The Future of Ayurveda: Harnessing the Power of Artificial Intelligence

23-252

Project Proposal Report

Jayasinghe J.A.S.C – IT20216078

B.Sc. (Hons) Degree in Information Technology Specializing in Software Engineering

Department of Computer Science and Software Engineering

February 2023

Declaration

I declare that this is our own work, and this proposal does not incorporate without acknowledgment any material previously submitted for a degree or diploma in any other university or Institute of higher learning, and to the best of our knowledge and belief, it does not contain any material previously published or written by another person except where the acknowledgment is made in the text. Also, I hereby grant to Sri Lanka Institute of Information Technology, the nonexclusive right to reproduce and distribute my dissertation, in whole or in part, in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Name	Student ID	Signature
Jayasinghe J.A.S.C	IT20216078	

Abstract

Many existing social media platforms enable us to share knowledge and information in various domains. Due to generating massive data from such platforms, it faces many issues regarding user content management, such as maintaining the uniqueness and delivering personalized and timeliness content. Also, there is room for collecting knowledge and information relating to various domains from the massive data of such platforms. It is expected to address and explore more such problems and possibilities during the project by developing a social network limited to sharing only health-related content.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

BERT	Bidirectional Encoder Representations from Transformers
AWS	Amazon Web Services
CQRS	Command and Query Responsibility Segregation

Introduction

Background Literature

Social media analytics are warranted to provide empirical insights regarding various channels' credibility, recency, uniqueness, frequency, and salience [1].

Credibility often refers to the quality, trustworthiness, and integrity of information. Especially when it comes to health-related information, it is a sensitive point. The credibility of information in social media health content can vary greatly. On the other hand, social media can be a great source of information, providing access to health professionals, medical organizations, and peer-reviewed research, including ordinary people. Social media can also be a breeding ground for misinformation, rumors, and conspiracy theories. Hence, it's important to be cautious when evaluating health information on social media and to take the time to verify the credibility of the source. Some factors to consider when assessing the credibility of information on social media should be the source, evidence, accuracy, tone, and audience,

1. Source – Who is the author or organization behind the content? Are they reputable sources with expertise in the field?
2. Evidence – Is the information based on scientific evidence or personal opinion? Is the evidence presented in a clear and transparent way, with references to sources?
3. Accuracy – Is the information accurate and up to date? Are there any apparent errors or inconsistencies in the information?
4. Tone – Is the tone intended for a general audience, or is it targeted towards a specific group with a particular agenda or bias?

By carefully evaluating these factors, you can better assess the credibility of health information on social media and make informed decisions about your health.

The term recency often refers to the timeliness of the information. Health information is constantly evolving, and new research is continually being published. As a result, it's important to ensure that the information you're reading on social media is up-to-date and based on the latest research. When evaluating the recency of information in social media health content, consider the following:

1. Date of publication: When was the content published? Is it recent or outdated?
2. Updates: Has the content been updated to reflect new information or research? If so, how frequently is it updated?
3. Relevance: Is the information still relevant and applicable to current health practices and guidelines?
4. Sources: Are the sources cited in the content current and reputable?
5. Verification: Has the information been verified by other sources or health professionals?

By considering these factors, you can ensure that the health information you're consuming on social media is accurate, reliable, and up to date. It's also important to stay informed about the latest developments in the field by following credible health organizations and professionals on social media.

Uniqueness refers to whether the information presented on social media is novel or has been previously published or widely disseminated elsewhere. While unique information can be valuable and provide new insights, it can also be a red flag if the source of information needs to be more credible or if the claims made need to be supported by evidence. Therefore, verifying the source and credibility of unique information in social media health content is important. When evaluating the uniqueness of information in social media health content, consider the following:

1. Source: Who is the source of the information? Are they credible and reliable authorities in the field?
2. Evidence: Can evidence support the claims made in the unique information? Has the information been peer-reviewed or validated by other experts in the field?
3. Consistency: Is the unique information consistent with other credible sources of information in the field? If not, what are the reasons for the inconsistency?
4. Bias: Is the unique information presented unbiased and objectively, or is there a potential for bias or conflicts of interest?

Considering these factors, you can better assess the credibility and value of unique information in social media health content. It's important to approach unique information critically and seek additional sources and perspectives before deciding about your health.

The frequency of information refers to how often the content is posted, shared, or updated on social media. While frequent posting can indicate that the content is actively updated and relevant, it can also be a red flag if the information needs to be based on credible sources or evidence. Therefore, it's important to verify the source and credibility of information that is frequently posted on social media. When evaluating the frequency of information in social media health content, consider the following:

1. Source: Who is the source of the information? Are they credible and reliable authorities in the field?
2. Evidence: Can evidence support the claims made in the frequent posts? Has the information been peer-reviewed or validated by other experts in the field?
3. Consistency: Are the frequent posts consistent with other credible sources of information in the field? If not, what are the reasons for the inconsistency?
4. Bias: Are the frequent posts presented unbiased and objectively, or is there a potential for bias or conflicts of interest?
5. Relevance: Are the frequent posts relevant and applicable to current health practices and guidelines? Or are they promoting products or services that are not evidence-based?

