

MULTI-OBJECTIVE EXTRACTIVE TEXT SUMMARIZATION

*Report submitted to the SASTRA Deemed to be University as the
requirement for the course*

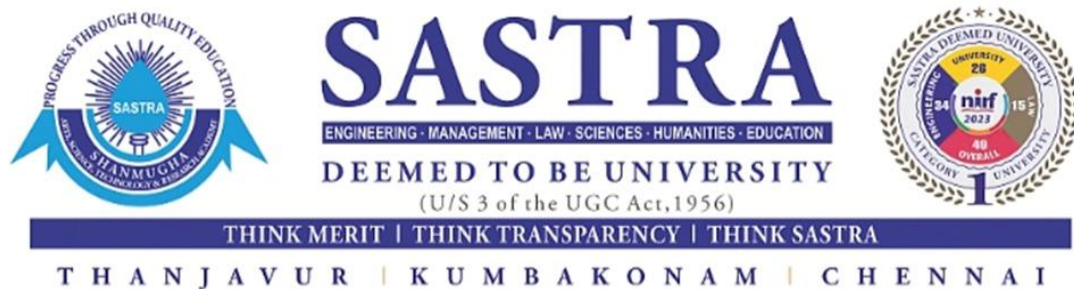
CSE400: PROJECT WORK

Submitted by

AKSHAYA ROHITH M
(Reg No: 224003008, B.Tech., CSE)

HARIHARAN SUBRAMANIAN
(Reg No: 224003032, B.Tech., CSE)

MAY 2024



SRINIVASA RAMANUJAN CENTRE
KUMBAKONAM, TAMIL NADU INDIA - 612001



SRINIVASA RAMANUJAN CENTRE

KUMBAKONAM – 612 001

Bonafide Certificate

This is to certify that the report titled “**Multi-Objective Extractive Text Summarization**” submitted as a requirement for the course, **CSE 400: FINAL PROJECT** for B.Tech. is a bonafide record of the work done by Mr. AKSHAYA ROHITH.M(224003008, B.Tech.,-CSE), Mr. HARIHARAN SUBRAMANIAN(224003032, B.Tech.,-CSE) during the academic year 2023-24, in the Srinivasa Ramanujan Centre, under my supervision.

Signature of Project Supervisor:

Name with Affiliation : Smt. S. Hemamalini, AP-II/CSE/SRC/SASTRA

Date :

Final Project *Viva voce* held on 07.05.2024

Examiner 1

Examiner 2



We declare that the report titled “**Multi-Objective Extractive Text Summarization**” submitted by us is an original work done by us under the guidance of **Smt. S. Hemamalini, AP-II/CSE/SRC/SASTRA, SASTRA Deemed to be University** during the final semester of the academic year 2023-24, in the **Srinivasa Ramanujan Centre**. The work is original and wherever we have used materials from other sources, we have given due credit and cited them in the text of the report. This report has not formed the basis for the award of any degree, diploma, associate-ship, fellowship or other similar title to any candidate of any university.

Date : 07.05.2024

ACKNOWLEDGEMENTS

We pay our sincere obeisance to the God Almighty for his grace and infinite mercy and for showing on us his choicest blessings.

We would like to thank our honorable Chancellor **Prof. R. Sethuraman**, Vice Chancellor **Dr. S. Vaidhyasubramaniam** and **Dr. S. Swaminathan**, Dean – Planning and Development for the encouragement and strategic support at every step of our college life.

We extend our sincere thanks to **Dr. R Chandramouli**, Registrar, SASTRA Deemed to be University for providing the opportunity to pursue this project.

We express our deepest thanks to **Dr. V. Ramaswamy**, Dean and **Dr. A. Alli Rani**, Associate Dean, Srinivasa Ramanujan Centre for their constant support and suggestions when required without any reservations.

We express our gratitude to HOD in-charge **Dr. V. Kalaichelvi**, ACP/CSE for his constant support and valuable suggestions for the completion of the project.

We exhibit our pleasure in expressing our thanks to **Smt. S. Hemamalini**, AP-II/CSE, our guide for her ever-encouraging spirit and meticulous guidance for the completion of the project.

We would like to place on record the benevolent approach and painstaking efforts of guidance and correction of **Dr. J. Sangeetha** and **Smt. D. Rekha**, the project coordinators and all department staff to whom we owe our hearty thanks forever.

Without the support of our parents and friends this project would never have become reality.

We dedicate this work to our well-wishers, with love and affection.

LIST OF FIGURES

Figure No.	Title	Page No.
2.1	Architecture Diagram	5
2.2	Objectives, Constraints and Solution	5
3.1	Pre-processing of data	6
3.2	Calculations involved in Sentence Representation	7
3.3	Representation of sentences in term space	8
3.4	Caching of cosine similarity values between various sentences	9
3.5	Steps in MOABC	10
3.6	Colony Initialization for MOABC	11
3.7	Steps in Employed Bee Phase	11
3.8	Version 1 - Mutation process	12
3.9	Average Sentence Similarity Check	13
3.10	NDS information updation in a voting dictionary	14

3.11 – 3.12	Domination criteria check comparison	15
3.13	Steps in Onlooker Bee Phase	18
3.14	Crowding Distance expression in graphical form	20
3.15-3.16	Ranking and crowding distances in a pareto view	21
3.17	Steps in Scout Bee Phase	23
3.18	Steps involved in NSGA-II algorithm	24
3.19	Sample Pareto Front	26
3.20	NDS Archive	26
3.21	NDS Insert Function	27
3.22	Various Single Solution Extraction methods experimented	28
3.23	Objectives Expression	33
3.24	Sentence Repair Scheme	35
3.25	Topic Modelling using Hierarchical Dirichlet Process (HDP)	36

