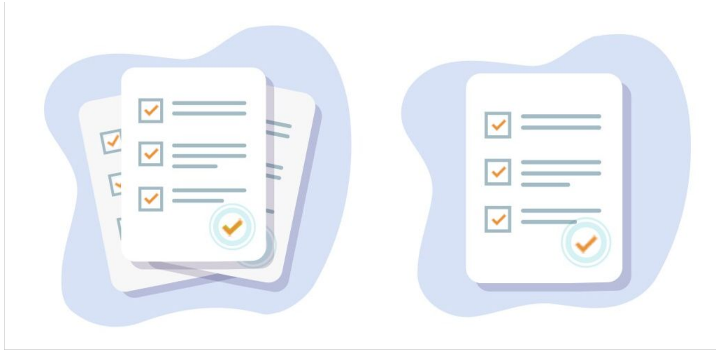
Towards Multi-document Summarisation in Low-resource Settings



Presented by:

Kushan Hewapathirana – 229333P
MSc in Computer Science (Part-time)

Supervisors:

1. Dr. Nisansa de Silva
2. Mr. C.D. Athuraliya

Outline

- Introduction
- Research Problem
- Literature Survey
- Benchmarking Baselines for MDS Datasets
- Enhancing MDS for The Medical Domain
- M2DS: Multilingual Dataset for MDS
- Adapter-based Opinion Summarisation
- Conclusions

Outline

- 💡 Introduction
- 🔍 Research Problem
- 📄 Publications
- 📚 Literature Survey
- 📊 Benchmarking Baselines for MDS Datasets
- 🩺 Enhancing MDS for The Medical Domain
- 🈳 M2DS: Multilingual Dataset for MDS
- 🧩 Adapter-based Opinion Summarisation
- 🎯 Conclusions

💡 Introduction

Introduction to Multi-document Summarisation

- MDS is the task of generating a single coherent summary from a set of related documents [1-2].
- It requires merging, deduplicating, and condensing information that may be complementary, redundant, or conflicting [1-3].
- Why is MDS Challenging?
  - Input documents often differ in style, length, and focus [1-2].
  - Must handle redundancy, contradiction, and incomplete overlap [3].
  - Summaries need to be faithful, informative, and well-structured [1].

[1] C. Ma, W. E. Zhang et al., "Multi-document summarization via deep learning techniques: A survey," ACM Computing Surveys (CSUR), 2020
[2] M. Afsharizadeh, H. Ebrahimpour-Komleh et al., "A survey on multi-document summarization and domain- oriented approaches," Journal of Information Systems and Telecommunication (JIST), vol. 1, no. 37, p. 68,2022
[3] P. J. Liu, M. Saleh, E. Pot, B. Goodrich, R. Sepassi, L. Kaiser, and N. Shazeer, "Generating wikipedia by summarizing long sequences," arXiv preprint arXiv:1801.10198, 2018

🔍 Research Problem

Research Gaps

- Limited Dataset Availability:
  - Most MDS datasets are **domain-specific**, **English-only**, and focused on **news** [9].
  - Especially lacking in multilingual and opinion-rich contexts [1, 10].

- Multilingual and Low-Resource Adaptation
  - Much of MDS research focuses on English, often using **translated** data [9].
  - Low-resource languages are largely ignored [9].

- Evaluation Across Diverse Domains:
  - Many models are tested on **only a few datasets** or **domains** [1, 10].
  - Generalisability remains an open challenge [1].

[1] P. J. Liu, M. Saleh, E. Pot, B. Goodrich, R. Sepassi, L. Kaiser, and N. Shazeer, “Generating wikipedia by summarizing long sequences,” arXiv preprint arXiv:1801.10198, 2018
[9] G. F. S. Eberhard, David M. and e. Charles D. Fennig, “Ethnologue: Languages of the americas and the pacific,” (No Title), 2023.
[10] R. Wolhandler, A. Cattan, O. Ernst, and I. Dagan, “How “multi” is multidocument summarization?” in, 2022, pp. 5761–5769. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing

Research Objectives

- Benchmark state-of-the-art MDS models across diverse datasets and languages
- Introduce the Multilingual Dataset for Multi-document Summarisation (M2DS), encompassing low-resource languages (R2).
- Introduce a novel, state-of-the-art adapter-based technique for opinion summarisation (R3).

	R1: Benchmark state-of-the-art MDS models across diverse datasets and languages	R2: Introduce the Multilingual Dataset for Multi-document Summarisation (M2DS), encompassing low-resource languages	R3: Introduce a novel, state-of-the-art adapter-based technique for opinion summarisation
Benchmarking Baselines for MDS Datasets (Phase I)	✓		
Medical Domain Adaptation (Phase II)	✓		
Multilingual Dataset for MDS: M2DS (Phase III)	✓	✓	
Adapter-based MDS (Phase IV)	✓		✓

📄 Publications

Publications

- **Multi-document Summarization: A Comparative Evaluation** - *IEEE ICIIS 2023, Sri Lanka* **(CORE C)**
  - Cited in '*MDCure: A Scalable Pipeline for Multi-Document Instruction-Following*' by Yale University, Google Research, and Allen Institute for AI

- **M2DS: Multilingual Dataset for Multi-document Summarisation** - *ICCCI 2024, Germany (Springer Nature Switzerland)* **(CORE B)**

- **Adapter-based Fine-tuning for PRIMERA** - *ADScAI 2025, Sri Lanka*


ADScAI

Adapter-based Fine-tuning for PRIMERA

Domain Adaptation for Multi-Document Summarisation: A Case Study in the Medical Research Domain

Multi-document Summarization: A Comparative Evaluation

Google Research

Ai2

Yale University

Publications (continued)

- **Domain Adaptation for Multi-document Summarisation: A Case Study in the Medical Research Domain -** *ICCCI 2025, Vietnam (Springer Nature Switzerland)* (Under review) **(CORE B)**

- **Adapter-based Multi-document Summarisation: Opinion Summarisation Use Case** - (In-progress Targeting *IJCNLP-AACL 2025*) **(CORE B)**

Adapter-based Multi-document Summarisation: Opinion Summarisation Use Case

Kushan Hewapathirana[1,2][0009-0008-1580-0699], Nisansa de Silva[1][0000-0002-5361-4810], and C.D. Athuraliya[2][0009-0007-4696-5210]

[1] Dept. of Computer Science & Engineering, University of Moratuwa, Sri Lanka
{kushan.22,nisansa}@cse.mrt.ac.lk
[2] ConscientAI, Sri Lanka
{kushan,cd}@conscient.ai

Abstract. This study explores the integration of adapter-based fine-tuning techniques to enhance the performance of PRIMERA model in the domain of opinion summarisation. PRIMERA, a state-of-the-art multi-document summarisation (MDS) model, exhibits impressive capabilities attributed to its pre-training on extensive MDS corpora. Leveraging adapter architectures, this research demonstrates significant performance improvements when PRIMERA is extended to the opinion summarisation domain by utilising state-of-the-art adapters. Hence, PRIMERA is positioned as a strong contender among opinion summarisation models, surpassing the state-of-the-art model, ADASUM. This work not only underscores the effectiveness of adapter methods but also pioneers a bridge
between the domains of MDS and opinio
gence offers potential solutions to data
summarisation while elevating MDS mo
sights and techniques from the opinion
directions entail further exploration of
tomized for opinion summarisation datas
and techniques tailored to the domain's
a promising avenue. Additionally, ongoin
tween MDS and opinion summarisation
sation models, fostering knowledge excha
opment across domains.

