



**INTELLIGENT DOCUMENT ANALYSIS AND NATURAL  
LANGUAGE PROCESSING: A CONVERSATIONAL AI  
APPROACH FOR FILE-BASED KNOWLEDGE  
EXTRACTION USING AUTOMATION SYSTEM**

**UNDERGRADUATE THESIS**

**Submitted as one of the requirements to obtain  
Sarjana Komputer**

**By:**

**IVAN YOHANES SIREGAR**

**001202000050**

**FACULTY OF COMPUTING  
INFORMATICS STUDY PROGRAM**

**CIKARANG**

**SEPTEMBER, 2023**

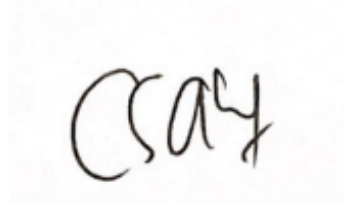
**INTELLIGENT DOCUMENT ANALYSIS AND NATURAL  
LANGUAGE PROCESSING: A CONVERSATIONAL AI  
APPROACH FOR FILE-BASED KNOWLEDGE EXTRACTION  
USING AUTOMATION SYSTEM**

By

IVAN YOHANES SIREGAR

001202000050

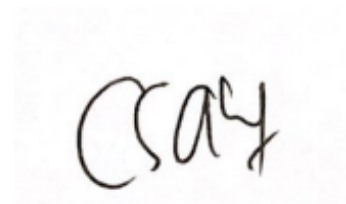
Approved:



---

Cutifa Safitri, Ph.D.

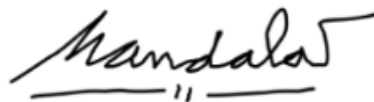
Thesis Advisor



---

Cutifa Safitri, Ph.D.

Program Head of Informatics



---

Rila Mandala, Ph.D.

Dean of Faculty of Computing

## **PANEL OF EXAMINER APPROVAL**

The Panel of Examiners declare that the undergraduate thesis entitled "**Intelligent Document Analysis and Natural Language Processing: A Conversational AI Approach for File-based Knowledge Extraction using Automation System**" that was submitted by **IVAN YOHANES SIREGAR** majoring in **Informatics** from the Faculty of Computer Science was assessed and approved to have passed the Oral Examination on Thursday September 21, 2023.

### **Panel of Examiner**



**ABDUL GHOFIR**

### **Chair of Panel Examiner**



**RUSDIANTO ROESTAM**

### **Examiner I**

## STATEMENT OF ORIGINALITY

In my capacity as an active student at President University and as the author of the final project stated below:

Name : IVAN YOHANES SIREGAR

Student ID number : 001202000050

Study Program : Informatics

Faculty : Computer Science

I hereby declare that my final project entitled “**Intelligent Document Analysis and Natural Language Processing: A Conversational AI Approach for File-based Knowledge Extraction using Automation System**” is to the best of my knowledge and belief, an original piece of work based on sound academic principles. If there is any plagiarism detected in this final project, I am willing to be personally responsible for the consequences of these acts of plagiarism and will accept the sanctions against these acts in accordance with the rules and policies of President University.

I also declare that this work, either in whole or in part, has not been submitted to another university to obtain a degree.

Cikarang, 25<sup>th</sup> September 2023



IVAN YOHANES SIREGAR

**SCIENTIFIC PUBLICATION APPROVAL FOR ACADEMIC  
INTEREST**

As an academic community member of the President's University, I, the undersigned:

Name : IVAN YOHANES SIREGAR

Student ID number : 001202000050

Study program : Informatics

for the purpose of development of science and technology, certify, and approve to give President University a non-exclusive royalty-free right upon my final report with the title:

**“Intelligent Document Analysis and Natural Language Processing: A Conversational AI Approach for File-based Knowledge Extraction using Automation System”**

With this non-exclusive royalty-free right, President University is entitled to converse, to convert, to manage in a database, to maintain, and to publish my final report. There are to be done with the obligation from President University to mention my name as the copyright owner of my final report.

