Faculty of Computer Science and Information Technology

WEB-BASED ARTICLE SUMMARIZATION WITH MACHINE LEARNING TECHNIQUES

Lim Wu Tong

Bachelor of Computer Science with Honours (Software Engineering)

2023

UNIVERSITI MALAYSIA SARAWAK

THESIS STATUS ENDORSEMENT FORM

TITLE Web-based Article Summarization with Machine Learning Techniques

ACADEMIC SESSION: 2022/2023

LIM WU TONG

(CAPITAL LETTERS)

hereby agree that this Thesis* shall be kept at the Centre for Academic Information Services, Universiti Malaysia Sarawak, subject to the following terms and conditions:

1. The Thesis is solely owned by Universiti Malaysia Sarawak
2. The Centre for Academic Information Services is given full rights to produce copies for educational purposes only
3. The Centre for Academic Information Services is given full rights to do digitization in order to develop local content database
4. The Centre for Academic Information Services is given full rights to produce copies of this Thesis as part of its exchange item program between Higher Learning Institutions [or for the purpose of interlibrary loan between HLI]
5. ** Please tick (√)

□ CONFIDENTIAL (Contains classified information bounded by the OFFICIAL SECRETS ACT 1972)

□ RESTRICTED (Contains restricted information as dictated by the body or organization where the research was conducted)

☑ UNRESTRICTED

Validated by

(AUTHOR'S SIGNATURE) (SUPERVISOR'S SIGNATURE)

Permanent Address

1, TAMAN KEMUDI
JALAN KUALA KEDAH
06600 ALOR SETAR, KEDAH

Date: 01/07/2023 Date: ____________

Note * Thesis refers to PhD, Master, and Bachelor Degree
** For Confidential or Restricted materials, please attach relevant documents from relevant organizations / authorities

WEB-BASED ARTICLE SUMMARIZATION WITH MACHINE LEARNING TECHNIQUES

LIM WU TONG

This project is submitted in partial fulfilment of the requirements for the degree of Bachelor of Computer Science with Honours (Software Engineering)

Faculty of Computer Science and Information Technology

UNIVERSITI MALAYSIA SARAWAK

2023

RINGKASAN ARTIKEL BERASASKAN WEB DENGAN TEKNIK PEMBELAJARAN MESIN

LIM WU TONG

Projek ini merupakan salah satu keperluan untuk Ijazah Sarjana Muda Sains Komputer dengan Kepujian (Kejuruteraan Perisian)

Fakulti Sains Komputer dan Teknologi Maklumat

UNIVERSITI MALAYSIA SARAWAK

2023

DECLARATION

I hereby declare that this project is my original work. I have not copied from any other student's work or from any other sources except where due reference or acknowledgement is not made explicitly in the text, nor has any part had been written for me by another person.

……………………………….

(LIM WU TONG) 01 JULY 2023

Matric No: 72789

ACKNOWLEDGEMENT

Firstly, I would like to express my deep appreciation to my supervisor, Dr Mohammad bin Hossin, who has provided invaluable assistance, guidance, and advice throughout the development of this project. His expertise and support have been instrumental in helping me to complete my work. At the same time, I would also like to thank my examiner, Prof. Dr Narayanan Kulathuramaiyer, for his valuable comments and suggestions that have helped to improve the quality of my project.

I would also like to acknowledge the help provided by Prof. Dr Wang Yin Chai, the final year project coordinator, who has given me the necessary information and guidance to complete my final year project. In addition, I would like to thank all my friends and coursemates who have supported me and provided me with the information needed to finish this project.

