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

Mandala

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, 25th 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, 25th 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

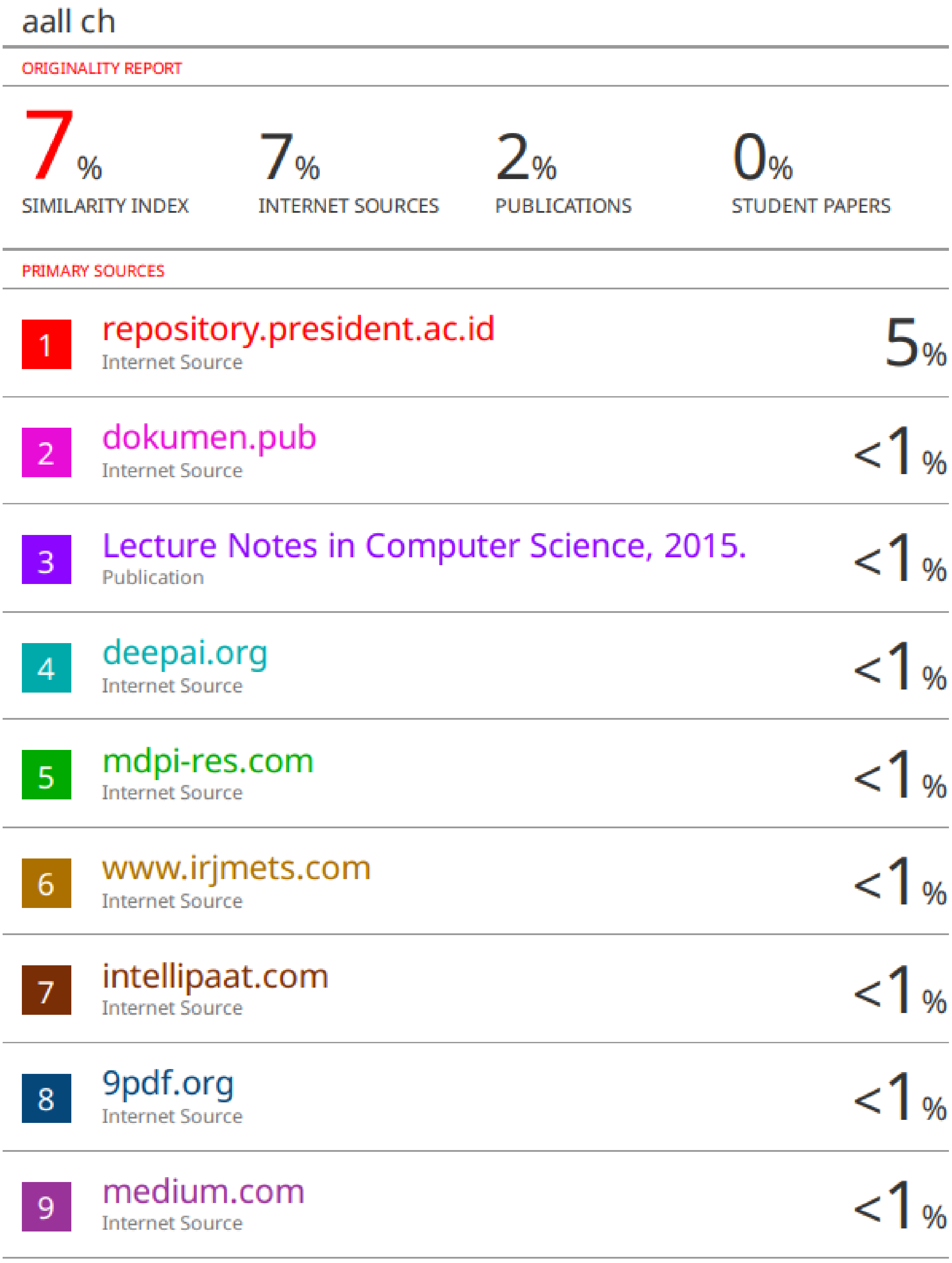
Student ID number 001202000050

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

Cikarang, 25th September 2023

Cutifa Safitri, Ph.D.