Due to the high frequency of generating data, maintaining uniqueness has become a challenge for existing social media platforms. It's important to approach frequent posts critically and seek additional sources and perspectives before making decisions about your health. By considering these factors, you can better assess the value of information frequently posted on social media.

Salience refers to the information’s applicability to the individual and their specific health needs and circumstances. While social media can provide health information, not all may be relevant or applicable to an individual's health concerns or conditions. Therefore, seeking information that is salient and applicable to your needs is important. When evaluating the salience of information in social media health content, consider the following:

1. Relevance: Is the information directly relevant to your health concerns or conditions?
2. Applicability: Is the information applicable to your circumstances, such as age, gender, or health history?
3. Quality: Is the information high quality and based on credible sources and evidence?
4. Consistency: Is the information consistent with other credible sources of information in the field? If not, what are the reasons for the inconsistency?

5. Bias: Is the information presented unbiased and objectively, or is there a potential for bias or conflicts of interest?

By considering these factors, you can better assess the relevance and credibility of information in social media health content that is salient to your health needs. It's important to seek information from credible sources and consult with a healthcare professional before deciding about your health. It is possible to perform deeper natural language processing (NLP) to understand who, what, when, where, why, and how referencing salience. On the other hand, salience also leads to privacy concerns.

Research Problem

We have carried out user research to discover problems have to face in existing social media platforms when seeking health-related information. Participants of this user research have sound experience in various existing social media platforms and live in Sri Lanka's geographical area. The user research has been carried out as a questionnaire. While conducting user research, it identified existing platforms that are generally common in seeking health-related information as follows.