Finally, I am especially grateful to my parents and siblings for their unwavering support and encouragement. Their love and encouragement have given me the strength and motivation to complete this project. Without their help, this project would not have been possible.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

The motivation behind this project is the increasing amount of information available on the internet, which makes it difficult for people to sift through and find the relevant information they need. Text summarization can help to address this problem by condensing lengthy texts into shorter summaries that convey the main points and ideas of the original text. However, traditional text summarization methods often produce summaries that are too short or lack coherence, which can make them difficult to understand. Machine learning techniques have the potential to overcome these limitations and produce more accurate and coherent summaries. In order to develop the web-based article summarization system, various machine learning techniques were studied and compared. The Naive Bayes, Neural Network, and decision tree techniques were chosen for their ability to handle both numerical and categorical data, and their robustness to noise and missing values. These techniques were implemented using the Python programming language and the scikit-learn library. The front-end of the system was developed using the Django framework, along with HTML, CSS and JavaScript for styling and interactive elements. The performance of the system was evaluated using a dataset of articles and their corresponding summaries. The quality of the summaries was assessed using metrics such as ROUGE and expert evaluation, while the preferredness were evaluated through user surveys and time efficiency observed from the system. The results showed that the system was able to produce summaries that were of good quality, preferred by users, and efficient in terms of time. Overall, the web-based article summarization system with machine learning techniques demonstrated the potential to be a useful tool for condensing and summarizing texts in a more accurate and coherent manner.

ABSTRAK

Motivasi di sebalik projek ini adalah peningkatan jumlah maklumat yang tersedia di internet, yang menyukarkan orang ramai untuk menyaring dan mencari maklumat berkaitan yang mereka perlukan. Rumusan teks boleh membantu menangani masalah ini dengan memekatkan teks yang panjang lebar menjadi ringkasan yang lebih pendek yang menyampaikan perkara utama dan idea teks asal. Walau bagaimanapun, kaedah ringkasan teks tradisional sering menghasilkan ringkasan yang terlalu pendek atau kurang koheren, yang boleh menyukarkannya untuk difahami. Teknik pembelajaran mesin mempunyai potensi untuk mengatasi batasan ini dan menghasilkan ringkasan yang lebih tepat dan koheren. Untuk membangunkan sistem ringkasan artikel berasaskan web, pelbagai teknik pembelajaran mesin telah dikaji dan dibandingkan. Teknik Naive Bayes, Neural Network dan Decision Tree dipilih kerana keupayaannya mengendalikan kedua-dua data berangka dan kategori, serta keteguhannya terhadap bunyi dan nilai yang hilang. Teknik-teknik ini telah dilaksanakan menggunakan bahasa pengaturcaraan Python dan scikit-learn. Bahagian hadapan sistem telah dibangunkan menggunakan rangka kerja Django, bersama-sama dengan HTML, CSS dan JavaScript untuk penggayaan dan elemen interaktif. Prestasi sistem telah dinilai menggunakan set data artikel dan ringkasan yang sepadan. Kualiti ringkasan dinilai menggunakan metrik seperti ROUGE dan penilaian pakar, manakala keutamaan dinilai melalui tinjauan pengguna dan kecekapan masa diperhatikan daripada sistem. Hasil kajian menunjukkan sistem ini mampu menghasilkan ringkasan yang berkualiti, digemari pengguna, dan cekap dari segi masa. Secara keseluruhannya, sistem ringkasan artikel berasaskan web dengan teknik pembelajaran mesin menunjukkan potensi untuk menjadi alat yang berguna untuk memekatkan dan meringkaskan teks dengan cara yang lebih tepat dan koheren.

Chapter 1: Introduction

1.1 Preliminaries

A research article presents the results of original research, evaluates its contribution to the corpus of knowledge in a specific discipline, and is published in a peer-reviewed scholarly journal. The publication of university professors' research plays a crucial part in deciding whether they are given tenure (Hall, 2017). Research article were primarily read by other researchers and students. Research articles are crucial for researchers and students nowadays. Project preparation is the most common reason for students to read research papers (Akmal, Dhivah, & Mulia, 2020). Students and researcher working on a project must read lengthy research articles in order to choose their topics and conduct their research.

