

LB/TH/43/2025

TH6011

**AUTOMATIC GENERATION OF RESEARCH  
PAPER ABSTRACTS USING DEEP-HYBRID  
MODELS**

R.P.D. Kumarasinghe

219354V

Master of Science in Computer Science

Department of Computer Science & Engineering  
Faculty of Engineering

University of Moratuwa  
Sri Lanka

February 2025

**AUTOMATIC GENERATION OF RESEARCH  
PAPER ABSTRACTS USING DEEP-HYBRID  
MODELS**

R.P.D. Kumarasinghe

219354V

Thesis submitted in partial fulfillment of the requirements for the degree  
Master of Science in Computer Science

Department of Computer Science & Engineering  
Faculty of Engineering

University of Moratuwa  
Sri Lanka

February 2025

## DECLARATION

I declare that this is my own work and this Thesis does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or Institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature:


Date: 28/02/2025

The supervisor should certify the Thesis with the following declaration.

The above candidate has carried out research for the Master of Science in Computer Science Thesis under my supervision. I confirm that the declaration made above by the student is true and correct.

Name of Supervisor: Nisansa de Silva

Signature of the Supervisor:

 Digitally signed by Nisansa  
de Silva  
Date: 2025.02.28 13:09:33  
+05'30'

Date: 28/02/2025

## **DEDICATION**

I dedicate this to all who strive to seek knowledge, believing that every step forward is a step towards a better world.

## **ACKNOWLEDGEMENT**

I would like to express my deepest gratitude to my research supervisor, Dr. Nisansa de Silva, for their unwavering guidance, support, and encouragement throughout the course of this thesis. Their insightful advice, constructive feedback, and immense patience have been invaluable in shaping the direction and quality of this research. I am deeply thankful for the time and effort they dedicated to helping me grow academically and professionally, and for always challenging me to think critically and aim for excellence.

Lastly, I extend my appreciation to my family and friends for their constant encouragement and understanding during this challenging process. Their support made this accomplishment possible.

## **ABSTRACT**

Condensing important information into a summary is crucial for readers navigating lengthy documents. In the context of research papers, the **abstract** serves as a concise overview of the study. This thesis focuses on enhancing research paper summarization by introducing a novel section-wise relevance matrix. To address the token size limitations of Large Language Models (LLMs), such as GPT-Neo, we developed a two-fold approach. First, we employed extractive summarization to condense lengthy texts into key sentences, followed by the application of abstractive summarization to generate coherent and concise summaries from these extracts.

Our approach, combining both extractive and abstractive techniques, leverages section-wise involvement ratios, with particular attention to the abstract section, improving the accuracy and quality of generated summaries. We introduced a pioneering dataset of research papers organized into sections, which plays a crucial role in this summarization process. Experimental results demonstrated that our method produces high-quality summaries while effectively overcoming token limitations, offering significant potential for summarizing long documents in low-resource and cost-effective environments.

However, challenges arise when section-wise segmentation is unclear, impacting the accuracy of summaries. This research underscores the need for further refinements and offers a promising framework for enhancing summarization techniques, benefiting researchers, educators, and information seekers alike.

**Keywords:** NLP, ATS, Summarization, Text Generation, LLM

## TABLE OF CONTENTS

Declaration of the Candidate & Supervisor	i
Dedication	ii
Acknowledgement	iii
Abstract	iv
Table of Contents	v
List of Figures	vii
List of Tables	viii
List of Abbreviations	viii
1 Introduction	1
1.1 Background	1
1.2 Research Problem	2
1.3 Research Objectives	3
2 Publications	4
3 Literature Survey	5
3.1 Popular tools	5
3.2 Validation Mechanisms	6
3.2.1 BLEU	7
3.2.2 ROUGE	8
3.2.3 METEOR	9
3.2.4 BERTScore	10
3.3 Related review papers	10
3.4 Paper review	17
3.5 Datasets	19
4 Methodology	21
4.1 Dataset	22
4.1.1 Dataset Generation	23
4.1.2 Cleaned Dataset	26

4.1.3	Raw Dataset	27
4.2	Approach	29
4.3	Pre-Summarization	33
4.4	Relevance Matrix	37
4.5	Encoding Data	39
4.6	Model Tuning	40
4.7	Prediction	42
4.8	Training Flow	44
5	Results	45
5.0.1	Pre-summarization methods comparison	45
5.0.2	Evaluation with Datasets	47
5.0.3	arXiv Dataset	47
5.0.4	PubMed Dataset	47
5.0.5	New Dataset	48
6	Conclusion	49
	References	51

## LIST OF FIGURES

<b>Figure</b>	<b>Description</b>	<b>Page</b>
Figure 3.1	Text summarization classification	12
Figure 3.2	Extractive high level architecture	13
Figure 3.3	Abstractive high level architecture	14
Figure 3.4	Automatic Text Summarization Approaches with their methods	15
Figure 3.5	Extractive summarization techniques	16
Figure 3.6	Abstractive summarization techniques	17
Figure 4.1	Research paper primary tag ( <b>arXiv</b> tag) portions in the dataset	22
Figure 4.2	Tuning GPT-Neo for abstract generation	30
Figure 4.3	Predicting abstracts with GPT-Neo	31
Figure 4.4	Token size portions for GPT-Neo model feeding	32
Figure 4.5	Token size distribution of abstract sections	33
Figure 4.6	Data preparation overview	35
Figure 4.7	Average involvement for the abstract with percentages of each section of research papers in our dataset	37
Figure 4.8	Paper encoding	39
Figure 4.9	Fine-tuning GPT-Neo	42
Figure 4.10	Training flow. 1. Generation of tfrecords, 2. Fine-tuning GPT-Neo	44
Figure 5.1	ROUGE scores comparison of pre-summarization method	45

## LIST OF TABLES

<b>Table</b>	<b>Description</b>	<b>Page</b>
Table 3.1	Review of identified survey papers on automatic text summarization	10
Table 3.2	ROUGE score of the text summarization methods on DUC 2007 dataset	16
Table 3.3	Popular datasets	20
Table 5.1	ROUGE scores comparison of the models based on the pre-summarization method	45
Table 5.2	ROUGE scores of average vector-based pre-summarizing	46
Table 5.3	ROUGE scores of Lex Rank-based pre-summarizing	46
Table 5.4	ROUGE scores of Text Rank-based pre-summarizing	46
Table 5.5	ROUGE scores of LSA-based pre-summarizing	47
Table 5.6	ROUGE scores comparison of the models on datasets. R1: ROUGE-1, R2: ROUGE-2, R3: ROUGE-3, RL: ROUGE-L	48

## LIST OF ABBREVIATIONS

<b>Abbreviation</b>	<b>Description</b>
BLEU	Bilingual Evaluation Understudy
CPUs	Central Processing Units
GPUs	Graphics Processing Units
I/O	Input/Output
LCS	Longest Common Subsequence
LLMs	Large Language Models
LSA	Latent Semantic Analysis
NLP	Natural Language Processing
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
TPUs	Tensor Processing Units