This statement I made in truth.

Cikarang, 25<sup>th</sup> September 2023



IVAN YOHANES SIREGAR

**ADVISOR APPROVAL FOR JOURNAL/INSTITUTION'S  
REPOSITORY**

As an academic community member of the President's University, I, the undersigned:

Name : Cutifa Safitri, Ph.D.

ID number : 20190900815

Study program : Informatics

Faculty : Computing

declare that following thesis:

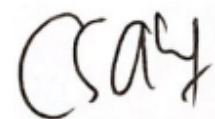
Title of thesis : **Document Analysis and Natural Language Processing: A  
Conversational AI Approach for File-based Knowledge  
Extraction using Automation System**

Thesis author : IVAN YOHANES SIREGAR

Student ID number : 001202000050

will be published in **institution's repository**.

Cikarang, 25<sup>th</sup> September 2023



Cutifa Safitri, Ph.D.

# PLAGIARISM CHECK RESULT

aall ch

## ORIGINALITY REPORT

7%

SIMILARITY INDEX

7%

INTERNET SOURCES

2%

PUBLICATIONS

0%

STUDENT PAPERS

## PRIMARY SOURCES

1	<a href="https://repository.president.ac.id">repository.president.ac.id</a> Internet Source	5%
2	<a href="https://dokumen.pub">dokumen.pub</a> Internet Source	<1%
3	Lecture Notes in Computer Science, 2015. Publication	<1%
4	<a href="https://deepai.org">deepai.org</a> Internet Source	<1%
5	<a href="https://mdpi-res.com">mdpi-res.com</a> Internet Source	<1%
6	<a href="https://www.irjmets.com">www.irjmets.com</a> Internet Source	<1%
7	<a href="https://intellipaat.com">intellipaat.com</a> Internet Source	<1%
8	<a href="https://9pdf.org">9pdf.org</a> Internet Source	<1%
9	<a href="https://medium.com">medium.com</a> Internet Source	<1%

# GPTZERO RESULT



This text is likely to be written by a **human**

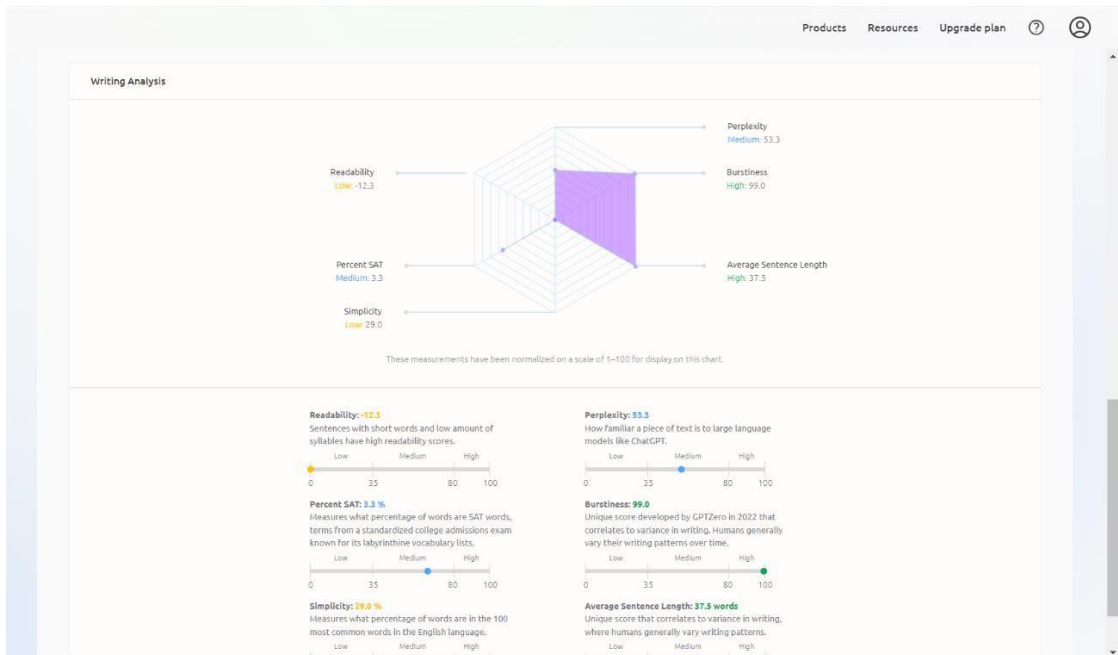
There is a **5%** probability this text was entirely written by AI ⓘ