In the age of the technology, a vast volume of research articles has produced by researchers and students. Hence, article summarization is important for research and student while doing their research study. One kind of information management is summarization (Jones, 1993). A summary is defined as a text that is derived from one or more texts, provides essential information from the original texts, and is no longer than half the length of the original texts, and frequently much shorter. (Allahyari, Pouriyeh, Assefi, Safaei, Trippe, Gutierrez, & Kochut, 2017). Summarization is the process of creating a shortened rendition of a lengthy literary work. It extracts the most essential information from the original text and eliminates irrelevant details and data. A summarization helps you to understand the article better (Ansari, 2022). People may quickly and readily understand a text without having to read it in its entirety if a brief summary is provided (Chuang, & Yang, 2000). It will be simpler to understand what the article is about and to understand the author's main points after reading through the article's summary. However, manual article summary might be inefficient and inconvenient due to its time-consuming and would need more work force to complete it. Therefore, this Web-based Article Summarization system is introduced in this project.

By using this Web-based Article Summarization system, any lengthy article may be summarized into a short and understandable format. This approach employs machine learning techniques to improve the precision and speed of the summarization process. Through this system, researchers, students, and even professors may save time and obtain a deeper grasp of the article, allowing them to do their study more efficiently.

1.2 Problem Statement

Students engaged in research are required to read research articles in order to choose topics and perform experiments (Subramanyam, 2013). Knowing what has been discovered and what questions remain may help when planning a research project. Therefore, article summary may help in the management of the article's content. When manually summarizing an article, the process usually starts with reading many times and paying close attention to focus on the main argument of the article. The length of the summary article must then be determined by extracting the main concept and any relevant supporting details. However, the process of manually article summarization will cause the time consuming and gain the high workload for students and researchers. The average researcher spends 49–61 hours per year reading articles, with an average of 20 minutes each reading and an estimated 145–184 reads per year (Tenopir, King, Clarke, Na, & Zhou, 2007). A typical researcher probably spends hundreds of hours per year reading articles (Nüst, Boettiger, & Marwick, 2018). Therefore, a web-based system is required to automatically summarise articles so that students and researchers can spend less time on the task.

Furthermore, numerous researches have shown strong evidence that distracted reading has a negative impact on the learning process (Schmidt, 2020). When reading and summarizing the article, some students and researchers are not focusing on the most important aspect of the information being discussed in the article. This will actually occur since the they were not paying attention, and they were also distracted by other details in the article that were not important. The difficulty of reading papers from 2015 is greater than that of those from the 1900s, and the issue is not related to words (Ball, 2017). Because of the difficulty of the article, it's possible that students and researchers may have difficulty understanding or will misunderstand the article's main idea when they read it or attempt to summarize it. Therefore, the web-based article summarization system will include machine learning techniques to provide a more accurate and precise summary of the research article.

In addition, the main concept presented in a single research article is the most challenging part to understand, and it is challenging to summarize it manually. When reading a research article, students and researchers may get discouraged to read it and attempt to summarize it since it contains unfamiliar terminology, aspects, and even methodologies. Therefore, different machine learning techniques will be included into the web-based article summarization system in order to provide users with the option of selecting their preferred article summarize result after processing by the different machine learning techniques.

1.3 Project Objectives

The Project aims to design and develop a web-based article summarization system.

The objectives of the project are:

i. To design a web-based system which integrates machine learning algorithm for article summarization.

ii. To implement the proposed web-based system integrating with program-based Python machine learning techniques to produce article summaries.

iii. To analyse the effectiveness and efficiency of proposed machine learning techniques based on summarization quality, preferredness and time efficiency.

1.4 Scope

The system is available to all users. However, the focus of this project's development will be on assisting final year students who need to do research for their final-year projects. By using this system, users will speed up their research and gain a deep understanding of the article they research.

1.5 Significant of project

The proposed web-based system for summarizing articles has the potential to greatly benefit a variety of users. Students, researchers, and professors alike may find it useful when conducting research and reviewing articles. By providing a short summary of the research, the system aims to reduce the time spent on manual summarization and improve the user's understanding of the article. This can increase efficiency during the research process.