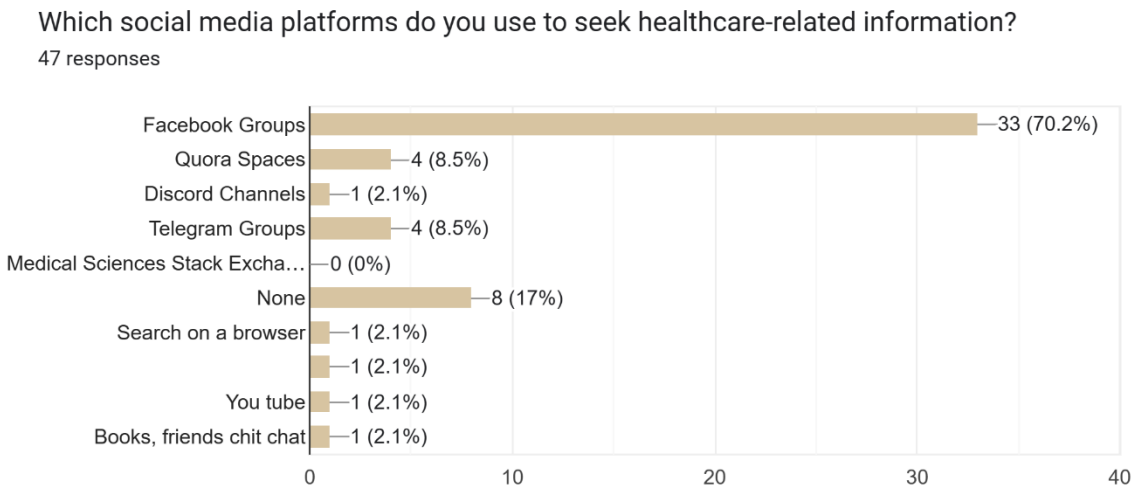

Figure 1: User Research Response

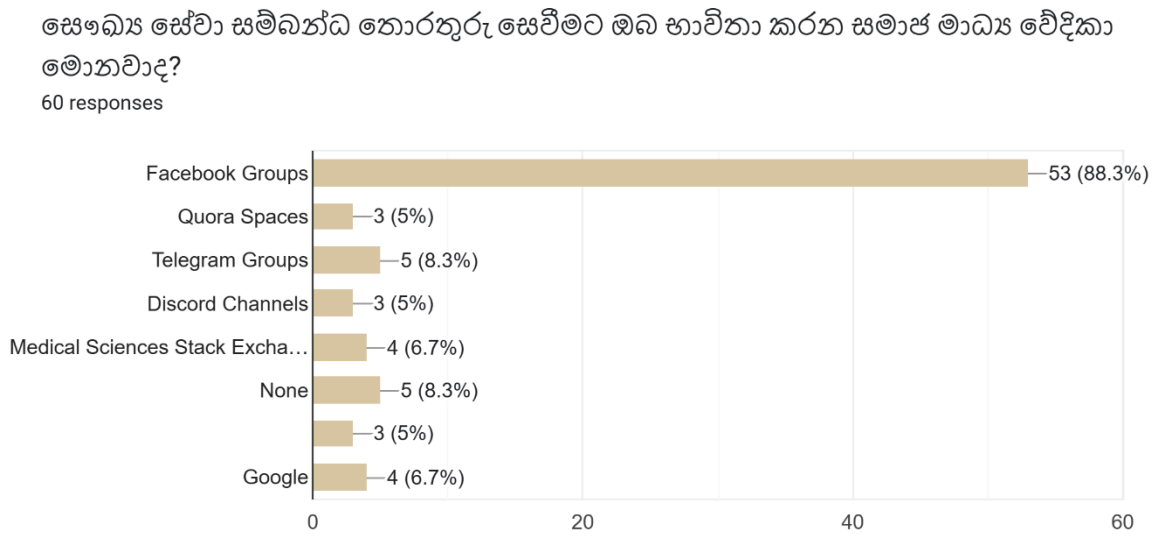

Figure 2: User Research Response

Most responses do not use social media platforms to seek health-related knowledge. Among existing social media platforms, most of the users are interested in using Facebook community groups to seek health-related knowledge. But on the other hand, users who participated in the user research are likely to try out a social network that will be specific to share health-related content only.

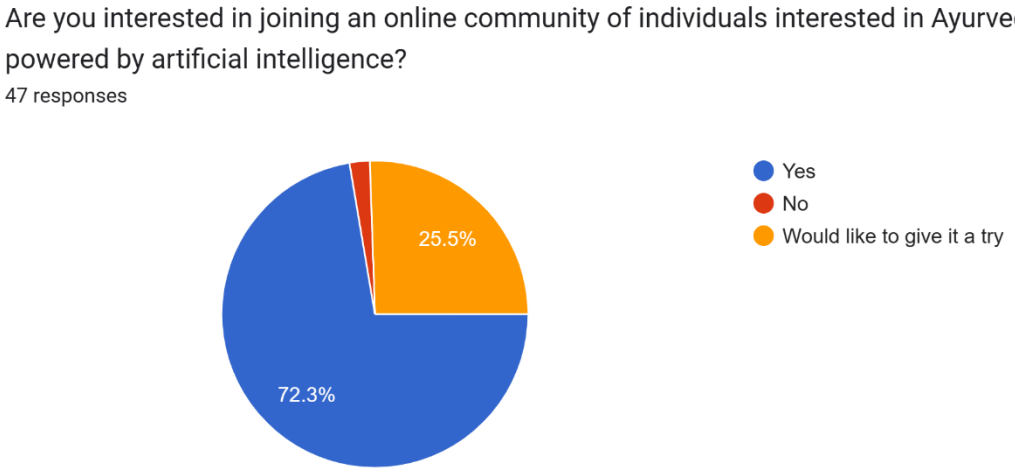

Figure 3: User Research Response

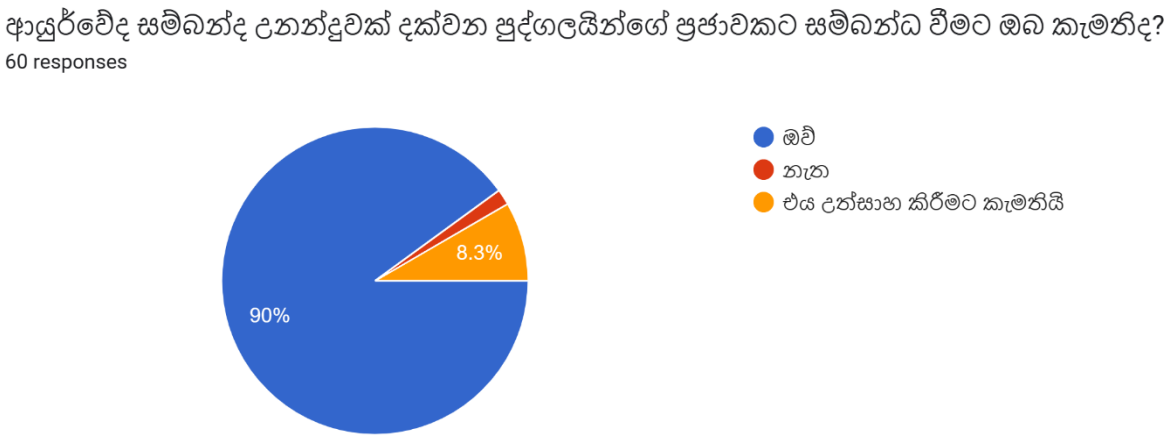

Figure 4: User Research Response

As problems that participants are facing, it collected the following responses.

can't verify the information are correct
Because the correct information cannot be found
None
none
Yes. Often provides incorrect information
Most of the content are not written by expertise
Lack of correct information and misinformation
It takes some time to get an answer
Too many guesses
Information is rare
No
The information isn't always accurate
It was difficult to finding facts
No idea.
Treatments for gastritis
There are so many annoying posts on that fb grps.
Too many suggestions but not specifics
fake details
I ALWAYS CHOOSE Medical practitioner
RESPONSES ARE VERY LATE