Keywords: Multi-document Summaris
cessing · Pre-trained Models · Opinion S

1 Introduction

In the contemporary digital landscape, onlin
ing consumer choices by providing comprehe
and services in the form of personal opinio
Amplayo et al. (2021) [3], act as valuable g
derstanding of both positive and negative as

Domain Adaptation for Multi-Document Summarisation: A Case Study in the Medical Research Domain

Kushan Hewapathirana[1,2][0009-0008-1580-0699], Nisansa de Silva[1][0000-0002-5361-4810], C.D. Athuraliya[2][0009-0007-4696-5210], and Piumi Kandanaarachchi[3][0009-0004-9592-9935]

[1] Dept. of Computer Science & Engineering, University of Moratuwa, Sri Lanka
{kushan.22,nisansa}@cse.mrt.ac.lk
[2] ConscientAI, Sri Lanka
{kushan,cd}@conscient.ai
[3] District General Hospital - Hambantota, Sri Lanka
piumikandanaarachchi@gmail.com

Abstract. Efficiently summarising a large volume of medical research papers is crucial for healthcare professionals to stay updated with the latest advancements in the field. Text summarisation, particularly abstractive text summarisation, a prominent and challenging task in natural language generation (NLG), aims to condense vast amounts of information into concise and coherent summaries. In this paper, we focus on the application of text summarisation in the medical domain and present our research on fine-tuning the PRIMERA multi-document summarisation (MDS) model specifically for medical research paper summarisation. By adapting PRIMERA to the medical domain using the MS^2 dataset, which comprises 20,000 reviews summarizing 470,000 medical research studies, we improve the model to generate accurate and concise summaries. Extensive fine-tuning experiments and human evaluations validate the effectiveness of our approach. Our fine-tuned PRIMERA model achieves a notable increase in ROUGE-1 scores from 12.8 to 33.0, ROUGE-2 scores from 2.0 to 6.5, and ROUGE-L scores from 8.1 to 22.6. Domain expert human evaluation further confirms improvements in relevance, coherence, and accuracy, demonstrating the model's ability to capture critical information and generate high-quality summaries tailored to the medical research domain.

Keywords: Multi-document Summarisation · Natural Language Processing · Medical Text Summarisation

1 Introduction

Multi-document summarisation is a challenging Natural Language Processing (NLP) task that aims to generate a summary by combining information from multiple sources. MDS involves handling conflicting, duplicate or complementary information to produce a summary that represents the overall content [12].

📚 Literature Survey

Model Evolution in MDS

- **RNN-based models:**
  - **Examples:** *R2N2, STDS, GRU-based encoder-decoder architecture, and RL-MMR [7-10].*
  - **Importance:** Ranking sentence importance, incorporating subtopic information, minimising diversity of opinions, including relevance measures.

- **CNN-based models:**
  - **Examples:** *PriorSum, HNet, DPP, MV-CNN, TCSum [11-14].*
  - **Importance:** Semantic and syntactic feature representation in MDS.

[7] Z. Cao, F. Wei, L. Dong, S. Li, and M. Zhou, "Ranking with recursive neural networks and its application to multi-document summarization," in Proceedings of the AAAI conference on artificial intelligence, vol. 29, no. 1, 2015.
[8] X. Zheng, A. Sun, J. Li, and K. Muthuswamy, "Subtopic-driven multi-document summarization," in EMNLP-IJCNLP, 2019, pp. 3153–3162
[9] A. Bražinskas, M. Lapata, and I. Titov, "Unsupervised opinion summarization as copycat-review generation," in ACL, Jul. 2020, pp. 5151–5169
[10] Mao, Y. Qu, Y. Xie, X. Ren, and J. Han, "Multi-document summarization with maximal marginal relevance-guided reinforcement learning," arXiv preprint arXiv:2010.00117, 2020
[11] Chen, "Convolutional neural network for sentence classification," Master's thesis, University of Waterloo, 2015
[12] Dos Santos and M. Gatti, "Deep convolutional neural networks for sentiment analysis of short texts," in Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers, 2014, pp. 69–78
[13] Z. Cao, F. Wei, S. Li, W. Li, M. Zhou, and H. Wang, "Learning summary prior representation for extractive summarization," in ACL, 2015, pp. 829–833
[14] H. Jin, T. Wang, and X. Wan, "Multi-granularity interaction network for extractive and abstractive multi-document summarization," in Proceedings of the 58th annual meeting of the association for computational linguistics, 2020, pp. 6244–6254

Model Evolution in MDS (continued)

- **Transformer-based models:**
  - **Examples:** *R2N2, STDS, GRU-based encoder-decoder architecture, and RL-MMR [14].*
  - **Importance:** Retain long-range dependencies and parallelisation advantage

- **Pre-trained language models:**
  - **Examples:** *BERTSUM [15], BART [16], T5 [17], PEGASUS [18], Longformer [19], BigBird [20], CDLM [21], PRIMERA [22], DAMEN [23]*
  - **Importance:** Broad contextual understanding of language, improved generalisability, fine-tuning efficiency

[14] H. Jin, T. Wang, and X. Wan, "Multi-granularity interaction network for extractive and abstractive multi-document summarization," in ACL, 2020, pp. 6244–6254
[15] Liu and M. Lapata, "Text summarization with pretrained encoders," arXiv preprint arXiv:1908.08345, 2019
[6] M. Lewis, Y. Liu, and others, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," arXiv preprint, 2019
[17] C. Raffel, and others, "Exploring the limits of transfer learning with a unified text-to-text transformer," The Journal of Machine Learning Research, vol. 21, 2020
[18] J. Zhang, and P. Liu, "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization," in International Conference on Machine Learning. PMLR, 2020
[19] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," arXiv preprint arXiv:2004.05150, 2020.
[20] M. Zaheer, G. Guruganesh, K. A. Dubey,,., "Big bird: Transformers for longer sequences," Advances in neural information processing systems, vol. 33, pp. 17 283–17 297, 2020
[21] A. Caciularu, A. Cohan, I. Beltagy, M. E. Peters, A. Cattan, and I. Dagan, "Cdlm: Cross-document language modeling," arXiv preprint arXiv:2101.00406, 202
[22] W. Xiao, I. Beltagy et al., "Primera: Pyramid-based masked sentence pre-training for multi-document summarization," in ACL, 2022, pp. 5245–5263
[23] G. Moro, and D. Freddi, "Discriminative marginalized probabilistic neural method for multi-document summarization of medical literature," in ACL 2022, pp. 180–189

Phase I:
📊 Benchmarking Baselines for MDS Datasets

Publications from this phase:
Multi-document Summarization: A Comparative Evaluation
IEEE ICIIS 2023, Sri Lanka (CORE C)

Motivation: Establishing a Benchmark

- Limited evaluation of SOTA MDS models across **diverse domains** and **document types** [1-3]
- Existing models struggle with **hybrid inputs** (mix of long and short documents) [1-3]
- Insufficient insight into performance on **complex** and **newer datasets**
  - Limits our ability to identify model weaknesses and guide future research [1,2]

⇒ Establish a strong benchmark by evaluating existing MDS models across a variety of datasets

[1] C. Ma, W. E. Zhang et al., "Multi-document summarization via deep learning techniques: A survey," ACM Computing Surveys (CSUR), 2020
[2] M. Afsharizadeh, H. Ebrahimpour-Komleh et al., "A survey on multi-document summarization and domain- oriented approaches," Journal of Information Systems and Telecommunication (JIST), vol. 1, no. 37, p. 68,2022
[3] A. M. Abid, "Multi-document text summarization using deep belief network," 2022.

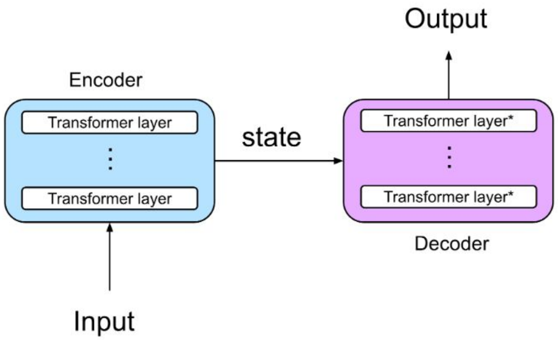
Current SOTA: Comparison of Evaluated Models

- PRIMERA [18]
  - Based on LED, enhanced with Entity Pyramid + Gap Sentence Generation (GSG)
  - Pre-training focuses on entity-rich sentence generation
  - Selects representative cluster sentences using Cluster ROUGE
  - Strong prior results in MDS across news and scientific datasets
- PEGASUS [22]
  - Pretrained using standard GSG objective
  - Focuses on sentence masking + reconstruction
  - Performs well in sentiment-heavy summarisation tasks (e.g., Rotten Tomatoes)
  - Slightly less robust on long or multi-source inputs

[18] W. Xiao, I. Beltagy et al., “Primera: Pyramid-based masked sentence pre-training for multi-document summarization,” in ACL, 2022, pp. 5245–5263
[22] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, “Pegasus: Pre-training with extracted gap-sentences for abstractive summarization,” in ICML. PMLR, 2020, pp. 11 328–11 339

Current SOTA: Comparison of Evaluated Models (continued)

- LED (Longformer Encoder-Decoder) [19]
  - Scalable transformer with sparse attention (local + global)
  - Often used as a baseline in long-input summarisation
  - Linear efficiency with respect to input length
  - Available in base and large variants (e.g., LED-base, LED-large)



[19] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," arXiv preprint arXiv:2004.05150, 2020.