In addition to its potential benefits for individual users, this system may also have a broader impact on the field of research as a whole. By streamlining the review process and allowing for more efficient analysis of articles, the system has the potential to facilitate faster and more comprehensive understanding of current developments in a given field.

Overall, the development of this web-based system for summarizing articles has the potential to greatly benefit both individual researchers and the research community as a whole. By providing a quick and easy way to review and understand research articles, the system has the potential to improve the efficiency and effectiveness of the research process.

1.6 Project Outcome

The functional web-based system that can summarize articles. By using this system, the uploaded article will be shortened and simplified so that it is short and easy to understand. Other from that, the system will use 3 different summarization machine learning algorithms to allow users to choose the quality, preferredness, and efficiency of the system's summarization.

1.7 Thesis Organization

Six chapters make up the report: Introduction, Literature Review, Methodology, Implementation, Evaluation, Testing, and Results, and Conclusion and Future Work.

1.7.1 Chapter 1: Introduction

This chapter presents an overview of the undertaking. This chapter discusses the problem statement, objectives, scope, significance of the project, project schedule, and anticipated outcome. The problem statement analyses the current problem at hand. The objectives define the requirements of the project, whereas the scope specifies the limitations or range to be covered during the two semesters of Final Year Project.

1.7.2 Chapter 2: Literature Review

Chapter 2 provides a literature review of the existing and similar systems. This chapter will analyse the existing system's constraints. The comparative analysis of the existing systems is followed by a review of the proposed system's advantages. The entire review of existing systems is conducted using reliable sources such as articles, journals, etc. This chapter will also provide a brief description of the project's implementation, including the algorithm, techniques, and technology used.

1.7.3 Chapter 3: Methodology

Chapter 3 focuses on the methodology used in the project's development. This chapter also includes the requirement analysis, in which the user and system needs are described in detail. In addition, the system design phase will be included at the end of this chapter, which acts as the blueprint for the whole project.

1.7.4 Chapter 4: Implementation

Chapter 4 describes the proposed system's implementation in detail. The proposed system's structure and prototype are discussed in detail.

1.7.5 Chapter 5: Evaluation, Testing and Result

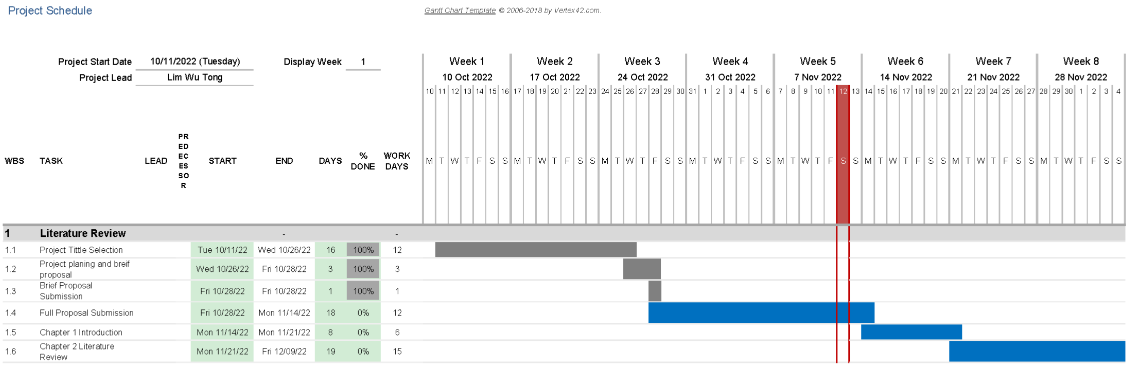
Chapter 5 begins with an introduction to the evaluation and testing phase, providing an overview of the testing plan and strategy. The chapter concludes with a summary of the testing phase and its impact on the system's functionality.

1.7.6 Chapter 6: Conclusion and Future Works

Chapter 6 is the last chapter of the project and provides a summary of the whole project. This chapter provides a summary of the project's achievements, limitations, and suggestions for future enhancements.