Figure 5: User Research Response

නිවැරැදි තොරතුරු ලැබෙන්නෙ නැති ගැටලුව

not accurate

නිවරදි තොරතුරු නොමැතිකම

Fake links

හැකිය

පිළිතුර වල නිවැරදි බව

Deta slowly

නියමිත ප්‍රමිතිය පිළිබඳ ,

No problems

Fake Details

අවශ්‍ය දේ නොතිබීම

Niwaradi thorathiru labenne natha

Correct details

තොරතුරු සොයා ගැනීමට අපහසු වීම

සමහරවිට අත්‍යවශ්‍ය දේ නොතිබීම

Privacy issues, trust issues

තොරතුරු වල නිරවද්‍යතාව.

ගැටළුවක් නැ

විශ්වාසනීයත්වය පිලිබඳ ගැටලු.

Figure 6: User Research Response

Regarding the responses, most were concerned with trustworthiness, quality, duplications, and search issues of content. Considering the responses, it is expected to address two concerns in the proposed social network implementation: handling duplications and enhanced searching. A Natural Language Processing based solution based on the BERT framework will be introduced to the proposed social network implementation to address the two concerns mentioned earlier.

There will be a knowledge base to facilitate chatbot service regarding the Ayurvedic health domain as a part of this project. Hence, it is expected to propose a solution to extract information from publicly shared content on the proposed social media platform and update the aforementioned existing knowledge base. This solution is expected to improve the results returned by the chatbot service for user queries.

The proposed social network implementation will be limited to sharing only health-related content.

Research Gap

Due to the prevailing social and health situation worldwide, there has been a significant increase in searches for health-related topics on social networks and popular search engines like Google.

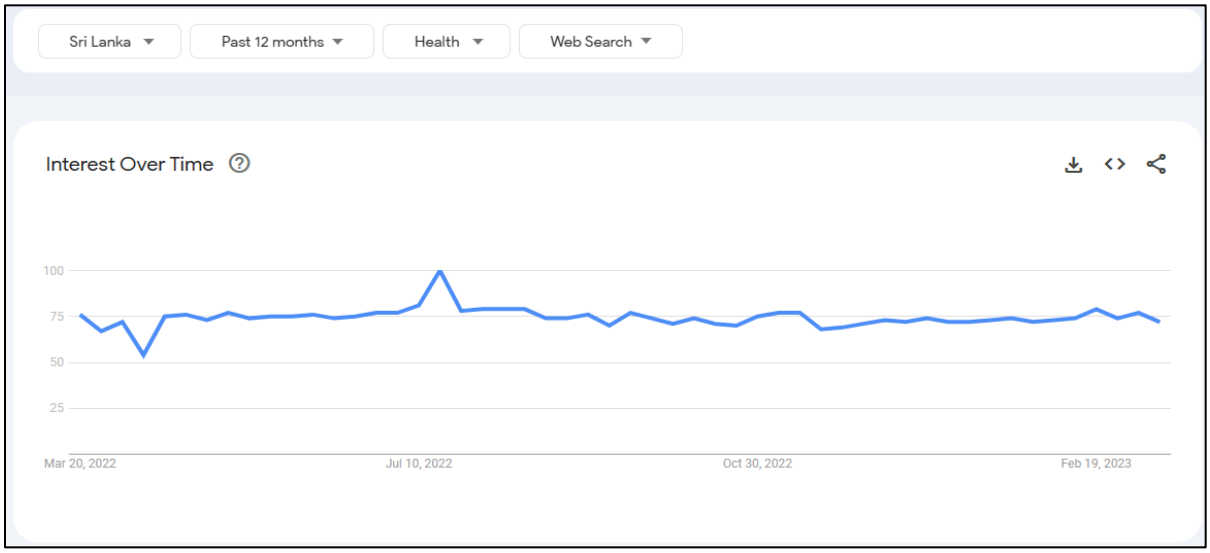

Figure 7: Google Search Trends in Health Domain

Over the past few years, significant progress has been made in developing information and communication technologies capable of providing content-sharing services through social networks. As a result, people progressively rely on social networks to obtain health-related information and merchandise, explore details regarding health and well-being, and receive guidance and exchange personal encounters. Existing social media platforms have some options for such a use case. Such choices and increasing social network usage opened an opportunity for many analytical possibilities.