Methodology: Overview of Datasets

Dataset	Total number of documents	Average number of documents per cluster	Domain
Multi-News [21]	56K [18]	3.5 [18]	News articles [21]
Multi-Xscience [22]	40K [18]	2.8 [18]	Related-work section in scientific articles [22]
Wikisum [23]	1.5M [18]	40 [18]	Wikipedia articles [23]
BigSurvey-MDS [24]	430K [18]	61.4 [18]	Human-written survey papers on various domains [24]
MS^2 [25]	470K [25]	23.5 [25]	Reviews of scientific publications in medical domain [25]
Rotten Tomato Dataset [10]	244K [10]	26.8 [10]	Movie reviews [10]

[10] S. Leon, “Rotten tomatoes movies and critic reviews dataset,” https://bit.ly/RTdataset, 2020, (Accessed on 06/24/2023)
[18] W. Xiao, I. Beltagy et al., “Primera: Pyramid-based masked sentence pre-training for multi-document summarization,” in ACL, 2022, pp. 5245–5263
[21] A. R. Fabbri, I. Li, T. She, S. Li, and D. R. Radev, “Multi-news: A large- scale multi-document summarization dataset and abstractive hierarchical model,”
[22] Lu, Y. Dong, and L. Charlin, “Multi-xscience: A large- scale dataset for extreme multi-document summarization of scientific articles,” in EMNLP, 2020, pp. 8068–8074
[23] P. J. Liu, M. Saleh, E. Pot, B. Goodrich, R. Sepassi, L. Kaiser, and N. Shazeer, “Generating wikipedia by summarizing long sequences,” arXiv preprint arXiv:1801.10198, 2018
[24] S. Liu, J. Cao, R. Yang, and Z. Wen, “Generating a structured summary of numerous academic papers: Dataset and method,” arXiv preprint arXiv:2302.04580, 2023.
[25] J. DeYoung, I. Beltagy, M. van Zuylen, B. Kuehl, and L. Wang, “MSˆ2: Multi-Document Summarization of Medical Studies,” in EMNLP, 2021, pp. 7494–7513.
[26] J. DeYoung, S. C. Martinez, I. J. Marshall, and B. C. Wallace, “Do multi- document summarization models synthesize?” arXiv preprint arXiv:2301.13844, 2023

Evaluation

Datasets	Metric	Models		
		PRIMERA [18]	PEGASUS [13]	LED [15]
Multi-News [21]	ROUGE-1	**42.0** [18]	32.0 [18]	17.3 [18]
	ROUGE-2	**13.6** [18]	10.1 [18]	3.7 [18]
	ROUGE-L	**20.8** [18]	16.7 [18]	10.4 [18]
Multi-XScience [22]	ROUGE-1	**29.1** [18]	27.6 [18]	14.6 [18]
	ROUGE-2	**4.6** [18]	**4.6** [18]	1.9 [18]
	ROUGE-L	**15.7** [18]	15.3 [18]	9.9 [18]
WikiSum [23]	ROUGE-1	**28.0** [18]	24.6 [18]	10.5 [18]
	ROUGE-2	**8.0** [18]	5.5 [18]	2.4 [18]
	ROUGE-L	**18.0** [18]	15.0 [18]	8.6 [18]
BigSurvey-MDS [24]	ROUGE-1	23.9	**38.9** [24]	39.8 [24]
	ROUGE-2	4.1	9.0 [24]	**9.4** [24]
	ROUGE-L	11.7	**16.2** [24]	16.1 [24]
MS^2 [25]	ROUGE-1	12.8	12.7	**25.8** [33]
	ROUGE-2	2.0	1.5	**8.4** [33]
	ROUGE-L	8.1	8.3	**19.3** [33]
Rotten Tomatoes Dataset [29]	ROUGE-1	25.4 [10]	**27.4** [10]	25.6 [10]
	ROUGE-2	8.4 [10]	**9.5** [10]	8.0 [10]
	ROUGE-L	19.8 [10]	**21.1** [10]	19.6 [10]

[10] J. DeYoung, S. C. Martinez, I. J. Marshall, and B. C. Wallace, "Do multi- document summarization models synthesize?" arXiv preprint arXiv:2301.13844, 2023
[18] W. Xiao, I. Beltagy et al., "Primera: Pyramid-based masked sentence pre-training for multi-document summarization," in ACL, 2022, pp. 5245–5263
[21] A. R. Fabbri, I. Li, T. She, S. Li, and D. R. Radev, "Multi-news: A large- scale multi-document summarization dataset and abstractive hierarchical model,"
[22] Lu, Y. Dong, and L. Charlin, "Multi-xscience: A large- scale dataset for extreme multi-document summarization of scientific articles," in EMNLP, 2020, pp. 8068–8074
[23] P. J. Liu, M. Saleh, E. Pot, B. Goodrich, R. Sepassi, L. Kaiser, and N. Shazeer, "Generating wikipedia by summarizing long sequences," arXiv preprint arXiv:1801.10198, 2018
[24] S. Liu, J. Cao, R. Yang, and Z. Wen, "Generating a structured summary of numerous academic papers: Dataset and method," arXiv preprint arXiv:2302.04580, 2023.
[25] J. DeYoung, I. Beltagy, M. van Zuylen, B. Kuehl, and L. Wang, "MSˆ2: Multi-Document Summarization of Medical Studies," in EMNLP, 2021, pp. 7494–7513.
[29] S. Leon, "Rotten tomatoes movies and critic reviews dataset," https://bit.ly/RTdataset, 2020, (Accessed on 06/24/2023)

Phase II:
Medical Domain Adaptation

Publications from this phase:
Domain Adaptation for Multi-document Summarisation:
A Case Study in the Medical Research Domain
ICCCI 2025, Germany (Springer Nature Switzerland)
(Under review) CORE B

Motivation: Adapting MDS to Medicine

- Medical research is vast and growing rapidly — staying up to date is challenging [1, 25]
- Summarising complex medical papers requires **accuracy, domain knowledge,** and **factual consistency** [1, 2, 25]
- Generic MDS models often struggle with **domain-specific terminology** and **structure** [25]

→

Fine-tune the PRIMERA model on the MS² dataset to generate reliable summaries for the medical domain.

[1] C. Ma, W. E. Zhang et al., "Multi-document summarization via deep learning techniques: A survey," ACM Computing Surveys (CSUR), 2020
[2] M. Afsharizadeh, H. Ebrahimpour-Komleh et al., "A survey on multi-document summarization and domain- oriented approaches," Journal of Information Systems and Telecommunication (JIST), vol. 1, no. 37, p. 68,2022
[25] J. DeYoung, I. Beltagy, M. van Zuylen, B. Kuehl, and L. Wang, "MSˆ2: Multi-Document Summarization of Medical Studies," in EMNLP, 2021, pp. 7494–7513.

Related Work: Domain-Specific and Scientific MDS

- CGSUM [35]
  - Citation-guided summarisation for scientific papers
  - Constructs a dataset (CSSC) and ranks documents using citation sentences
  - Summaries are generated from the top-ranked related papers

- DAMEN [23]
  - Designed for medical MDS
  - Uses two BERT-based modules: Indexer and Discriminator
  - Encodes background information to improve domain relevance

[23] Moro, G., Ragazzi, L., et al.: Discriminative marginalized probabilistic neural method for multi-document summarization of medical literature. In: ACL, pp. 180–189 (2022
[35] Chen, J., Cai, C., et al.: Comparative graph-based summarization of scientific papers guided by comparative citations. In: Proceedings of the 29th International Conference on Computational Linguistics, pp. 5978–5988 (2022)

Related Work: Domain-Specific and Scientific MDS (continued)

- Diverse Beam Search (DBS) [36]
  - Enhances output diversity during generation
  - Produces multiple summary candidates, selects best match to target attributes (e.g., sentiment range or mean)

[36] DeYoung, J., Martinez, S.C., et al.: Do multi-document summarization models synthesize? arXiv preprint arXiv:2301.13844 (2023)

Methodology: Model Selection

- Selection Criteria:
  - Literature review insights (Phase I)
  - ROUGE performance across datasets
  - Model recency and publication venue

- Selected Model: PRIMERA
  - Outperformed others across most datasets
  - Generalises well without heavy dataset-specific tuning
  - Uses Entity Pyramid + Gap Sentence Generation [1-2, 22]

