Text Summarization

Precision Précis

Kathan Bhavsar - AU2040170
Neel Buddhdev - AU2040176
Rushali Moteria - AU2040210
Twinkle Popat - AU2040216

Introduction

- Information available online has grown exponentially.
- Summarization: Reducing a large body of text into a shorter version by capturing the key points and main ideas
- Techniques of summarisation: extractive technique and the abstractive technique
- Manual summarization: time-consuming and subjective
- Automated approaches: more efficient and consistent

Problem Statement

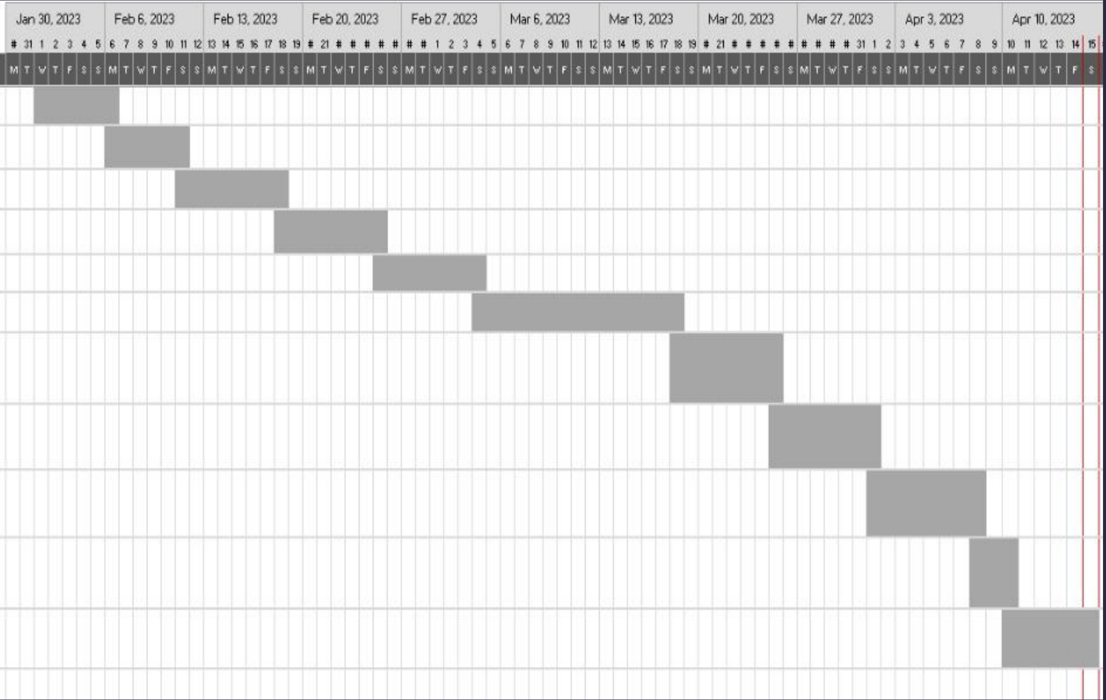
- The challenge lies in developing models that can accurately capture the most important information and select sentences that help to create a relevant summary while avoiding redundancy.

- Our project aims at addressing this problem by exploring various features and experimenting different SVM kernels to find the most efficient model for text summarisation.

GANTT Chart

1

TASK	PROGRESS	START	END
Selecting of problem and the dataset	100%	2-1-23	2-6-23
Tokenization and sentence scoring	100%	2-6-23	2-11-23
Performed Bag of words	100%	2-11-23	2-18-23
Research methods for summary	100%	2-18-23	2-25-23
TF-IDF to rank sentence	100%	2-25-23	3-4-23
SLA, count vectorizer ans SVM	100%	3-4-23	3-18-23
Resolved the dimension issue and analyzed the Mid Sem remarks	100%	3-18-23	3-25-23
Graph for Count Vectoriser vs. Sentence Length, Testing different kernels	100%	3-25-23	4-1-23
Creating functions for different features, and for forming the binary array	100%	4-1-23	4-8-23
Forming different models using different features, training and testing those model using different kernels	100%	4-8-23	4-10-23
Hyperparameter tuning of models with the highest accuracy	100%	4-10-23	4-15-23



Existing body of work

- The approach in the paper, “Extractive Text Summarization using Neural Networks” [4] is a fully data-driven approach that relies on neural networks for text summarization. Our model uses a algorithm that relies on feature engineering to identify important sentences for summarization.