1.8 Project Schedule

The development of a successful project mainly depends on the use of a precise project schedule. The identifying of the project milestone serves as a guideline for determining whether or not the project can be completed on time. The completion of this project requires a total of two semesters. The Gantt Chart was chosen as the tool to be used for arranging the various stages of the project's progression.

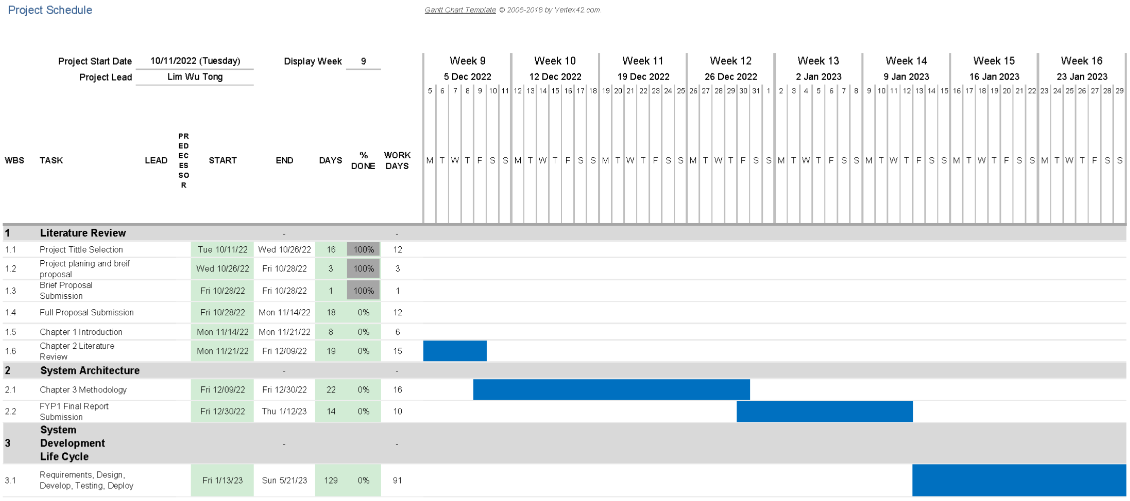


Project Schedule

Project Start Date 10/11/2022 (Tuesday)
Project Lead Lim Wu Tong
Display Week 1

WBS	TASK	LEAD	PREDECESSOR	START	END	DAYS	% DONE	WORK DAYS
1	**Literature Review**				-			-
1.1	Project Tittle Selection			Tue 10/11/22	Wed 10/26/22	16	100%	12
1.2	Project planing and breif proposal			Wed 10/26/22	Fri 10/28/22	3	100%	3
1.3	Brief Proposal Submission			Fri 10/28/22	Fri 10/28/22	1	100%	1
1.4	Full Proposal Submission			Fri 10/28/22	Mon 11/14/22	18	0%	12
1.5	Chapter 1 Introduction			Mon 11/14/22	Mon 11/21/22	8	0%	6
1.6	Chapter 2 Literature Review			Mon 11/21/22	Fri 12/09/22	19	0%	15