Sentences that are likely written by AI are **highlighted**.

The system allows users to have interactive conversations and extract knowledge from uploaded files. The application is developed using R for the AI code, UI Path for automation, and Tkinter for building the user interface. By applying advanced natural language processing techniques, the system interprets user queries and provides accurate responses based on the uploaded material. The research contributes to the field of knowledge extraction and retrieval by demonstrating a practical application that combines conversational AI, automation, and file-based analysis. Through experimental evaluations, the effectiveness of the system in extracting valuable insights from various documents is demonstrated. The findings emphasize the potential of such applications in improving information retrieval and decision-making processes. This research lays the groundwork for future advancements in intelligent document analysis, offering a valuable tool for knowledge extraction and facilitating more efficient access to information resources. Keywords: intelligent document analysis, natural language processing, conversational AI, knowledge extraction, information retrieval, automation ii ii DEDICATION I dedicate this Final Project to my family and myself who always provide peace, comfort, motivation, the best prayers, and set aside their finances. iii iii ACKNOWLEDGEMENT I want to say thank you to all the people who helped me with my Final Project. First, I want to thank God for giving me

... only the first 5000 characters are shown in the free version of GPTZero. If you need a higher limit please check the [Subscription plans](#) available.

2/19 sentences are likely AI generated. ⓘ





## **ABSTRACT**

This final project presents an innovative approach to intelligent document analysis and natural language processing using a conversational AI system. The system allows users to have interactive conversations and extract knowledge from uploaded files. The application is developed using R for the AI code, UI Path for automation, and TKinter for building the user interface. By applying advanced natural language processing techniques, the system interprets user queries and provides accurate responses based on the uploaded material. The research contributes to the field of knowledge extraction and retrieval by demonstrating a practical application that combines conversational AI, automation, and file-based analysis. Through experimental evaluations, the effectiveness of the system in extracting valuable insights from various documents is demonstrated. The findings emphasize the potential of such applications in improving information retrieval and decision-making processes. This research lays the groundwork for future advancements in intelligent document analysis, offering a valuable tool for knowledge extraction and facilitating more efficient access to information resources.

***Keywords: intelligent document analysis, natural language processing, conversational AI, knowledge extraction, information retrieval, automation***

## **ACKNOWLEDGEMENT**

I want to say thank you to all the people who helped me with my Final Project.

First, I want to thank God for giving me the strength to complete this school project.

I'm really thankful to Mam Cutifa Safitri for teaching me a lot and supporting me. She's been a big help, not just in school but in making me a better person.

I also want to thank my lecturers at President University for helping me learn and grow. I appreciate all the things they taught me, and I hope God continues to bless them.

My family has been amazing with their love, support, and prayers. They've given me the strength to keep going.

Lastly, I want to thank Brigitta Sheren Patricia, S.IP. and my friends who helped me along the way. Your prayers, kind words, and sharing of knowledge showed me how important it is to have a supportive community. It's amazing what we can achieve together.

I'm really grateful to all of you for being a part of my journey and helping me succeed. Thank you.