- The neural network approach uses a feedforward neural network to identify important sentences for summarization.

- The neural network approach proposed in the paper is scalable and can produce summaries of arbitrarily sized documents.

Our Approach

1 Following features are considered

Entity Count (x1)	Entities can indicate key topics or events in the document.
TF-IDF (x2)	Measures the frequency of a term and its importance in the text.
Count Vectoriser (x3)	Measures the importance of words based on their frequency in text.
Sentence Length (x4)	Longer sentences may contain more information.

2 Creating Binary Array

Threshold =

$$\frac{Total\ number\ of\ entity\ words\ in\ the\ article}{Total\ number\ of\ words\ in\ the\ article}$$

Sentence Score =

$$\frac{Total\ number\ of\ entity\ words\ in\ the\ sentence}{Total\ number\ of\ words\ in\ the\ sentence}$$

If sentence score > threshold, assign “1”, else assign “0”.

Our Approach

3 Training models using different kernels

The table below shows accuracies of different models.

Models	RBF	Polynomial	Sigmoid
Model1 (x1,x2)	0.6267	0.6268	0.5293
Model2 (x1,x3)	0.6223	0.6301	0.5275
Model3 (x1,x4)	0.6244	0.6284	0.5399
Model4 (x2,x4)	0.6290	0.6277	0.5292
Model5 (x3,x4)	0.6250	0.6267	0.5329

Our Approach

4 Hyperparameter Tuning of models

- Selecting the two SVM models which give the maximum accuracy.
- “C” is the regularization parameter which controls the tradeoff between low training error and low testing error.
- Larger value of “gamma” corresponds to lower influence of the training sample and vice-versa.

Final Results

Report for Model1 RBF: (The original score for Model1 RBF for 1000 sentences was 0.55)

	precision	recall	f1-score	support
0	0.61	0.93	0.74	97
1	0.87	0.45	0.59	103
accuracy			0.68	200
macro avg	0.74	0.69	0.66	200
weighted avg	0.74	0.68	0.66	200

Final Results

Report for Model2 RBF: (The original score for Model2 RBF for 1000 sentences was 0.55)

	precision	recall	f1-score	support
0	0.69	0.76	0.72	95
1	0.76	0.70	0.73	105
accuracy			0.73	200
macro avg	0.73	0.73	0.72	200
weighted avg	0.73	0.72	0.73	200

Final Results

ROUGE Scores:

ROUGE-1	0.846
ROUGE-2	0.473

Sentence Lengths:

Original Article	61
Human Summary	12
Generated Summary	23

Conclusion

- Features such as County entity, TF-IDF, sentence length and count vectorizer play an important role in capturing the context of the text data.
- Use of these features can improve the accuracy of text summarization.
- Use of SVM with a radial basis function and polynomial kernel yielded the best performance, and hyperparameter tuning further improved the accuracy of the model.

Role of each group member

Kathan Bhavsar

Feature extraction methods and Output generation for extractive summarization

Twinkle Popat

Model Training And Hyperparameter tuning

Rushali Moteria

Model training and Hyperparameter tuning

Neel Buddhdev

Rough scores evaluation and Feature extraction methods

References

[1] "NLP based Machine Learning Approaches for Text Summarization," 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)

https://ieeexplore.ieee.org/abstract/document/9076358

[2] S. Zaware, D. Patadiya, A. Gaikwad, S. Gulhane and A. Thakare, "Text Summarization using TF-IDF and Textrank algorithm," 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)

https://ieeexplore.ieee.org/document/9453071

[3] A. W. Palliyali, M. A. Al-Khalifa, S. Farooq, J. Abinahed, A. Al-Ansari and A. Jaoua, "Comparative Study of Extractive Text Summarization Techniques," 2021 IEEE/ACS 18th International Conference on Computer Systems and Applications (AICCSA)

https://ieeexplore.ieee.org/document/9686867

[4] Sinha, A., Yadav, A., & Gahlot, A. (n.d.). "Extractive Text Summarization using Neural Networks."

https://arxiv.org/abs/1802.10137