package for automatic evaluation of summaries*. ACL Anthology. Retrieved August 4, 2022, from https://aclanthology.org/W04-1013/
- community, T. H. F. D. (n.d.). *Cnn_dailymail · datasets at hugging face*. cnn_dailymail · Datasets at Hugging Face. Retrieved August 4, 2022, from https://huggingface.co/datasets/cnn_dailymail
- *Streamlit docs*. Streamlit documentation. (n.d.). Retrieved August 4, 2022, from https://docs.streamlit.io/