    → Selects salient sentences more effectively than prior methods

[1] C. Ma, W. E. Zhang et al., “Multi-document summarization via deep learning techniques: A survey,” ACM Computing Surveys (CSUR), 2020
[2] M. Afsharizadeh, H. Ebrahimpour-Komleh et al., “A survey on multi-document summarization and domain- oriented approaches,” Journal of Information Systems and Telecommunication (JIST), vol. 1, no. 37, p. 68,2022
[22] W. Xiao, I. Beltagy et al., “Primera: Pyramid-based masked sentence pre-training for multi-document summarization,” in ACL, 2022, pp. 5245–5263

Methodology: MS² Dataset for Biomedical MDS

- Designed for biomedical multi-document summarisation
- Contains 20,000 systematic literature reviews
  - Each review summarises findings from multiple primary studies [37]

- Why MS² matters:
  - Reflects real-world synthesis in healthcare and clinical domains
  - Enables generation of concise, evidence-based summaries
  - Ensures high-quality content aligned with biomedical research standards [37]

[37] DeYoung, J., Beltagy, I., van Zuylen, M., Kuehl, B., Wang, L.L.: Ms^2:Multi-document summarization of medical studies. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 7494–7513 (2021)

Methodology: Evaluation Process

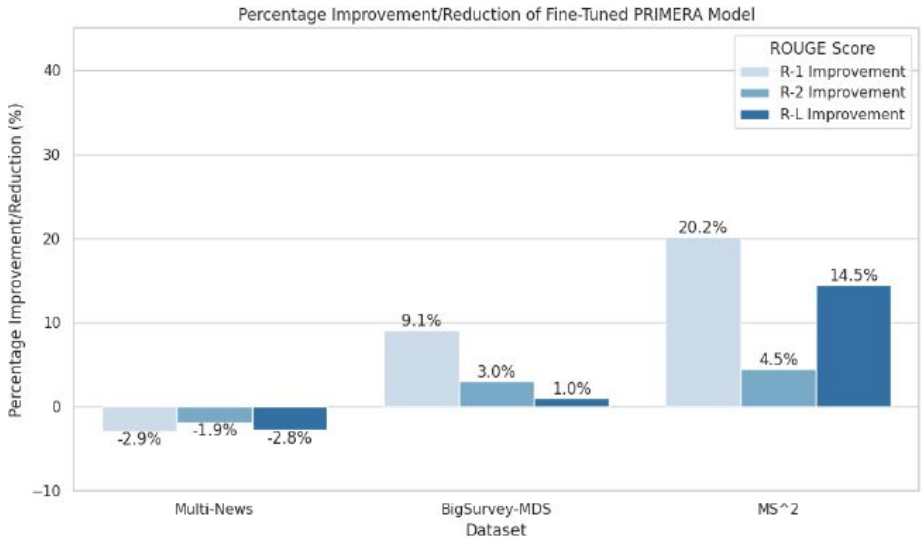
- Automated Evaluation [1-2, 28]:
  - **ROUGE-N**: Measures n-gram overlap with reference summaries
  - **ROUGE-L**: Captures sentence-level similarity using longest common subsequence

- Human Evaluation:
  - Assessed 50 summaries using 5 key criteria:
    → **Relevance, Coherence, Coverage, Conciseness, Accuracy**
  - 3 domain experts independently scored outputs
  - Cohen's Kappa used to measure inter-annotator agreement

[1] C. Ma, W. E. Zhang et al., "Multi-document summarization via deep learning techniques: A survey," ACM Computing Surveys (CSUR), 2020
[2] M. Afsharizadeh, H. Ebrahimpour-Komleh et al., "A survey on multi-document summarization and domain- oriented approaches," Journal of Information Systems and Telecommunication (JIST), vol. 1, no. 37, p. 68,2022
[28] [28] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in Text summarization branches out, 2004, pp. 74–81

Evaluation: Quantitative

Dataset		PRIMERA
Multi-News	R-1	39.1
	R-2	11.7
	R-L	18.0
BigSurvey-MDS	R-1	33.0
	R-2	7.1
	R-L	12.7
MS^2	R-1	33.0
	R-2	6.5
	R-L	22.6



Evaluation: Qualitative

Generated

Conclusions The results of this meta- analysis suggest that surgical ablation of the left atrial fibrillation (LAF) is associated with a significantly lower incidence of recurrent atrial arrhythmias compared with catheter ablation.However, there was no significant difference in the incidence of major adverse events between the two groups. Surgical ablation appears to be a safe and effective treatment strategy for the treatment of LAF.

Ground Truth

Subgroup analysis demonstrated similar trends , with higher freedom from AF in the surgical ablation group for paroxysmal AF patients . The incidence of pacemaker implantation was higher , while no difference in stroke or cardiac tamponade was demonstrated for the surgical versus catheter ablation groups . Current evidence suggests that epicardial ablative strategies are associated with higher freedom from AF , higher pacemaker implantation rates and comparable neurological complications and cardiac tamponade incidence to catheter ablative treatment .

Evaluation: Qualitative (continued)

Ground Truth Summary	Generated Summary
PDE5Is were significantly more effective than placebo or SSRIs for treating PE, while SSRIs were better than placebo. Combined treatment had better efficacy than PDE5Is alone.	There is no evidence to support the use of SSRIs for treating PE. There is insufficient evidence to determine if they are more effective than placebo or no treatment.
Silver-level evidence concerning the beneficial effects of mineral baths compared to no treatment. No clear effects for other balneological treatments were found.	Balneotherapy is an effective treatment for osteoarthritis of the knee in adults.
Oral cobalamin improves serum vitamin B12 and hematological parameters. Avoids discomfort and cost of injections. Supported for clinical use.	Oral cobalamin is effective for vitamin B12 deficiency in older adults. Insufficient evidence to support oral or intramuscular administration for B12 treatment.

Evaluation: Qualitative (continued)

	Relevance	Coherence	Coverage	Conciseness	Accuracy
Ground Truth	0.15	0.25	0.12	0.30	0.10
Generated Summary	-0.12	0.13	0.24	0.28	-0.10

Phase III:
🈳 Multilingual Dataset for MDS

Publications from this phase:
M2DS: Multilingual Dataset for Multi-document Summarisation
ICCCI 2024, Germany (Springer Nature Switzerland)
(CORE B)

Motivation: From English-only to Multilingual MDS

- Over **7,000 languages** exist, but most MDS research focuses on **English only** [30]
- Just ~**5% of the world** are native English speakers [30]
- Lack of **native multilingual MDS datasets** limits global applicability [30-31]

- First MDS dataset with native document-summary pairs in:
  - *English, Japanese, Korean, Tamil, Sinhala*
- Supports the development of low-resource and multilingual MDS models

[30] Eberhard, D.M., G.F.S., Fennig, C.D.: Ethnologue: languages of the Americas and the pacific (2023)
[31] Giannakopoulos, G.: Multi-document multilingual summarization and evaluation tracks in ACL 2013 multiling workshop. In: Proceedings of the Multiling 2013 Workshop on Multilingual Multi-document Summarization, pp. 20–28 (2013)

Related Work: Prior Work on Multilingual MDS

- MultiLing Workshop: Introduced one of the earliest multilingual MDS corpora [34-36]
  - **Languages:** Arabic, English, Greek, Chinese, Romanian, Czech, Hebrew, Spanish
  - **Approach: Machine Translated** English texts sentence-by-sentence

- Hyperplane-based Extractive Model [36]
  - Reformulated extractive MDS as a linear optimisation task
  - Extended classic vector space model to better handle multilinguality