PLAGIARISM CHECK RESULT



aall ch

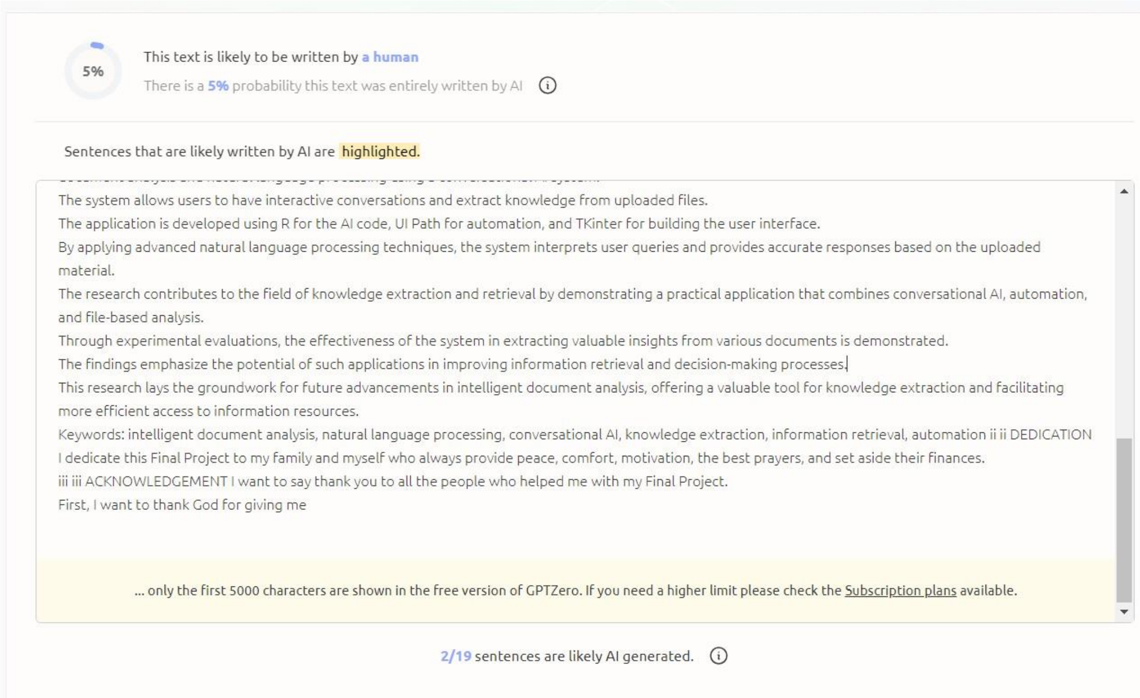
ORIGINALITY REPORT

7%	7%	2%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

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

GPTZERO RESULT



5%

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

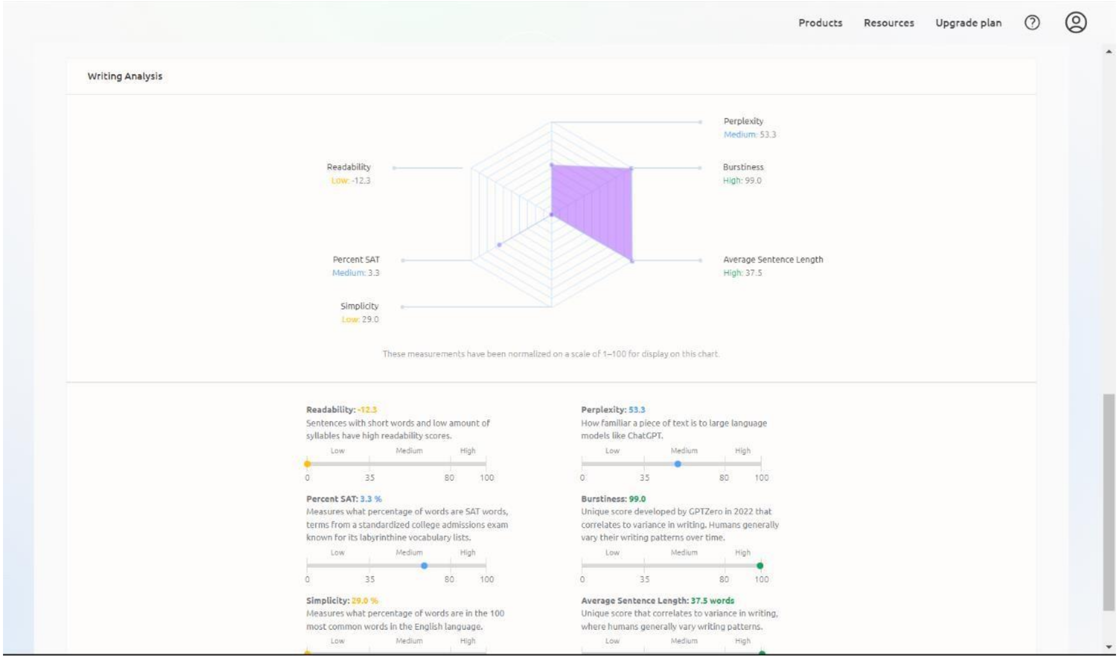
There is a **5%** probability this text was entirely written by AI

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.



These measurements have been normalized on a scale of 1–100 for display on this chart.

Readability: -12.3
Sentences with short words and low amount of syllables have high readability scores.

Perplexity: 53.3
How familiar a piece of text is to large language models like ChatGPT.

Percent SAT: 3.3 %
Measures what percentage of words are SAT words, terms from a standardized college admissions exam known for its labyrinthine vocabulary lists.

Burstiness: 99.0
Unique score developed by GPTZero in 2022 that correlates to variance in writing. Humans generally vary their writing patterns over time.

Simplicity: 29.0 %
Measures what percentage of words are in the 100 most common words in the English language.

Average Sentence Length: 37.5 words
Unique score that correlates to variance in writing, where humans generally vary writing patterns.

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

LIST OF FIGURES

LIST OF TABLE