Users in existing social media platforms such as Facebook and Quora face duplication of content frequently and immensely. Several situations have been reported that this has been negatively affected during elections in several regions on social media platforms such as Facebook and Reddit, which means many users have posted the same misleading information and persuaded people to believe it. On the other hand, this issue was kind of an unaccounted credibility of information. Furthermore, in a scenario that posts knowledge relating to a specific domain, achieving all the information in a single flow may be inconvenient if users post duplicating content. For example, in Stack Overflow, a community questions, and answers-based social network, there are many duplicated questions; therefore, they have many good answers in a couple of places, and users may have to spend some time searching for all and finding them. This may lead to poor user experience. There is existing research to detect duplicating content on social media platforms such as Quora, Twitter, and Stack Overflow. Most are based on machine learning techniques such as Support Vector Machine, Logistic Regression, Random Forest, and eXtreme Gradient Boosting. Research [2] has proposed three deep learning techniques, which are Word2Vec based: Conversational Neural Networks, Recurrent Neural Networks, and Long-Short-Term-Memory. The Word2Vec-based Long-Short-Term-Memory deep learning approach gained more accurate results than others. There are several deep learning approaches that aim specifically for Natural Language Processing tasks, such as transformers and generative pre-trained models. It is expected to follow a transformers-based deep learning approach using BERT (Bidirectional Encoder Representations from Transformers) open-source framework. Several reasons encouraged the selection of this approach, such as,

1. Word2Vec models produce context-independent embeddings, meaning that each word is represented by a single numeric vector, even if it has multiple senses. In contrast, BERT models generate context-dependent embeddings that enable us to have various numeric representations for the same word, depending on its context. As a result, BERT embeddings capture contextual nuances that context-independent embeddings cannot.

2. Word2Vec embeddings disregard word position information, while the BERT model considers the position of each word in a sentence as a direct input before computing its embedding.
3. Word2Vec pre-trained word embeddings are readily available for direct use, as they exist as a key-value mapping between words and vectors. The model is unnecessary; the input is a single word, while the output is the corresponding vector representation. In contrast, BERT generates contextual embeddings that require a sentence as input instead of a single word. The BERT model requires knowledge of the surrounding terms to generate a word vector, necessitating the presence of the trained model to produce embeddings based on input and context. The output provides a fixed-length vector representation of the entire input sentence.

Hence, this approach is expected to produce better accurate results than the Word2Vec-based Long-Short-Term-Memory deep learning approach.

Social media platforms can contain health content, including a vast array of information, patient experiences, health-related news, research findings, and expert opinions. This information can be used to update a knowledge base that stores various information and concepts relating to a domain. There is no existing research, or it may have yet to be publicly announced as a possibility on existing social media platforms. There is a knowledge base implementation as part of this research project to facilitate a chatbot service. It stores various information and medical concepts relating to Ayurvedic health treatments. Hence it is expected to propose a solution to update that knowledge base using content sharing in the social network implementation. A few things that need to be considered when proposing a solution for this functionality have been identified when proposing a solution for this functionality have been identified.

1. Existing research can extract medical concepts from a given text [3] [4]. But need to validate that they are showing the expected results in the Ayurvedic health domain. If not, it is possible to follow the same procedure that was followed in approaches mentioned in existing research to develop a matching model for the Ayurvedic health domain.
2. It requires identifying duplications.
3. It requires identifying contradictions.
4. It requires identifying new knowledge.
5. It requires labeling the updated knowledge as community knowledge with reference since collecting health-related information from the social network can only guarantee to be sometimes correct.

Compared to existing social network platforms, the proposed social network implementation can be distinguished as follows.

	Facebook Groups	**Reddit**	**Quora Spaces**	**Medical sciences stack exchange**	**Proposed implementation**
Sharing content limited to a specific domain	☒ (Manual process)	☒ (User selection)	☒ (Manual process)	☒ (User selection)	☒ (Only health-related content)
Detection of duplicates	☐	☐	☐	☒	☒
Labeling content (decided by user)	☒	☒	☒	☒	☒
Grouping based on the content	☒	☒	☒	☐	☒
Have both sharable questions and articles content types	☒	☒	☒	☒	☒
Collecting knowledge from sharable content	☐	☐	☐	☐	☒

Table 1: Comparison with Existing Solutions

Objectives

Main Objective

Implementing a social network that will be limited to sharing health-related content only.

Sub Objectives

- Detection of duplicates using the BERT framework, a transformers-based neural network architecture, and evaluating its performance metrics against existing studies.
- Propose a solution to update an existing knowledge base using the publicly shared content in the social network.

Methodology

Project Requirements

Functional and Non-Functional Requirements

Functional Requirements	**Non-Functional Requirements**
Manage shared content (posting, updating, and deleting) on the social network.	Performance – Enhanced search functionality should respond to each user's query, and knowledge base update functionality should be able to handle efficiently under minimum system requirements; otherwise, cloud services pricing may be increased.
An enhanced search functionality service to search for similar and related content.	Reliability – The system should be operational with minimum downtime.
Duplicating content detection.	Scalability – The system should be able to fluctuate users' request demands.
Scheduler service to update the knowledge base with recently posted content, including questions, articles, and comments on the social network.	Cost-effectiveness – The system should be optimized to use resources efficiently to minimize the cost of cloud service charges.

Table 2: Functional and Non-Functional Requirements

Software Requirements

- Node.js
- Python
- React Native
- Apache Kafka
- MySQL
- Apache Pinot

Personal Requirements

❖ An external resource person substantiated to collaborate with the project for extended knowledge requirements in health-related matters. (LinkedIn)

Data Requirements

❖ Required data will be gathered by publicly available content on existing social media platforms' health communities.

