**A WEB-BASED TEXT/DOCUMENT SUMMARIZING AND TOPIC MODELING SYSTEM: AN AI (NLP) APPROACH TO SUMMARIZING TEXT**

**BY:**

**EBEREONWU EINSTEIN MUNACHISO 19/0512**

**SHOYEMI OLASUBOMI TUTULORO 19/1167**

**UDO DAVID DAVID 19/1326**

**A PROJECT WORK SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF BACHELOR OF SCIENCE B.Sc. (HONS) IN SOFTWARE ENGINEERING**

**SUBMITTED TO:**

**THE DEPARTMENT OF SOFTWARE ENGINEERING**

**SCHOOL OF COMPUTING AND ENGINEERING SCIENCES**

**BABCOCK UNIVERSITY, ILISAN REMO**

**OGUN STATE, NIGERIA**

**MAY 2023**

**DECLARATION**

We declare that the project work, “A WEB-BASED TEXT/DOCUMENT SUMMARIZING AND TOPIC MODELING SYSTEM: AN AI (NLP) APPROACH TO SUMMARIZING TEXT” was carried out by the following people

**EBEREONWU EINSTEIN MUNACHISO 19/0512 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**SIGNATURE/ DATE**

**SHOYEMI OLASUBOMI TUTULORO 19/1167 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**SIGNATURE/ DATE**

**UDO DAVID DAVID 19/1326 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**SIGNATURE/ DATE**

**CERTIFICATION**

This is to certify that this project titled“A WEB-BASED TEXT/DOCUMENT SUMMARIZING AND TOPIC MODELING SYSTEM: AN AI (NLP) APPROACH TO SUMMARIZING TEXT” was carried out by the following students under the supervision of the Department of Software Engineering, Babcock University, Ilishan-Remo, Ogun State, Nigeria:

**EBEREONWU EINSTEIN MUNACHISO 19/0512** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**SHOYEMI OLASUBOMI TUTULORO 19/1167** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**UDO DAVID DAVID 19/1326** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**DR. ADIGUN TAIWO DATE**

**PROJECT SUPERVISOR**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**EXTERNAL EXAMINER DATE**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**DR. ADEKOLA OLUBUKOLA DATE**

**(HEAD OF SOFTWARE ENGINEERING DEPARTMENT)**

**DEDICATION**

This project is dedicated to God almighty, who has led us through our four years in his own university; Babcock and has granted us the grace to produce this project which is a proof of knowledge gained at our time in the university. We also dedicate this project to all our parents, family and friends who have added to our growth and knowledge in one way or the other throughout our 4 years of studies.

**ACKNOWLEDGEMENT**

We thank God for his guidance, protection and intervention in our lives throughout our journey in Babcock University and especially for the grace to overcome all challenges faced in our every endeavors. We sincerely appreciate our supervisor Dr. Adigun Taiwo for all his guidance and support throughout the creation of this project.

We are truly grateful to all our lecturers who have impacted in us, the knowledge and skills which we put into practice during the several stages of our project build. Lecturers such as Dr. Adigun Taiwo who is also our project supervisor, Dr. Adetunji Oluwatofunmi our able course advisor who taught us “Software Security Engineering, Introduction to Professional Ethics and Practice”, Professor Sunday Idowu who taught us “Introduction to Computer Science and Programming”, Dr. Maitanmi S. who taught us “Introduction to Web Technology and development”, Mr. Otuneme who taught us “Object Oriented Software Development”, Dr. Jet Akinsola who taught us “Algorithms and Data Structures”, Dr. Wunmi Ajayi who taught us “Software Requirements Engineering and Construction and Software Engineering Economics”, and all other lecturers whose names are not mentioned.

We also give a special thanks to our respective parents Dr. and Mrs. Ebereonwu, Mr. and Mrs. Udena, Mr. and Mrs. Shoyemi for their continual dedication to us and support throughout.

**ABSTRACT**

Students, lecturers, accountants, businessmen & women, politicians, researchers, and many more individuals of various backgrounds not listed above often have to read through large volumes of information or produce a summarized version of documents or other bodies of text. Such a task can easily be achieved with small volumes of text but becomes a hassle when individuals have to read through and summarize large volumes of text. For such a reason, the system “A WEB-BASED TEXT/DOCUMENT SUMMARIZING AND TOPIC MODELING SYSTEM: AN AI (NLP) APPROACH. TO SUMMARIZING TEXT” was built. It provides a top-notch topic modeling feature where the system detects the topic of the body of text and displays it to the user, and a language detection feature where the system can identify the language of the text entered before providing an appropriate summary. In an event where a language is selected, yet the text entered into the system is of a different language, it displays an alert asking the user to choose the appropriate language.