# TABLE OF CONTENTS

ABSTRACT.....	i
DEDICATION.....	ii
ACKNOWLEDGEMENT .....	iii
TABLE OF CONTENTS.....	iv
LIST OF FIGURES .....	viii
LIST OF TABLE .....	ix
1 CHAPTER I INTRODUCTION.....	1
1.1 Background .....	1
1.2 Problem Statement .....	2
1.3 Objectives.....	3
1.4 Scope and Limitation .....	4
1.4.1 This thesis has several scopes: .....	4
1.4.2 This thesis has several limitations.....	5
1.5 Project Methodology .....	6
1.6 Final Project Outline .....	8
2 CHAPTER II LITERATURE REVIEW .....	10
2.1 Document Analysis and Text Mining .....	10
2.2 File-based Knowledge Extraction .....	11
2.3 R Programming Language .....	13
2.4 UI Path and Automation.....	14
2.5 Tkinter .....	16
2.6 Collaboration of Agile methodology and Lean methodology.....	18
2.7 Related Works .....	19
2.7.1 Google's Knowledge Graph .....	19

2.7.2	BERT .....	20
2.7.3	Word2Vec .....	21
2.7.4	Comparison Overview .....	22
2.8	Algorithm Implementation .....	24
2.8.1	Natural Language Processing (NLP) Algorithms .....	25
2.8.2	Document Clustering Algorithms .....	26
2.8.3	Information Extraction Algorithm .....	29
2.8.4	Text Classification Algorithm .....	31
2.8.5	Text Summarization Algorithms .....	33
2.8.6	Algorithm Comparison .....	35
3	CHAPTER III SYSTEM ANALYSIS .....	38
3.1	System Overview .....	38
3.2	Function Analysis .....	39
3.3	Use Case Diagram .....	41
3.4	Use Case Narrative .....	42
3.5	Swim Lane Diagram .....	58
3.5.1	Swim Lane Diagram for Upload File .....	59
3.5.2	Swim Lane Diagram for Asking Question .....	60
3.5.3	Swim Lane Diagram for Analyze Document .....	61
3.5.4	Swim Lane Diagram for Extract Entities .....	62
3.5.5	Swim Lane Diagram for Generate Summary .....	63
3.5.6	Swim Lane Diagram for Perform Sentiment Analysis .....	63
3.5.7	Swim Lane Diagram for Perform Clustering .....	64
3.5.8	Swim Lane Diagram for Extract Information .....	65
3.5.9	Swim Lane Diagram for Question Answered .....	66
3.6	Hardware and Software Requirement .....	67
3.6.1	Hardware Requirement .....	67
4	CHAPTER IV SYSTEM DESIGN .....	70
4.1	User Interface Design .....	70

4.1.1	Upload File.....	70
4.1.2	Asking Question.....	71
4.1.3	Background Document Analysis Process .....	72
4.1.4	Entity Extraction Process .....	74
4.1.5	Summarization Process .....	75
4.1.6	Sentiment Analysis Process .....	76
4.1.7	Clustering Process .....	78
4.1.8	Information Extraction Process .....	80
4.1.9	Question Answered Interface.....	81
5	CHAPTER V SYSTEM IMPLEMENTATION.....	83
5.1	User Interface .....	83
5.1.1	Upload File.....	84
5.1.2	Asking Question.....	85
5.1.3	Question Answered .....	86
5.2	System Detail .....	87
5.2.1	Upload File.....	88
5.2.2	Asking Question.....	89
5.2.3	Analyze Document.....	91
5.2.4	Entity Extraction .....	92
5.2.5	Summarization Process .....	94
5.2.6	Sentiment Analysis Process .....	97
5.2.7	Clustering Process .....	99
5.2.8	Information Extraction .....	101
5.2.9	Question Answered .....	103
5.3	Industrial Scope.....	106
		vii
5.3.1	Waiting Issue.....	106
5.3.2	Inventory Issue .. ..	107
5.3.3	Scrap Issue .....	108
6	CHAPTER VI SYSTEM T VALUATION .....	110

6.1	Testing Environment.....	110
6.2	Suggest Evaluation.....	115
7	CHAPTER VII CONCLUSION AND FUTURE WORKS .....	119
7.1	Conclusion.....	119
7.2	Future Works.....	120
	REFERENCES .....	123