System Architecture

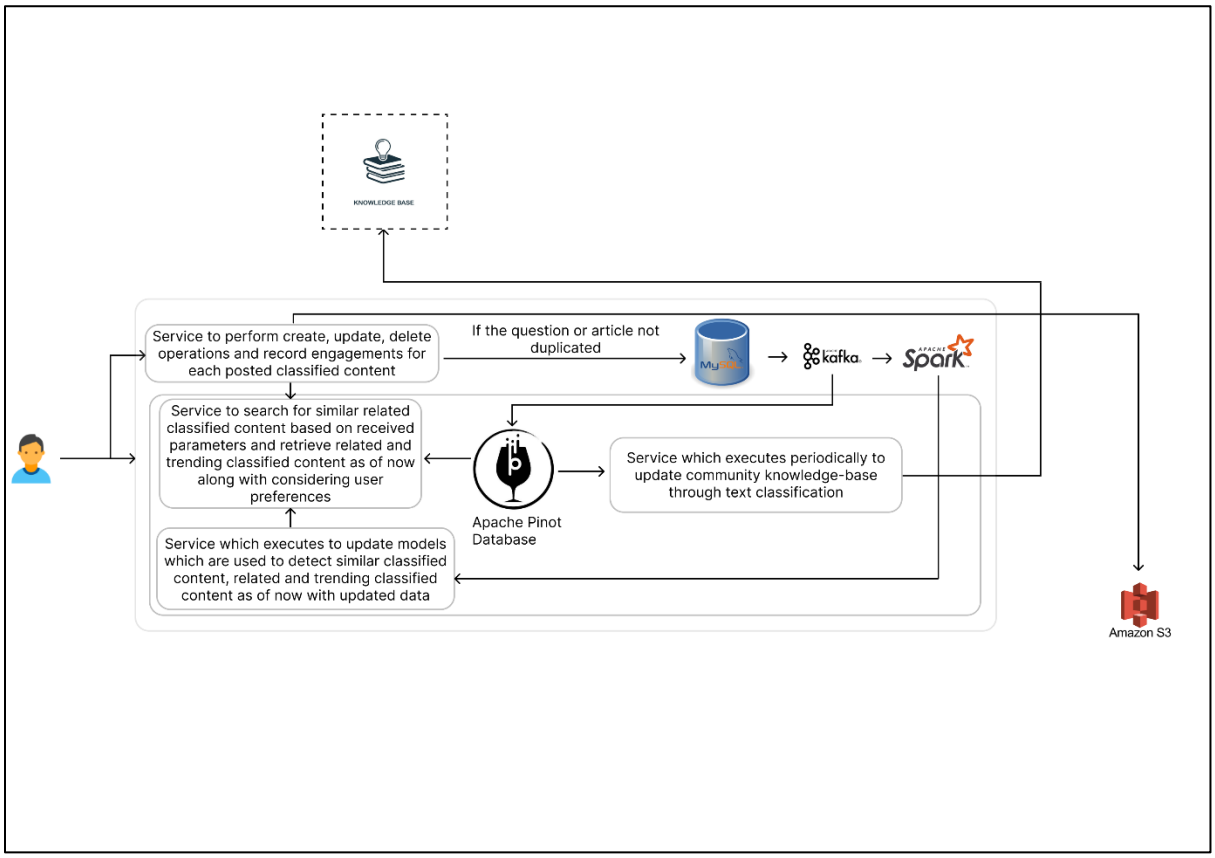

Figure 8: High-Level System Architecture Diagram

It is expected to develop a social network limited to sharing health-related content.

The overall deployment architecture will be based on Microservices architecture. The services have been broken down as follows.

1. Service which handles CREATE, UPDATE, and DELETE operations against the users' content.
2. Service which contains the enhanced search functionality.
3. Service which updates enhanced search functionality model based on existing data in the social network.
4. Service which periodically executes to update the existing knowledge base based on existing periodically added data in the social network.

MySQL database will be used as a persistent relational schema storage while Apache Pinot database will be used for low-latency query purposes. Apache Pinot database was selected for query purposes due to the following advantages.

1. Real-time analytics: Apache Pinot is optimized for real-time analytics, which means it can handle high volumes of incoming data and provide near-instant query results.
2. Scalability: Pinot is highly scalable and can handle large amounts of data. It can also be easily distributed across multiple nodes, making it suitable for use in a distributed computing environment.
3. Low latency: Pinot is optimized for low query latency, so it can quickly process and return query results.
4. Flexibility: Pinot can handle various data types and formats, making it suitable for use with a wide range of data sources. It also supports various query types, including range, filter, and aggregation queries.

It is expected to use Apache Kafka as an event stream processing technology and Apache Spark Engine as a data engineering technology for Natural Language Processing-based functionalities.