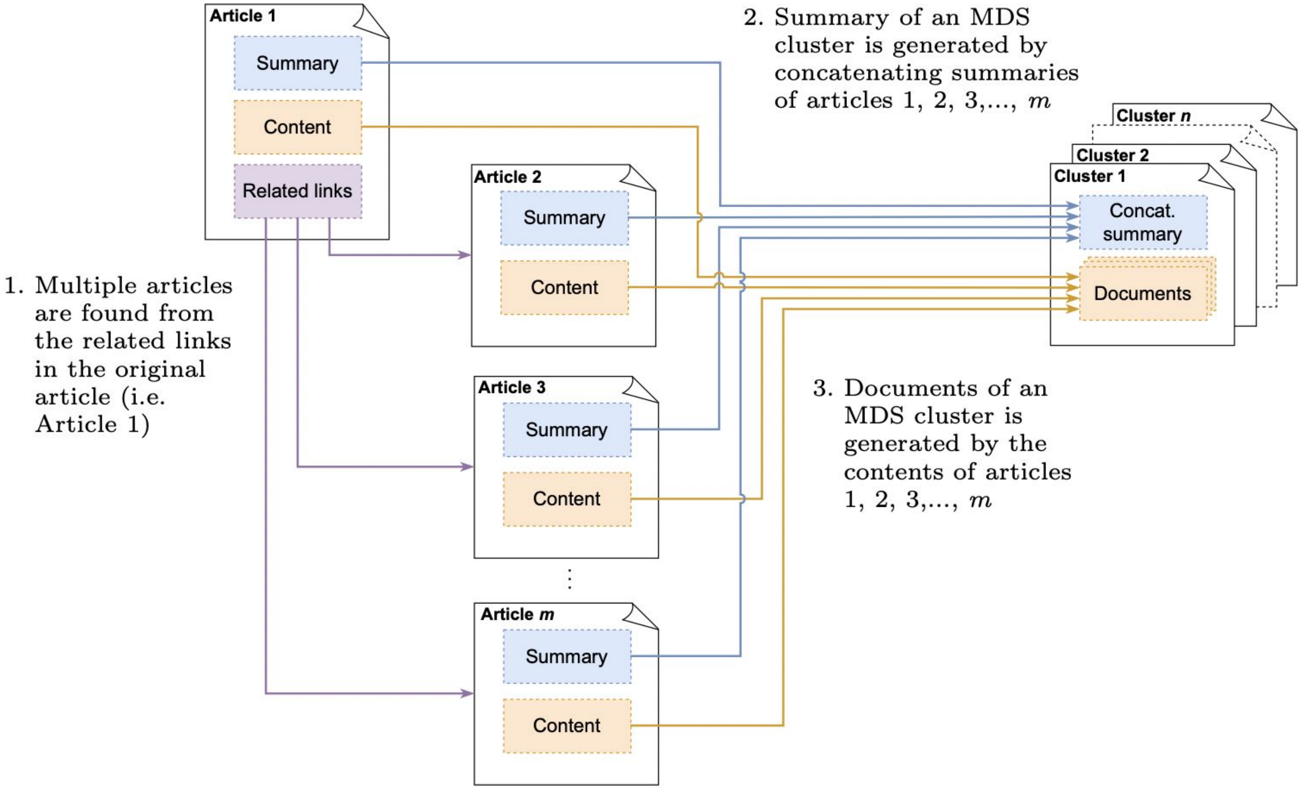
[34] G. Giannakopoulos, "Multi-document multilingual summarization and evaluation tracks in acl 2013 multiling workshop," in Proceedings of the multiling 2013 workshop on multilingual multi-document summarization, 2013, pp. 20–28.
[35] L. Li, C. For˘ascu, M. El-Haj, and G. Giannakopoulos, "Multi-document multilingual summarization corpus preparation, part 1: Arabic, english, greek, chinese, romanian," in Proceedings of the multiling 2013 workshop on multilingual multi-document summarization, 2013, pp. 1–12.
[36] L. Marina and V. Natalia, "Multilingual multi-document summarization with poly," in Proceedings of the MultiLing 2013 Workshop on Multilingual Multidocument Summarization, 2013.

Related Work: Prior Work on Multilingual MDS (continued)

- Multilingual Single-document Summarisation (SDS)
  - **MLSUM:** 1.5M news articles across several languages [38]
  - **XL-Sum:** 1.35M articles in 44 languages [39]
  - **WikiLingua:** Parallel SDS dataset with high language coverage [40]
  - **MLGSum / M3LS:** Large-scale multi-source or multi-modal multilingual datasets [41]
- **Cross-Lingual Summarisation (CLS)** [34, 39, 42]
  - Generates summaries in a different language than source documents
  - Grew alongside multilingual SDS efforts
  - CLS is now a common extension of multilingual SDS datasets

[34] G. Giannakopoulos, "Multi-document multilingual summarization and evaluation tracks in acl 2013 multiling workshop," in Proceedings of the multiling 2013 workshop on multilingual multi-document summarization, 2013, pp. 20–28.
[38] D. Scialom, Thomas et al., "Mlsum: The multilingual summarization corpus," in EMNLP, 2020, pp. 8051–8067.
[39] T. Hasan, Bhattacharjee et al., "Xl-sum: Large-scale multilingual abstractive summarization for 44 languages," in ACL-IJCNLP 2021, 2021, pp. 4693–4703.
[40] F. Ladhak, E. Durmus, C. Cardie, and K. Mckeown, "Wikilingua: A new benchmark dataset for cross-lingual abstractive summarization," in EMNLP 2020, 2020, pp. 4034–4048.
[41] D. Wang, J. Chen, H. Zhou, X. Qiu, and L. Li, "Contrastive aligned joint learning for multilingual summarization," in ACL-IJCNLP 2021, 2021, pp. 2739– 2750.
[42] Y. Verma, A. Jangra, R. Verma, and S. Saha, "Large scale multi-lingual multimodal summarization dataset," in ACL, 2023, pp. 3602–3614.

Methodology: Dataset Creation



Methodology: Comparison of M2DS with Other Datasets

Dataset	No. of documents	No. of clusters	Avg. no. of documents per cluster	Domain
Multi-News•	56.0k*	16.0k	3.5*	News articles•
Multi-Xscience◦	40.0k*	14.0k	2.8*	Related work section in scientific articles◦
Wikisum¦	1.5M*	37.5k	40.0*	Wikipedia articles¦
BigSurvey-MDS¢	430.0k*	7.0k	61.4*	Human-written survey papers on various domains¢
PEERSUM‖	11.9k‖	1.5k	7.8‖	Peer reviews of scientific publications
MSˆ2†	470.0k†	20.0k	23.5†	Reviews of scientific publications in medical domain†
Rotten Tomato Dataset↑	244.0k‡	9.0k	26.8‡	Movie reviews‡
M2DS	180.0k	51.5k	3.5	News articles
- English	67.0k	17.0k	3.9	
- Tamil	32.0k	10.0k	3.2	
- Japanese	29.0k	11.0k	2.6	
- Korean	27.0k	8.0k	3.4	
- Sinhala	23.5k	5.5k	4.2	

The sources are as follows: *Xiao et al.(2022) [41], †DeYoung et al.(2021) [6], ‡DeYoung et al.(2023) [7], •Fabbri et al.(2019) [10], ◦Lu et al.(2020) [27], ‖Li et al.(2022) [22], ¦Liu et al.(2018) [24], ¢Liu et al.(2023) [25], ↑Leon et al.(2020) [19].

Evaluation: Fine-tuned, for All Languages

Language		Models						
		LEAD-3	**RANDOM**	**CENTROID**	**PRIMERA**	**PEGASUS**	**LED**	**Llama 2**
Sinhala	**R-1**	0.06	5.7	4.5	5.7	4.1	3.6	**20.2**
	R-2	0.0	0.05	0.1	2.2	2.1	1.9	**6.5**
	R-L	0.06	5.1	3.9	3.2	2.8	2.9	**17.3**
Japanese	**R-1**	3.5	2.3	1.9	6.3	5.7	5.9	**7.7**
	R-2	0.0	0.01	0.05	**3.2**	1.3	1.4	0.8
	R-L	3.5	1.9	1.7	4.1	3.3	2.7	**6.8**
Korean	**R-1**	2.4	1.4	1.3	5.4	5.5	4.6	**8.5**
	R-2	0.4	0.02	0.03	1.1	**1.4**	0.8	1.0
	R-L	2.3	1.3	1.3	2.3	2.9	1.9	**8.1**
Tamil	**R-1**	6.8	1.6	2.2	4.4	3.8	3.7	**10.2**
	R-2	0.9	0.0	0.06	1.1	0.7	0.4	**3.1**
	R-L	6.2	1.6	1.9	2.2	1.7	1.3	**9.8**
English	**R-1**	1.2	6.4	7.6	**28.7**	22.5	20.5	20.8
	R-2	0.0	0.05	3.8	12.3	9.9	10.1	**13.5**
	R-L	1.1	5.7	7.6	17.1	14.7	15.2	**19.2**

The dataset dataset will be uploaded to https://huggingface.co/datasets/KushanH/m2ds and https://osf.io/7gjtm/files/osfstorage. The code and pre-trained models will be available on https://github.com/KushanMH/m2ds.