## LIST OF FIGURES

<i>Figure 1.1 Agile Methodology</i> .....	6
<i>Figure 1.2 Lean Methodology</i> .....	6
<i>Figure 2.1 Google Knowledge Graph</i> .....	20
<i>Figure 2.2 BERT</i> .....	21
<i>Figure 2.3 Word2Vec (CBOW and Skip-gram)</i> .....	22
<i>Figure 2.4 Named entity recognition (NER) Algorithm</i> .....	25
<i>Figure 2.5 Sentiment Analysis Algorithm</i> .....	26
<i>Figure 2.6 Topic Modeling Algorithm</i> .....	26
<i>Figure 2.7 Clustering Algorithm</i> .....	28
<i>Figure 2.8 Relation Extraction Algorithm</i> .....	30
<i>Figure 2.9 Named Entity Linking algorithm</i> .....	31
<i>Figure 2.10 Text Classification Algorithms</i> .....	32
<i>Figure 2.11 Abstractive Summarization Algorithm</i> .....	34
<i>Figure 3.1 Use Case Diagram Application</i> .....	42
<i>Figure 3.2 Swim Lane Diagram of Upload File</i> .....	59
<i>Figure 3.3 Swim Lane Diagram of Asking Question</i> .....	60
<i>Figure 3.4 Swim Lane Diagram of Analyze Document</i> .....	61
<i>Figure 3.5 Swim Lane Diagram of Extract Entities</i> .....	62
<i>Figure 3.6 Swim Lane Diagram of Generate Summary</i> .....	63
<i>Figure 3.7 Swim Lane Diagram of Perform Sentiment Analysis</i> .....	64
<i>Figure 3.8 Swim Lane Diagram of Perform Clustering</i> .....	65
<i>Figure 3.9 Swim Lane Diagram of Extract Information</i> .....	66
<i>Figure 3.10 Swim Lane Diagram for Question Answered</i> .....	67
<i>Figure 4.1 Upload File Interface</i> .....	71
<i>Figure 4.2 Asking Question Interface</i> .....	72
<i>Figure 4.3 Question Answered UI</i> .....	81
<i>Figure 5.1 Data Correctness Survey</i> .....	106
<i>Figure 5.2 Waiting Issue Chart</i> .....	107
<i>Figure 5.3 Inventory Issue Chart</i> .....	108
<i>Figure 5.4 Scrap Issue Chart</i> .....	109

## LIST OF TABLE

<i>Table 2.1 Feature Comparison Table</i> .....	23
<i>Table 3.1 Table of Function Description</i> .....	39
<i>Table 3.2 Use Case Narrative for “Upload File” Use Case</i> .....	42
<i>Table 3.3 Use Case Narrative for “Asking Question” Use Case</i> .....	44
<i>Table 3.4 Use Case Narrative for “Analyze Document” Use Case</i> .....	46
<i>Table 3.5 Use Case Narrative for “Extract Entities” Use Case</i> .....	48
<i>Table 3.6 Use Case Narrative for “Generate Summary” Use Case</i> .....	50
<i>Table 3.7 Use Case Narrative for “Perform Sentiment Analysis” Use Case</i>	51
<i>Table 3.8 Use Case Narrative for “Perform Clustering” Use Case</i> .....	53
<i>Table 3.9 Use Case Narrative for “Extract Information” Use Case</i> .....	55
<i>Table 3.10 Use Case Narrative for “Question Answered” Use Case</i> .....	56
<i>Table 6.1 Uploading a Document Scenario</i> .....	111
<i>Table 6.2 Asking Question Scenario</i> .....	112
<i>Table 6.3 Generating a Summary Scenario</i> .....	112
<i>Table 6.4 Performing Sentiment Analysis Scenario</i> .....	113
<i>Table 6.5 Extracting Entities Scenario</i> .....	113
<i>Table 6.6 Clustering Documents Scenario</i> .....	114
<i>Table 6.7 Extracting Information from Multiple Documents Scenario</i> .....	114
<i>Table 6.8 Real-time Analysis Scenario</i> .....	115