It is expected to use AWS S3 service to store images and videos uploaded to the social network. It may also be used to keep older models built for enhanced search functionality and knowledge base update functionality.

Furthermore, it is expected to explore the possibility of incorporating an AutoML pipeline for enhanced search functionality service to keep learning models up to date with recently updated content in the social network instead of a separate standalone service written from scratch, as mentioned earlier.

Implementation Design

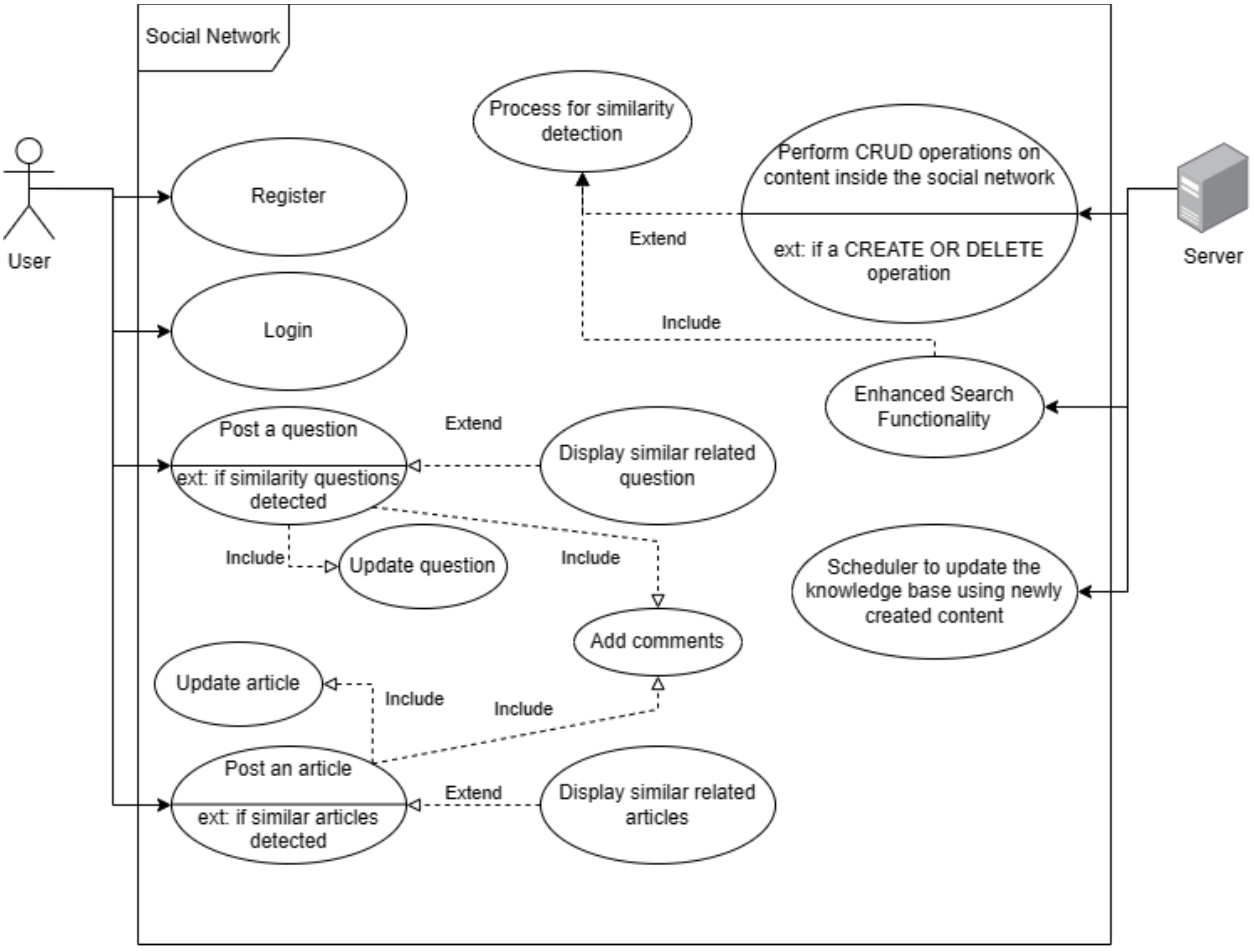

Figure 9: Use Case Diagram

The above use case diagram provides an overview of the functional and technical aspects of the server system, highlighting both the external and internal actors involved separately.