Evaluation: Zero-shot vs Fine-tuned (English)

Language		Models					
		PRIMERA	**PRIMERA** (fine-tuned)	**PEGASUS**	**PEGASUS** (fine-tuned)	**LED**	**LED** (fine-tuned)
English	**R-1**	23.6	**28.7**	18.6	22.5	17.1	20.5
	R-2	8.8	**12.3**	9.1	9.9	7.1	10.1
	R-L	13.6	**17.1**	12.4	14.7	13.2	15.2

Phase IV:
🧩 Adapter-based Opinion Summarisation

Publications from this phase:
- Adapter-based Multi-document Summarisation: Opinion Summarisation Use Case - (In-progress Targeting IJCNLP-AACL 2025) (CORE B)
- Adapter-based Fine-tuning for PRIMERA - ADScAI 2025, Sri Lanka

Motivation: Tackling Opinion Summarisation

- Online reviews offer valuable consumer insights — a mix of subjective opinions and factual info [7]

- Opinion summarisation datasets are rare and small
  - Annotation is costly and difficult due to subjective variation
  - Summaries must represent diverse perspectives across reviews [7-8, 43]

To overcome the challenge of scarce annotated data in opinion summarisation, we explore whether a powerful multi-document summariser can be adapted.

[7] A. Bražinskas, M. Lapata, and I. Titov, "Few-shot learning for opinion summarization," in EMNLP, 2020, pp. 4119–4135.
[8] N. Oved and R. Levy, "Pass: Perturb-and-select summarizer for product reviews," in ACL-IJCNLP, 2021, pp. 351–365.
[43] R. K. Amplayo and M. Lapata, "Informative and controllable opinion summarization," in EACL: Main Volume, 2021, pp. 2662–2672.

Opinion Summarisation

- Opinion summarisation is a specialised branch of MDS.
- It focuses on summarising diverse sets of reviews or opinion-rich texts about specific target entities, such as movies, products, or services [4-6].
- It provides concise representations of opinions, combining both sentiment analysis and summarisation tasks [5].
- Creating large-scale annotated datasets is challenging due to the need for summarising diverse products or services, each accompanied by numerous reviews [6].

[4] G. Di Fabbrizio, A. Stent, and R. Gaizauskas, “A hybrid approach to multi-document summarization of opinions in reviews,” in INLG, 2014, pp. 54–63.
[5] A. Bražinskas, M. Lapata, and I. Titov, “Unsupervised opinion summarization as copycat-review generation,” in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 5151–5169.
[6] A. Bražinskas, Nallapati et al., “Efficient few-shot fine-tuning for opinion summarization,” in NAACL, 2022, pp. 1509–1523

Related Work: Models in Opinion Summarisation

- LexRank [43]
  - Early extractive model using graph-based sentence centrality
  - Selects key sentences without training data
- MEANSUM [44]
  - Unsupervised abstractive model
  - Uses auto-encoders to generate summaries without domain-specific features
- COPYCAT [5]
  - Hierarchical latent variable model
  - Captures both product-level and review-level semantics for generation

[5] Bra˘zinskas, A., Lapata, M., Titov, I.: Unsupervised opinion summarization as copycat-review generation. In: ACL. pp. 5151–5169 (2020)
[43] Erkan, G., Radev, D.R.: Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of artificial intelligence research 22, 457–479 (2004)\
[44] Chu, E., Liu, P.: Meansum: A neural model for unsupervised multi-document abstractive summarization. In: ICML. pp. 1223–1232. PMLR (2019)[45] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in Text summarization branches out, 2004, pp. 74–81

Related Work: Models in Opinion Summarisation (continued)

- FEWSUM [7]
  - Few-shot learning approach
  - Uses lexical cues to align review content with summaries
- PASS [45]
  - Built on T5, fine-tuned using Amazon/Yelp gold summaries
  - Leverages transfer learning for better domain alignment
- ADASUM [6]
  - Current state-of-the-art
  - Based on BART with adapters for domain-specific tuning
  - Efficient and modular for low-resource opinion summarisation

[6] A. Bražinskas, Nallapati et al., “Efficient few-shot fine-tuning for opinion summarization,” in NAACL, 2022, pp. 1509–1523
[7] A. Bražinskas, M. Lapata, and I. Titov, “Few-shot learning for opinion summarization,” in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 4119–4135
[45] Oved, N., Levy, R.: Pass: Perturb-and-select summarizer for product reviews. In:ACL-IJCNLP. pp. 351–365 (2021)

Related Work: Adapters

- Bottleneck Adapters [46-47]
  - Insert compact “bottleneck” layers into transformer blocks
  - Enable efficient fine-tuning by training only a small number of parameters
  - Original model weights remain frozen
- AdapterFusion [48]
  - Combines multiple pre-trained adapters
  - Enables knowledge transfer and fusion across domains or tasks

[46] Bapna, A., Firat, O.: Simple, scalable adaptation for neural machine translation.In: EMNLP-IJCNLP. pp. 1538–1548 (2019)
[47] Houlsby, N., Giurgiu, et al.: Parameter-efficient transfer learning for nlp. In: ICML. pp. 2790–2799. PMLR (2019)[35] L. Li, C. Forˇascu, M. El-Haj, and G. Giannakopoulos, “Multi-document multilingual summarization corpus preparation, part 1: Arabic, english, greek, chinese, romanian,” in Proceedings of the multiling 2013 workshop on multilingual multi-document summarization, 2013, pp. 1–12.
[48] Pfeiffer, J., Kamath, et al.: Adapterfusion: Non-destructive task composition for transfer learning. In: ACL. pp. 487–503 (2021)

Related Work: Adapters (continued)

- MAD-X [49]
  - Adapter-based framework for multilingual transfer
  - Uses language-specific adapters for scalable cross-lingual adaptation
- LoRA (Low-Rank Adaptation) [50]
  - Injects low-rank matrices into attention layers
  - Achieves parameter-efficient fine-tuning with strong performance

[49] Pfeiffer, J., Vuli´c, I., Gurevych, et al.: Mad-x: An adapter-based framework for multi-task cross-lingual transfer. In: EMNLP. pp. 7654–7673 (2020)
[50] Hu, E.J., Wallis, et al.: Lora: Low-rank adaptation of large language models. In: ICLR (2021)

Methodology: Model Selection

- Selection Criteria:
  - Performance in Phase I (e.g., ROUGE scores)
  - Recency & impact (publication year and venue)
  - Generalisation capability across domains

- Selected Model: PRIMERA
  - Strongest cross-domain performance in Phase I
  - Built on LED, allowing efficient long-input handling [22]
  - Maintains document boundaries using special `<doc-sep>` tokens [22]
  - Minimises reliance on dataset-specific patterns — ideal for domain adaptation [1]
  - Uses Entity Pyramid + Gap Sentence Generation [1-2, 22]

    → Selects salient sentences more effectively than prior methods

[1] C. Ma, W. E. Zhang et al., "Multi-document summarization via deep learning techniques: A survey," ACM Computing Surveys (CSUR), 2020
[2] M. Afsharizadeh, H. Ebrahimpour-Komleh et al., "A survey on multi-document summarization and domain- oriented approaches," Journal of Information Systems and
[22] Telecommunication (JIST), vol. 1, no. 37, p. 68,2022
W. Xiao, I. Beltagy et al., "Primera: Pyramid-based masked sentence pre-training for multi-document summarization," in ACL, 2022, pp. 5245–5263

Methodology: Datasets

- Amazon Dataset [51]:
  - Categories:
    → Electronics, Clothing, Home & Kitchen, Health & Personal Care
  - Preprocessing steps:
    → Removed reviews <20 or >120 words
    → Balanced the number of review clusters per category

amazon

- Yelp Dataset [52]:
  - 300 summaries created via Amazon Mechanical Turk (MTurk)
  - Each summary combines ≈ 8 customer reviews
  - Designed for evaluating abstractive opinion summarisation quality

yelp

[51] He, R., McAuley, J.: Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In: proceedings of the 25th international conference on world wide web. pp. 507–517 (2016)
[52] Braˇzinskas, A., Nallapati, et al.: Efficient few-shot fine-tuning for opinion summarization. In: NAACL. pp. 1509–1523 (2022)

Methodology: Datasets (continued)

Split	Amazon [51]		Yelp [52]	
	Reviews	Golden summaries	Reviews	Golden summaries
Train	84	30	90	30
Test	36	12	90	30
Evaluation	60	23	120	40

[51] He, R., McAuley, J.: Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In: proceedings of the 25th international conference on world wide web. pp. 507–517 (2016)
[52] Braˇzinskas, A., Nallapati, et al.: Efficient few-shot fine-tuning for opinion summarization. In: NAACL. pp. 1509–1523 (2022)

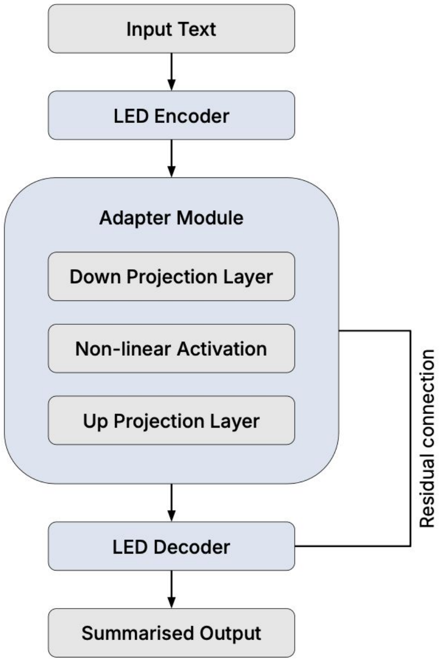
Methodology: Adapter Selection

- Conducted an empirical ablation study to assess the impact of different adapter architectures
- Integrated each adapter type into the PRIMERA model for evaluation on the opinion summarisation task
- Six adapters—Bottleneck adapter [46-47], AdapterFusion [48], Compacter [53], Invertible adapter [54], LoRA [50], and Mix-and-match adapter [55] —were selected.