It focuses on a particular field of AI called Natural Language Processing (NLP) for the summarization and language detection aspect of the system, the UI/UX was designed using Figma and implemented using HTML, CSS, and JavaScript, and the server side was built with Python using the Flask framework.

The project initially accepted only written texts, could only summarize English text and didn’t have the language detection feature. However, new feature recommendations were made and were added: document summarizing. This feature allows a user to select a word document

(.docx) or a text document (.txt) file and summarizes the content of such file. A download summary feature lets a user download the summary as a word document (.docx). A language detection/ language modeling feature allow user detect what language the body of text for summary is written in and finally, a summarize to another language feature allows users to summarize the entered text and get the summary in another language.

**Keywords: AI, NLP, Topic Modeling, Language Detection, Language Modeling.**

**Word count: 338.**

**TABLE OF CONTENT**

DECLARATION…………………………………………………………………...i

CERTIFICATION………………………………………………………………....ii

DEDICATION…………………………………………………………………….iii

ACKNOWLEDGEMENT…………………………………………………………iv

ABSTRACT……………………………………………………………………......v

LIST OF FIGURES……………………………………………………………….vi

1. CHAPTER ONE: INTRODUCTION..................................................................1
   1. BACKGROUND OF STUDY……………………………………………..1
   2. STATEMENT OF PROBLEM…………………………………………….2
   3. AIM AND OBJECTIVES OF THE STUDY…….………………………...3
   4. METHODOLOGY………………………………………………………....4
   5. SCOPE OF STUDY………………………………………………………..4
   6. SIGNIFICANCE OF STUDY……………………………………………...5
   7. RESEARCH JUSTIFICATION…………………………………………....5
   8. DEFINITION OF TERMS………………………………………………....6
   9. ORGANIZATION OF THE PROJECT…………………………………….7
2. CHAPTER TWO: LITERATURE
   1. REVIEW……………………………………………………………………

CHAPTER THREE: SYSTEM ANALYSIS AND DESIGN

CHAPTER FOUR: IMPLEMENTATION

CHAPTER FIVE: SUMMARY, CONCLUSION AND RECOMMENDATION….

REFERENCES

APPENDICES

**CHAPTER ONE: INTRODUCTION**

1. **BACKGROUND OF THE STUDY**

In this modern world, technology is constantly increasing the efficiency, feasibility and ease of ways to carry out time consuming and brain tasking activities. With the increasing amount of information to consume and limited time to do so, extracting the key information embedded in large body of text has become almost unachievable for individuals from various backgrounds be it education, medicine, tourism, law, and many more.

However, because of technological advancements and the existence of Artificial Intelligence (AI), with a focus on Natural Language Processing (NLP), the work of summarizing has become a very seamless task; systems are now very capable of churning large bodies of text while also withholding the semantics. Individuals used to have to read through multiple lines of text, absorb and digest the information it contained before proceeding to write a summary of essential points included in the original body of text or take any more actions, but due to the advancements of AI, this is no longer the case.

Artificial intelligence (AI) is a field of research that includes technology emulating or attempting to clone the intellectual quotient of humans. Natural language processing, computer vision, self-driving vehicles, fraud detection and prevention, and many more applications can benefit from AI. AI as a subject of study may be traced back to a symposium held at Dartmouth in Hanover, New Hampshire in 1956. Although the goal to create an artificially intelligent system had been devised, it was not simple to implement, which lead to

a drop in government financing, this resulted in a period known as the "AI winter," which lasted from 1970 to 1980. Following many AI achievements, such as when IBM's Deep Blue became the first computer to defeat a chess champion, and when an AI system won the game show "Jeopardy," research began to accelerate. Ever since then, AI has been found useful in several fields of study and will continue to remain relevant. As earlier stated, this project is made possible due to the NLP branch of AI.