3.26	Overall Process flow in ETS	37
4.1	Consolidated Performance Measure of all Experiments performed	46
4.2	Pareto Front Trend	48
4.3	Objective trend over iterations	49

LIST OF TABLES

Table No.	Title	Page No.
3.1	Table of distance measures and expressions	32
3.2	Parameters	38
4.1	Comparison of different Mutation Operator Versions with Single Solution Extraction methods	39
4.2	Comparison between MOABC and NSGA with Single Solution Extraction methods	40
4.3	Comparison of Various Replacement Operators with Single Solution Extraction methods	40
4.4	Comparison of TFISF and SBERT with Single Solution Extraction methods	41
4.5	Comparison of Topic Modelling versions with Single Solution Extraction methods	41
4.6	Mutation Operators Comparison across various documents	42
4.7	Comparison of both MOABC and NSGA algorithms over various documents	43
4.8	Comparison of different versions of Replacement Operators over various documents	43
4.9	Comparison of TFISF vs SBERT Sentence Representation over various documents	44
4.10	Comparison of Different Topic modelling Versions over various documents	45

4.11	Consolidated Performance Measure of all Experiments performed	46
4.2.1	Comparison of Our proposed Method with Contemporary Methods	47

ABBREVIATIONS

ECA	:	Evolutionary Computing Algorithm
ETS	:	Extractive Text Summarization
MOABC	:	Multi Objective Artificial Bee Colony Algorithm
NSGA	:	Non-dominated Sorting Algorithm
TF-ISF	:	Term Frequency Inverse Sentence Frequency
NDS	:	Non-Dominated Solution
ROUGE	:	Recall-Oriented Understudy for Gisting Evaluation
BCTSO	:	Binary Crowded Tournament Selection Operator
RCR	:	Ranking and Crowding based Replacement
HDP	:	Hierarchical Dirichlet Process
SBERT	:	Sentence Bidirectional Representation Transformer

ABSTRACT

Extractive summarization is a technique for automatically creating summaries of text documents by selecting the most important sentences from the original document. Redundancy reduction minimizes repetition in the summary, preventing information overload and improving conciseness. Coverage ensures that the summary faithfully conveys the main points and key concepts of the original text.

Current systems approach text summarization within a single-objective modelling framework, that are known to only capture homogenous facets of content and simple relationships within the text, potentially affecting the overall effectiveness and richness of the generated summaries. Multi-objective optimization strategies have shown to effectively tackle conflicting objectives such as the above. The Multi-Objective Artificial Bee Colony algorithm (MOABC) emerges as a promising approach, effectively handling both objectives through its population-based exploration. The repair operator is revised to a less computationally intensive repair operator. Then, some slight modifications in the mutation and replacement operators are proposed and its performance in the search process is recorded. SBERT transformers as a possible sentence embedding structure is also investigated and finally topic modelling is incorporated. In this project, ROUGE scores are used as the performance evaluation criteria.

Keywords: Extractive Summarization, Redundancy Reduction, Coverage, Multi-Objective Artificial Bee Colony Algorithm, ROUGE.

Table of Contents

Title	Page No.
Bonafide Certificate	ii
Declaration	iii
Acknowledgment	iv
List of Figures	v
List of Tables	viii
Abbreviations	x
Abstract	xi
1. Summary of Base Paper	1
1.1 Introduction	1
1.2 Demerits of Existing System	2
1.3 Merits of Proposed System	2
1.4 Dataset	2
1.5 Motivation	3
2. Objective	4
3. Methodology	6
3.1 Data Preprocessing and representation	6
3.1.1 Data Pre-processing	6
3.1.2 Sentence Representation	7
3.1.3 Similarity Measure	8
3.2 Multi-Objective Artificial Bee Colony Algorithm	9
3.2.1 Colony Initialization	10
3.2.2 Sending Employee Bee Phase	11
3.2.3 Sending Onlooker Bee Phase	17
3.2.4 Sending Scout Bee Phase	22
3.3 Non-Dominated Sorting Algorithm	24
3.3.1 Mutation Operation	24
3.3.2 Ranking and Crowding Distance	25
3.3.3 Binary Crowded Tournament Selection	25
Operator	
3.4 Pareto Front and NDS Archive	25
3.5 Single Solution Extraction	27
3.5.1 Consensus	28
3.5.2 Largest Hypervolume	29
3.5.3 Topmost Coverage	30
3.5.4 Topmost Redundancy Reduction	30
3.5.5 Shortest Distance to Ideal Point	30
3.5.6 Shortest Distance to All Points	31
3.6 Objective Expression	32
3.7 Sentence Scoring Schemes	33
3.8 Sentence Repair Scheme	34
3.9 Topic Modelling	35

3.10 Evaluation Metrics	36
3.11 Overall Process Flow	37
3.12 Parameters	38
4. Results	39
4.1 Experiments	39
4.1.1 Comparison of Performance of Single Solution Extraction methods	39
4.1.2 Other Experiments	42
4.1.2.1 Comparison of Mutation Operator	42
4.1.2.2 Comparison of Algorithms	43
4.1.2.3 Comparison of Different Replacement Operators	43
4.1.2.4 Comparison of TFISF vs SBERT Sentence Representations	44
4.1.2.5 Comparison of Different Topic Modelling Versions	44
4.2 Inferences and Results	46
4.2.1 Contemporary Methods Comparison	47
4.3 Trends in Pareto Front and Objective	48
5. Conclusion and Future Work	50
5.1 Conclusion	50
5.2 Future Work	50
6. Source Code	51
7. Output	106
8. References	109
9. Appendix	112
9.1 Base Paper	112