[46] Bapna, A., Firat, O.: Simple, scalable adaptation for neural machine translation.In: EMNLP-IJCNLP. pp. 1538–1548 (2019)
[47] Houlsby, N., Giurgiu, et al.: Parameter-efficient transfer learning for nlp. In: ICML. pp. 2790–2799. PMLR (2019)[35] L. Li, C. Forˇascu, M. El-Haj, and G. Giannakopoulos, "Multi-document multilingual summarization corpus preparation, part 1: Arabic, english, greek, chinese, romanian," in Proceedings of the multiling 2013 workshop on multilingual multi-document summarization, 2013, pp. 1–12.
[48] Pfeiffer, J., Kamath, et al.: Adapterfusion: Non-destructive task composition for transfer learning. In: ACL. pp. 487–503 (2021)
[50] Hu, E.J., Wallis, et al.: Lora: Low-rank adaptation of large language models. In: ICLR (2021)
[53] Karimi Mahabadi, R., Henderson, J., Ruder, S.: Compacter: Efficient low-rank hypercomplex adapter layers. Advances in Neural Information Processing Systems 34, 1022–1035 (2021)
[54] Pfeiffer, J., Vuli´c, I., Gurevych, et al.: Mad-x: An adapter-based framework for multi-task cross-lingual transfer. In: EMNLP. pp. 7654–7673 (2020)
[55] He, J., Zhou, C., et al.: Towards a unified view of parameter-efficient transfer learning. In: ICLR (2021

Methodology: Adapter Selection (continued)

- Adapters chosen based on their architectural diversity
- Enables a comprehensive comparison of efficiency and performance under low-resource conditions [53-55]



[53] Karimi Mahabadi, R., Henderson, J., Ruder, S.: Compacter: Efficient low-rank hypercomplex adapter layers. Advances in Neural Information Processing Systems 34, 1022–1035 (2021)

[54] Pfeiffer, J., Vuli´c, I., Gurevych, et al.: Mad-x: An adapter-based framework for multi-task cross-lingual transfer. In: EMNLP. pp. 7654–7673 (2020)

[55] He, J., Zhou, C., et al.: Towards a unified view of parameter-efficient transfer learning. In: ICLR (2021

Evaluation: ROUGE Scores on Amazon and Yelp Datasets

Model	Amazon			Yelp		
	R1	**R2**	**RL**	**R1**	**R2**	**RL**
LEXRANK	27.72*	5.06*	17.04*	26.96*	4.93*	16.13*
MEANSUM	26.63*	4.89*	17.11*	27.50*	3.54*	16.09*
COPYCAT	27.85*	4.77*	18.86*	28.12*	5.89*	18.32*
FEWSUM	33.56*	7.16*	21.49*	37.29*	9.92*	22.76*
PASS	37.43*	8.02*	23.34*	36.91*	8.12*	23.09*
ADASUM	39.78 *	**10.80***	25.55*	38.82 *	11.75*	25.14*
PRIMERA	36.44	6.71	24.47	35.21	7.12	22.71
+ Bottleneck Adapter†	38.92	8.72	25.78	37.11	7.34	24.33
+ AdapterFusion‡	**40.11**	9.63	**25.61**	38.78	8.99	23.66
+ Compacter•	38.98	7.72	22.34	37.62	7.82	23.44
+ Invertible Adapter‖	39.07	9.11	23.45	**42.12**	**12.91**	**27.57**
+ LoRA¦	37.22	6.17	21.93	36.89	6.92	22.34
+ Mix-and-Match Adapter¢	38.76	8.34	25.16	39.49	10.23	26.66

The sources are as follows: ∗Braˇzinskas at el.(2022) [9], †Houlsby at el.(2019) [21], ‡Pfeiffe at el.(2021) [39], •Karimi at el.(2021) [23], ‖Pfeiffe at el.(2020) [40], ¦Hu at el.(2021) [22], ¢He at el.(2021) [18]

Evaluation: ROUGE Scores on Multi-news Dataset

Method	ROUGE-1	ROUGE-2	ROUGE-L
PRIMERA (Vanilla Model) [18]	42.0 [18]	13.6 [18]	20.8 [18]
PRIMERA (Fully Fine-Tuned) [18]	**49.9** [18]	**21.1** [18]	**25.9** [18]
Bottleneck Adapter [46]	47.1	17.4	21.6
AdapterFusion [48]	48.5	18.2	22.7

[18] W. Xiao, I. Beltagy et al., "Primera: Pyramid-based masked sentence pre-training for multi-document summarization," in ACL, 2022, pp. 5245–5263
[46] Bapna, A., Firat, O.: Simple, scalable adaptation for neural machine translation.In: EMNLP-IJCNLP. pp. 1538–1548 (2019)
[48] Pfeiffer, J., Kamath, et al.: Adapterfusion: Non-destructive task composition for transfer learning. In: ACL. pp. 487–503 (2021)

🎯 Conclusions

Summary of Contributions

- Benchmarked state-of-the-art MDS models across multiple domains, revealing limited generalisation outside their training contexts
- Fine-tuned a leading MDS model (PRIMERA) for the medical domain, achieving strong ROUGE scores and expert-verified improvements
- Developed an adapter-based approach for opinion summarisation, enabling effective use of tiny, low-resource datasets
- Introduced M2DS, the first native multilingual MDS dataset, supporting evaluation in five languages
- Contributed adapter support for LED/PRIMERA to the open-source `adapter-transformers` library, facilitating low-cost adaptation