Natural Language Processing (NLP) is a subfield of Artificial Intelligence which came into existence in the middle of 20th century it enables computers to interpret spoken words or written texts in a way comparable to that of humans. NLP has been used to achieve several ground-breaking achievements including but not limited to text translation from one language to another, development of chat bots, text summarization as in this case, and personal-assistants such as Apple's Siri, Amazon's Alexa, and Google's Google Assistant, which are capable of having real-time non-human controlled conversations with individuals, carrying out specific tasks such as setting a reminder for 12:30, calling a friend, texting a friend, and much more can all be achieved simply by asking a home assistant which are able to understand humans due to their AI integration.

**1.2 STATEMENT OF PROBLEM**

With the advent of computer literacy, we have so much information in circulation. Ranging from phones, texts, social media platforms, law enforcement agencies, churches, school study materials etc. Looking at the school study materials for instance, there’s usually a tremendous amount of course load to read and limited time to do so, to enable them take tests

and exams in turn. We as academics already know that not every detail in the material is usually relevant. Although it’s important to look closely to find the relevant details. This project is geared at providing a long-term solution of extracting the key underlying information embedded in this heavily detailed content made for consumption. In the sense the material is fed into the system by either copying and pasting, manual upload or drag and drop, the system then assimilates and digests the content with respect to time and space and then extracts the key information necessary for consumption by the reader. Moreso, the system also enables auto-detection of the language of the said body of text, summarize to other languages, detect topic of discussion of any text entered for summary. It removes the constraint of solely English language summarization as other systems have and puts the choice of language in the hands of the user.

* 1. **AIM AND OBJECTIVES OF THE STUDY**

The aim of this study is to develop and implement a web-based text/document summarizing and topic modeling system using an AI (NLP) approach to summarizing text. The objectives of the above-mentioned system are to:

1. Be able to give comprehensive and accurate summaries of text provided in it for summarization.
2. Be able to detect the subject/ topic of the provided body of text.
3. Be able to detect the language the provided body of text is written in.
4. Be able to summarize text to another language.

**1.4 METHODOLOGY**

This study uses the methodologies below to achieve the aforementioned objectives:

1. An intense and critical analysis of already existing closely related software systems was done in order to aid in the design of the system from a user’s point of view.
2. The prototyping Software Development Life Cycle (SDLC) model was adopted as user input was heavily required in order to improve software quality.
3. The web app will be divided into server (to be built using Flask) and client side (with HTML, CSS and JavaScript.)
4. A Machine Learning mode is to be trained to detect languages and return the result as an end point.
5. An NLP system is to detect the most popular non-stop word and return that as the topic of the body of text.

**1.5 SCOPE OF STUDY**

This study focuses on developing a web-based text/document summarizing and topic modeling system: an AI (NLP) approach to summarizing text. The system will also be able to provide the topic of a body of text, detect text language, and summarize to other languages. However, the project scope will be limited to the following:

1. Summarizing documents: It will summarize documents written in a few languages, not all.
2. Detecting languages: It will detect a select few numbers of languages; languages that were used to train the model for its detection
3. Topic modeling: It will analyze and provide the topic/ subject matter of the body of text for only a reasonable amount of text.
4. Only written text, .txt files and .docx file contents can be summarized directly by the system.
5. Summaries can only be downloaded as .docx files.

**1.6 SIGNIFICANCE OF STUDY**

The study focuses on developing an intelligent system that is able to quickly produce un-biased, un-ambiguous and relevant summaries (in a select few languages) for individuals at any point in time, show individuals the topic of discussion of any body of text in order to aid individuals decide if it is truly worth reading or not and finally provide individuals the ability to detect the language in which a given body of text is written in.

Thanks to the advent of AI, the project which is of immerse importance to academics, research and other fields will significantly reduce the amount of time it takes to summarize any given body of text therefore providing individuals key information alone and giving them more time to focus on the necessary information and be more efficient in carrying out the intended post summary activities with the key information.

**1.7 RESEARCH JUSTIFICATION**

Advantages of the proposed system over the previous system are as follows:

1. The system allows for language detection
2. The system allows for cross-language summarization
3. The system provides top 2 possible topics of the body of text.
4. The system allows for summarization of larger body of text in free version as compared to other existing systems.

