

LB/TH/33/2025
TH5911

**AUTOMATIC SUMMARY GENERATION BY DATA
EXTRACTION OF SUNBURST CHARTS**

Hinguralage Lahiru Pradeep Ariyasinghe
218520U

Master of Science in Artificial Intelligence

Department of Computational Mathematics
Faculty of Information Technology

University of Moratuwa
Sri Lanka

April 2024

AUTOMATIC SUMMARY GENERATION BY DATA EXTRACTION OF SUNBURST CHARTS

Hinguralage Lahiru Pradeep Ariyasinghe
218520U

Thesis/Dissertation submitted in partial fulfillment of the requirements for the
degree.

Master of Science in Artificial Intelligence

Department of Computational Mathematics
Faculty of Information Technology

University of Moratuwa
Sri Lanka

May 2024

DECLARATION

I declare that this is my own work, and this thesis/dissertation 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: 28/04/2024

The above candidate has carried out research for the PhD/MPhil/Master's thesis/dissertation under my supervision. I confirm that the declaration made above by the student is true and correct.

Name of Supervisor: Dr Thushari Silva

Signature of the Supervisor:

Date: 28/04/2024

DEDICATION

The research thesis is dedicated wholeheartedly to my beloved wife and parents who inspired, motivated and supported me throughout the study.

ACKNOWLEDGEMENT

I would like to thank my academic and research supervisor, Dr Thushari Silva, for her generous support and assistance. Also, I want to express my deep appreciation to Dr. Sagara Sumathipala., for their professional support and supervision during this study.

I am thankful to the academic staff at the Faculty of IT, Department of Computational Mathematics, University of Moratuwa, for their guidance and knowledge during our undergraduate years. Also, members of my family, our friends and batch mates who have supported us in numerous ways.

ABSTRACT

Sunburst charts are useful for visualizing hierarchical data structures. However, they can be challenging to interpret because of their nested hierarchy and circular layout. Viewers must determine the color and size of each segment and navigate through various hierarchy levels, which can be confusing, especially with many or unevenly distributed categories.

To address these challenges, this research proposes a method to extract summary data from sunburst charts, enabling quick identification of key trends and patterns. This approach is particularly beneficial for business analysts and marketers who need to analyze large datasets efficiently. By automating the extraction process, analysts can focus their efforts on areas of significance, saving time and improving efficiency.

Existing literature indicates a gap in tools for extracting data and generating summaries of sunburst charts. To fill this gap, the author implemented a system that leverages Machine Learning (ViT models), image processing, optical character recognition (OCR), and LLM. While several methods could have been used for this project, the chosen technologies offer a robust solution for data extraction and summary generation from sunburst charts.

Keywords: ViT Models, NLP, OCR, Machine Learning

TABLE OF CONTENTS

Declaration	1
Dedication	2
Acknowledgement.....	3
Abstract	4
Table of Contents	5
List of Figures	8
Chapter 1	9
Introduction	9
1.1 Introduction	9
1.2 Prolegomena.....	9
1.3 Problem Definition	9
1.3.1 Problem Identified	10
1.4 Background and Motivation	10
1.5 Gaps in Research	10
1.6 Contribution to the Body of Knowledge	10
1.6.1 Technological Contribution.....	10
1.6.2 Domain Contributions	11
1.7 Research Challenges.....	11
1.8 Aims	11
1.8 Objectives	12
1.9 The Novelty of the Research	12
1.10 Project Scope	12
1.11 Chapter Summary	12
Chapter 2	13
LITERATURE REVIEW ON EXISTING CHART SUMMARIZATION SYSTEMS	13
2.1 Introduction	13
2.2 Problem Domain.....	13
2.3 Existing Systems	14
2.6 Chapter Summary	17
Chapter 3	18

TECHNOLGY ADOPTED IN CHART SUMMARIZATION SYSTEM	18
3.1 Introduction	18
3.2 Vision Transformer (ViT) Models	18
3.3 Optical Character Recognition (OCR)	18
3.4 Generative Pre-training Transformer (GPT)	19
3.5 Python.....	19
3.6 Version Controlling System	19
3.7 Open-Source Computer Vision Library (OpenCV)	19
3.8 Microsoft Visual Studio	19
3.9 Chapter Summary	20
Chapter 4.....	20
APPROACH Of SUNBURST CHART SUMMARIZATION SYSTEM.....	20
4.1 Introduction	20
4.2 Hypothesis	20
4.3 Input.....	20
4.4 Output.....	20
4.5 Process.....	21
4.6 Users	21
4.7 Summary	21
Chapter 5	22
DESIGN Of SUNBURST CHART SUMMARIZATION SYSTEM.....	22
5.1 Introduction	22
5.2 Design Goals	22
5.3 System Architecture Design (High-level)	23
5.3.1 Design Diagram.....	23
5.4 System Design.....	24
5.4.1 Methodology of designed system.....	24
5.5 Design Diagrams	25
5.5.1 Activity Diagram.....	25
5.5.2 Data Flow Diagram	26
5.6 UI Wireframe Design	27
5.7 Chapter Summary	27

Chapter 6	28
IMPLEMENTATION Of SUNBURST CHART SUMMARIZATION SYSTEM ..	28
6.1 Introduction	28
6.2 Implementation of sub modules	28
6.2.1 Image pre-processing	28
6.2.2 Circles Detection	29
6.2.3 Segments Detection	30
6.2.4 Calculate Percentages and Colors	30
6.2.5 Text Detection and Extraction.....	31
6.2.6 Summary Generation.....	31
6.2 Chapter Summary.....	32
Chapter 7	33
EVALUATION OF CHART SUMMARIZATION SYSTEM FOR SUNBURST CHARTS.....	33
7.1 Introduction	33
7.1.1 Circles Detection	33
7.1.2 Segments Detection	33
7.1.3 Calculate Percentages and Colors	33
7.1.4 Texts Extraction	34
7.1.5 Summary Generation.....	34
7.4 Conclusion.....	34
Chapter 8	35
CONCLUSION, LIMITATIONS AND FURTHER WORKS.....	35
8.1 Introduction	35
8.2 Conclusion.....	35
8.3 Limitations.....	35
8.3.1 Preprocessing and Circles Detection.....	35
8.3.2 Texts Extraction	36
8.3.5 Summary Generation.....	36
8.2 Further Works.....	36
References	37
Appendix A – Source Code.....	40

LIST OF FIGURES

Figure	Description	Page
Figure 3.1	Process of the ViT Model	16
Figure 4.2	Process of the sunburst chart summarization	16
Figure 5.1	Architecture Diagram	18
Figure 5. 2	Activity diagram	19
Figure 6.3:	Data flow diagram L1	21
Figure 6.4:	Data flow diagram L2	21
Figure 6.5:	UI Wireframe	22