References

[1] C. Ma, W. E. Zhang et al., “Multi-document summarization via deep learning techniques: A survey,” ACM Computing Surveys (CSUR), 2020
[2] M. Afsharizadeh, H. Ebrahimpour-Komleh et al., “A survey on multi-document summarization and domain- oriented approaches,” Journal of Information Systems and Telecommunication (JIST), vol. 1, no. 37, p. 68,2022
[3] A. M. Abid, “Multi-document text summarization using deep belief network,” 2022.
[4] C. Ma, W. E. Zhang et al., “Dependency structure for news document summarization,” arXiv e-prints, pp. arXiv–2109, 2021
[5] R. Wolhandler, A. Cattan et al., “How” multi” is multi-document summarization?” arXiv preprint arXiv:2210.12688, 2022.
[6] R. Pasunuru, M. Liu et al., “Efficiently summarizing text and graph encodings of multi-document clusters,” in ACL, 2021, pp. 4768–4779.
[7] Z. Cao, F. Wei, L. Dong, S. Li, and M. Zhou, “Ranking with recursive neural networks and its application to multi-document summarization,” in Proceedings of the AAAI conference on artificial intelligence, vol. 29, no. 1, 2015.
[8] X. Zheng, A. Sun, J. Li, and K. Muthuswamy, “Subtopic-driven multi-document summarization,” in EMNLP-IJCNLP, 2019, pp. 3153–3162
[9] A. Bražinskas, M. Lapata, and I. Titov, “Unsupervised opinion summarization as copycat-review generation,” in ACL, Jul. 2020, pp. 5151–5169
[10] Mao, Y. Qu, Y. Xie, X. Ren, and J. Han, “Multi-document summarization with maximal marginal relevance-guided reinforcement learning,” arXiv preprint arXiv:2010.00117, 2020
[11] Chen, “Convolutional neural network for sentence classification,” Master’s thesis, University of Waterloo, 2015
[12] Dos Santos and M. Gatti, “Deep convolutional neural networks for sentiment analysis of short texts,” in Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers, 2014, pp. 69–78
[13] Z. Cao, F. Wei, S. Li, W. Li, M. Zhou, and H. Wang, “Learning summary prior representation for extractive summarization,” in ACL, 2015, pp. 829–833
[14] H. Jin, T. Wang, and X. Wan, “Multi-granularity interaction network for extractive and abstractive multi-document summarization,” in Proceedings of the 58th annual meeting of the association for computational linguistics, 2020, pp. 6244–6254
[15] Liu and M. Lapata, “Text summarization with pretrained encoders,” arXiv preprint arXiv:1908.08345, 2019
[16] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, and others, “Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension,” arXiv preprint arXiv:1910.13461, 2019
[17] C. Raffel, N. Shazeer, A. Roberts, K. Lee, and others, “Exploring the limits of transfer learning with a unified text-to-text transformer,” The Journal of Machine Learning Research, vol. 21, no. 1, pp. 5485–5551, 2020
[18] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, “Pegasus: Pre-training with extracted gap-sentences for abstractive summarization,” in International Conference on Machine Learning. PMLR, 2020, pp. 11 328–11 339
[19] I. Beltagy, M. E. Peters, and A. Cohan, “Longformer: The long-document transformer,” arXiv preprint arXiv:2004.05150, 2020.
[20] M. Zaheer, G. Guruganesh, K. A. Dubey,,., “Big bird: Transformers for longer sequences,” Advances in neural information processing systems, vol. 33, pp. 17 283–17 297, 2020
[21] A. Caciularu, A. Cohan, I. Beltagy, M. E. Peters, A. Cattan, and I. Dagan, “Cdlm: Cross-document language modeling,” arXiv preprint arXiv:2101.00406, 202
[22] W. Xiao, I. Beltagy et al., “Primera: Pyramid-based masked sentence pre-training for multi-document summarization,” in ACL, 2022, pp. 5245–5263
[23] G. Moro, L. Ragazzi, L. Valgimigli, and D. Freddi, “Discriminative marginalized probabilistic neural method for multi-document summarization of medical literature,” in ACL 2022, pp. 180–189
[24] C.-Y. Lin, “Rouge: A package for automatic evaluation of summaries,” in Text summarization branches out, 2004, pp. 74–81
[25] Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: a Method for Automatic Evaluation of Machine Translation,” in Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318
[26] P. J. Liu, M. Saleh, E. Pot, B. Goodrich, R. Sepassi, L. Kaiser, and N. Shazeer, “Generating wikipedia by summarizing long sequences,” arXiv preprint arXiv:1801.10198, 2018

References

[27] A. R. Fabbri, I. Li, T. She, S. Li, and D. R. Radev, “Multi-news: A large- scale multi-document summarization dataset and abstractive hierarchical model,”
[28] M. Koupaee and W. Y. Wang, “Wikihow: A large scale text summarization dataset,” arXiv preprint arXiv:1810.09305, 2018
[29] J. DeYoung, S. C. Martinez, I. J. Marshall, and B. C. Wallace, “Do multi- document summarization models synthesize?” arXiv preprint arXiv:2301.13844, 2023
[30] S. Leon, “Rotten tomatoes movies and critic reviews dataset,” https://bit.ly/RTdataset, 2020, (Accessed on 06/24/2023)
[31] S. Liu, J. Cao, R. Yang, and Z. Wen, “Generating a structured summary of numerous academic papers: Dataset and method,” arXiv preprint arXiv:2302.04580, 2023.
[32] L. L. Wang, J. DeYoung, and B. C. Wallace, “Overview of mslr2022: A shared task on multi-document summarization for literature reviews,” in Proceedings of the Third Workshop on Scholarly Document Processing, 2022, pp. 175–180
[33] Lu, Y. Dong, and L. Charlin, “Multi-xscience: A large- scale dataset for extreme multi-document summarization of scientific articles,” in EMNLP, 2020, pp. 8068–8074
[34] J. DeYoung, I. Beltagy, M. van Zuylen, B. Kuehl, and L. Wang, “MSˆ2: Multi-Document Summarization of Medical Studies,” in EMNLP, 2021, pp. 7494–7513.
[35] L. Li, C. Forˇascu, M. El-Haj, and G. Giannakopoulos, “Multi-document multilingual summarization corpus preparation, part 1: Arabic, english, greek, chinese, romanian,” in Proceedings of the multiling 2013 workshop on multilingual multi-document summarization, 2013, pp. 1–12.
[36] L. Marina and V. Natalia, “Multilingual multi-document summarization with poly,” in Proceedings of the MultiLing 2013 Workshop on Multilingual Multidocument Summarization, 2013.
[37] DeYoung, J., Beltagy, I., van Zuylen, M., Kuehl, B., Wang, L.L.: Ms^2:Multi-document summarization of medical studies. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 7494–7513 (2021)
[38] D. Scialom, Thomas et al., “Mlsum: The multilingual summarization corpus,” in EMNLP, 2020, pp. 8051–8067.
[39] T. Hasan, Bhattacharjee et al., “Xl-sum: Large-scale multilingual abstractive summarization for 44 languages,” in ACL-IJCNLP 2021, 2021, pp. 4693–4703.
[40] F. Ladhak, E. Durmus, C. Cardie, and K. Mckeown, “Wikilingua: A new benchmark dataset for cross-lingual abstractive summarization,” in EMNLP 2020, 2020, pp. 4034–4048.
[41] D. Wang, J. Chen, H. Zhou, X. Qiu, and L. Li, “Contrastive aligned joint learning for multilingual summarization,” in ACL-IJCNLP 2021, 2021, pp. 2739– 2750.
[42] Y. Verma, A. Jangra, R. Verma, and S. Saha, “Large scale multi-lingual multimodal summarization dataset,” in ACL, 2023, pp. 3602–3614.
w generation. In: ACL. pp. 5151–5169 (2020)
[43] Erkan, G., Radev, D.R.: Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of artificial intelligence research 22, 457–479 (2004)
[44] Chu, E., Liu, P.: Meansum: A neural model for unsupervised multi-document abstractive summarization. In: ICML. pp. 1223–1232. PMLR (2019)[45] C.-Y. Lin, “Rouge: A package for automatic evaluation of summaries,” in Text summarization branches out, 2004, pp. 74–81
[45] Oved, N., Levy, R.: Pass: Perturb-and-select summarizer for product reviews. In:ACL-IJCNLP. pp. 351–365 (2021)
[46] Bapna, A., Firat, O.: Simple, scalable adaptation for neural machine translation.In: EMNLP-IJCNLP. pp. 1538–1548 (2019)
[47] Houlsby, N., Giurgiu, et al.: Parameter-efficient transfer learning for nlp. In: ICML. pp. 2790–2799. PMLR (2019)[35] L. Li, C. Forˇascu, M. El-Haj, and G. Giannakopoulos, “Multi-document multilingual summarization corpus preparation, part 1: Arabic, english, greek, chinese, romanian,” in Proceedings of the multiling 2013 workshop on multilingual multi-document summarization, 2013, pp. 1–12.
[48] Pfeiffer, J., Kamath, et al.: Adapterfusion: Non-destructive task composition for transfer le
[49] Pfeiffer, J., Vuli´c, I., Gurevych, et al.: Mad-x: An adapter-based framework for multi-task cross-lingual transfer. In: EMNLP. pp. 7654–7673 (2020)
[50] Hu, E.J., Wallis, et al.: Lora: Low-rank adaptation of large language models. In: ICLR (2021)
[51] He, R., McAuley, J.: Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In: proceedings of the 25th international conference on world wide web. pp. 507–517 (2016)

References

[52] Bra˘zinskas, A., Nallapati, et al.: Efficient few-shot fine-tuning for opinion summarization. In: NAACL. pp. 1509–1523 (2022)
[53] Karimi Mahabadi, R., Henderson, J., Ruder, S.: Compacter: Efficient low-rank hypercomplex adapter layers. Advances in Neural Information Processing Systems 34, 1022–1035 (2021)
[54] Pfeiffer, J., Vuli´c, I., Gurevych, et al.: Mad-x: An adapter-based framework for multi-task cross-lingual transfer. In: EMNLP. pp. 7654–7673 (2020)
[55] He, J., Zhou, C., et al.: Towards a unified view of parameter-efficient transfer learning. In: ICLR (2021

Thank you.


Thank you...