**1.8 DEFINITION OF TERMS**

**AI:** Short for Artificial Intelligence is the ability of a machine to perform tasks thought to require human intelligence. (Encyclopedia Britannica, 2021)

**NLP:** Short for Natural language processing refers to the branch of computer science and more specifically, the branch of artificial intelligence concerned with giving computers the ability to understand text and spoken words in much the same way human beings can. (IBM Cloud Education, 2020)

**HTML:** Short for Hypertext Markup Language, is the standard markup language for documents designed to be displayed in a web browser. It can be assisted by technologies such as Cascading Style Sheets (CSS) and scripting languages such as JavaScript. Web browsers receive HTML documents from a web server or from local storage and render the documents into multimedia web pages. HTML describes the structure of a web page semantically and originally included cues for the appearance of the document. (Wikipedia, 1993)

**CSS:** Short for Cascading Style Sheets is a style sheet language which is used to describe the look and formatting of a document written in markup language. It provides an additional feature to HTML. It is generally used with HTML to change the style of web pages and user

interfaces. It can also be used with any kind of XML documents including plain XML, SVG and XUL.

**JavaScript: J**avaScript is a scripting language for creating dynamic web page content. It creates elements for improving site visitors’ interaction with web pages, such as dropdown menus, animated graphics, and dynamic background colors. (**Jordana A., 2022**)

**Python:** Python is a high-level programming language designed by Guido van Rossum which supports multiple programming paradigms such as Object-Oriented Programming (OOP), procedural programming. It can be used for several activities such as backend development, data analytics, machine learning, general scripting tasks and virtually anything one can think of.

**Flask:** Flask is what’s known as a WSGI framework. Mercifully pronounced “whiskey,” this stands for ​​Web Server Gateway Interface. Essentially, this is a way for web servers to pass requests to web applications or frameworks. Flask relies on the WSGI external library to function, as well as the Jinja2 template engine. (Derry, 2022)

**1.9 ORGANIZATION OF THE PROJECT**

This chapter has successfully introduced the web-based text/document summarizing and topic modeling system: an AI (NLP) approach to summarizing text including its inner workings, aims and objectives, methodology, scope of study and overall, the significance of the study.

**Chapter two** covers the literature review, it acknowledges previously existing work related to this study, their literature review, cited works and the shortcomings of the related studies including how this research can improve on such shortcomings.

**Chapter three** expatiates on the systems design methodology, its functional, non-functional, user and system requirements, logic flow which includes all necessary diagrams, design and finally, development tools used to build the system

**Chapter four** contains everything that has to do with system design implementation, testing strategies, minimum system requirements, software maintenance and evolution and a lot more

**Chapter five** discusses the summary, recommendations and conclusion aspect of the research work including how it can be applied in real world scenarios

**REFERENCES**

Burns. (2022, February 1). What is Artificial Intelligence (AI)? Definition, Benefits and Use Cases.

SearchEnterpriseAI. Retrieved September 30, 2022, from <https://www.techtarget.com/searchenterpriseai/definition/AI-Artificial-Intelligence>.

Lewis. (2014, December 4). A Brief History of Artificial Intelligence | Live Science. livescience.com.

Retrieved September 30, 2022, from <https://www.livescience.com/49007-history-of-artificial-intelligence.html>

Education, I. C. (2020, July 2). What is Natural Language Processing? What Is Natural Language

Processing? | IBM. Retrieved September 30, 2022, from <https://www.ibm.com/cloud/learn/natural-language-processing#:~:text=Natural%20language%20processing%20(NLP)%20refers,same%20way%20human%20beings%20can>.

Menzli, A. (2021, July 28). 10 NLP Projects to Boost Your Resume - neptune.ai. neptune.ai. Retrieved

September 30, 2022, from <https://neptune.ai/blog/10-nlp-projects/amp>

Britannica, The Editors of Encyclopaedia. "artificial intelligence summary". Encyclopedia Britannica,

24 Jul. 2021, <https://www.britannica.com/summary/artificial-intelligence/>. Accessed 1 October 2022.

Education, I. C. (2020, July 2). What is Natural Language Processing? What Is Natural Language

Processing? | IBM. Retrieved October 1, 2022, from <https://www.ibm.com/cloud/learn/natural-language-processing>

(1993, January 1). HTML - Wikipedia. HTML - Wikipedia. Retrieved October 1, 2022, from

<https://en.wikipedia.org/wiki/HTML>

A., J. (2021, August 3). What Is JavaScript? A Basic Introduction to JS for Beginners. Hostinger

Tutorials. Retrieved October 2, 2022, from <https://www.hostinger.com/tutorials/what-is-javascript>.

DEERY. (2022, August 1). The Flask Web Framework: A Beginner’s Guide. CareerFoundry. Retrieved

October 2, 2022, from <https://careerfoundry.com/en/blog/web-development/what-is-flask/>