Figure 1. 1 Gantt Chart Week 1-8



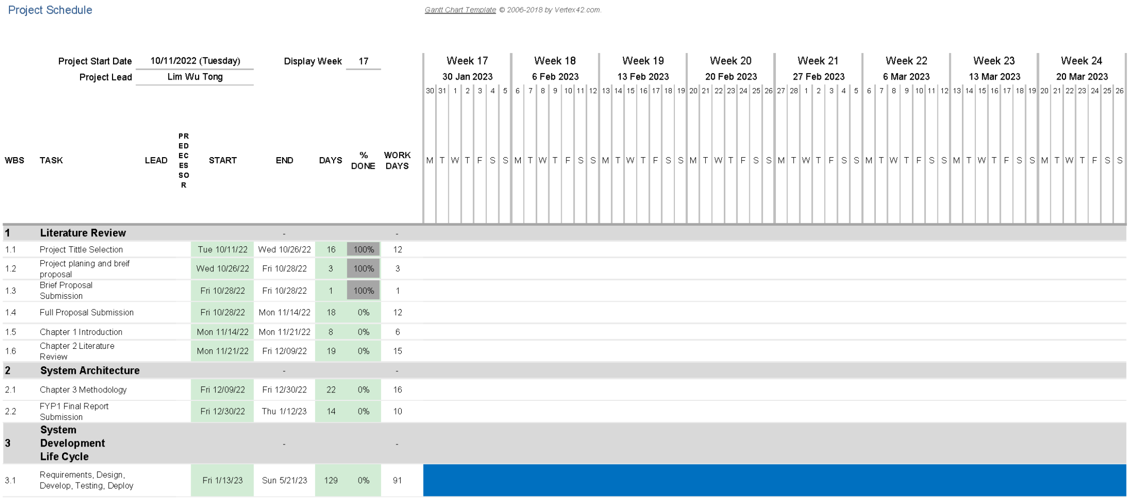
Project Schedule

Gantt Chart Template © 2006-2018 by Vertex42.com

Project Start Date	10/11/2022 (Tuesday)	Display Week	9
Project Lead	Lim Wu Tong		

WBS	TASK	LEAD	PREDECESSOR	START	END	DAYS	% DONE	WORK DAYS
1	**Literature Review**				-			-
1.1	Project Tittle Selection			Tue 10/11/22	Wed 10/26/22	16	100%	12
1.2	Project planing and breif proposal			Wed 10/26/22	Fri 10/28/22	3	100%	3
1.3	Brief Proposal Submission			Fri 10/28/22	Fri 10/28/22	1	100%	1
1.4	Full Proposal Submission			Fri 10/28/22	Mon 11/14/22	18	0%	12
1.5	Chapter 1 Introduction			Mon 11/14/22	Mon 11/21/22	8	0%	6
1.6	Chapter 2 Literature Review			Mon 11/21/22	Fri 12/09/22	19	0%	15
2	**System Architecture**				-			-
2.1	Chapter 3 Methodology			Fri 12/09/22	Fri 12/30/22	22	0%	16
2.2	FYP1 Final Report Submission			Fri 12/30/22	Thu 1/12/23	14	0%	10
3	**System Development Life Cycle**				-			-
3.1	Requirements, Design, Develop, Testing, Deploy			Fri 1/13/23	Sun 5/21/23	129	0%	91

Figure 1. 2 Gantt Chart Week 9-16



Project Schedule

Gantt Chart Template © 2006-2018 by Vertex42.com

Project Start Date	10/11/2022 (Tuesday)	Display Week	17
Project Lead	Lim Wu Tong		

WBS	TASK	LEAD	PREDECESSOR	START	END	DAYS	% DONE	WORK DAYS
1	**Literature Review**				-			-
1.1	Project Tittle Selection			Tue 10/11/22	Wed 10/26/22	16	100%	12
1.2	Project planing and breif proposal			Wed 10/26/22	Fri 10/28/22	3	100%	3
1.3	Brief Proposal Submission			Fri 10/28/22	Fri 10/28/22	1	100%	1
1.4	Full Proposal Submission			Fri 10/28/22	Mon 11/14/22	18	0%	12
1.5	Chapter 1 Introduction			Mon 11/14/22	Mon 11/21/22	8	0%	6
1.6	Chapter 2 Literature Review			Mon 11/21/22	Fri 12/09/22	19	0%	15
2	**System Architecture**				-			-
2.1	Chapter 3 Methodology			Fri 12/09/22	Fri 12/30/22	22	0%	16
2.2	FYP1 Final Report Submission			Fri 12/30/22	Thu 1/12/23	14	0%	10
3	**System Development Life Cycle**				-			-
3.1	Requirements, Design, Develop, Testing, Deploy			Fri 1/13/23	Sun 5/21/23	129	0%	91

Figure 1. 3 Gantt Chart Week 17-24