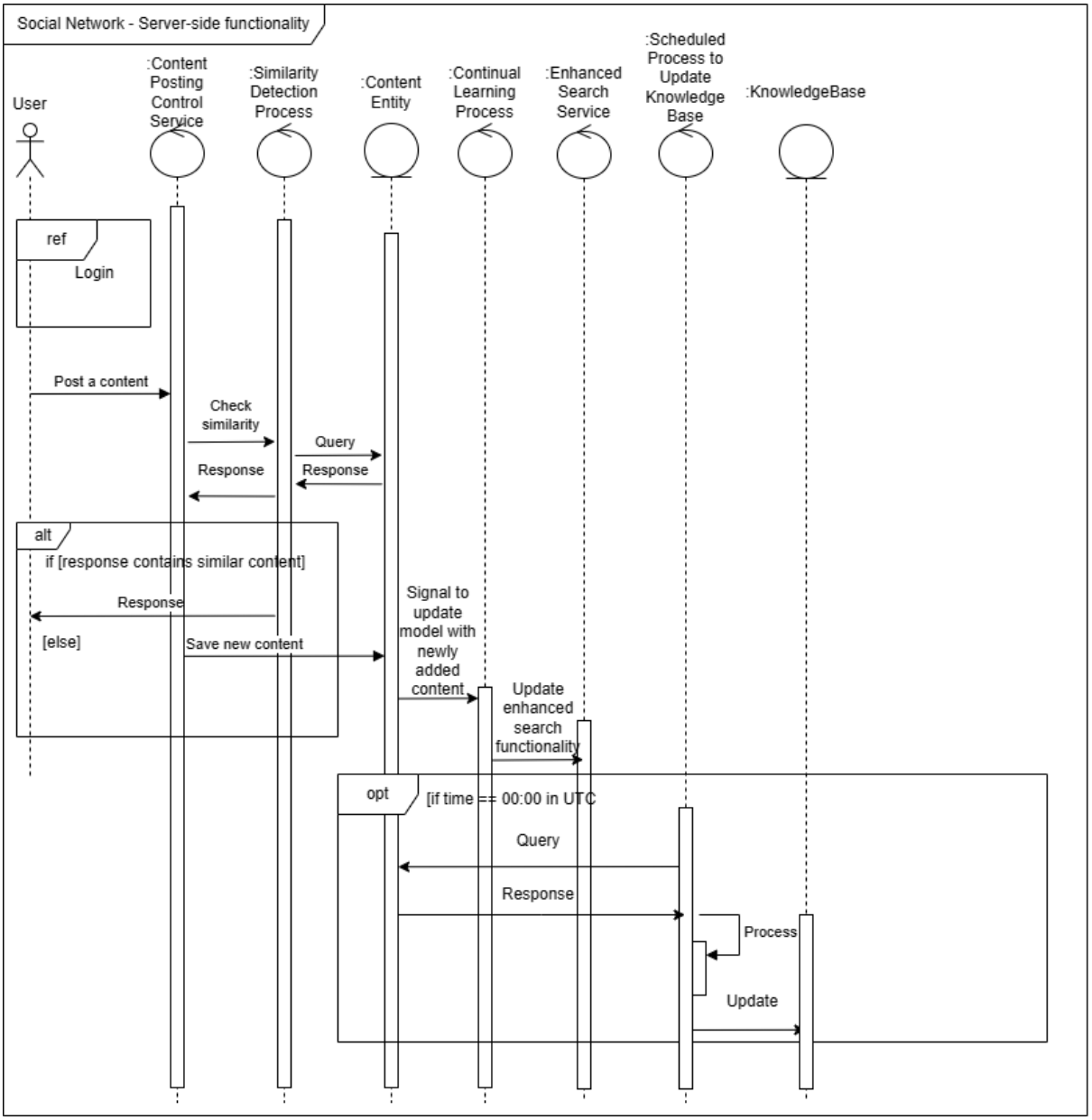

Figure 10: Sequence Diagram

The system's technical interactions are illustrated in the above sequence diagram.

A standalone identity service will be used to handle user accounts and user sessions.

Implementation of the service, which handles CREATE, UPDATE, and DELETE operations against the users' content and enhanced search functionality, will be based on Clean Architecture and CQRS Pattern. Internal process communication between the mentioned services will be facilitated by Apache Kafka when content is added, updated, or deleted in the social network.

Other earlier-mentioned services' internal implementation designs will be evaluated during the research.

Feasibility of the Solution

The following scopes have been considered regarding the feasibility of the solution.

- Technical Feasibility –
    - It will require knowledge of the following programming and scripting languages.
        - Python and JavaScript
    - It will require knowledge of the following architectures at the deployment level.
        - Kubernetes Architecture
- Economic Feasibility – Several cloud providers have been identified that required services made available more affordable and reliable.
- Schedule Feasibility – All the functionalities should be implemented within a given time frame. And it is expected can be done.
- Operation Feasibility – All the deployment operations are expected to automate using continuous integration and deployment tools. Furthermore, several incident management, performance monitoring, and error tracking services have been evaluated and will be used depending on their features and pricing models for a better development experience and any defect tracking.
- Legal Feasibility – Inside Sri Lanka, it is required to get prior approval from the technical commissioner in Ayurvedic Department if an electronic-based application is developing relating to Ayurveda. And approval has already been requested.

Business Evaluation

- Target Audience – Any interested user can join the social network without any aged frame since it is limited to sharing only health-related content. But people sensitive to extreme content may refrain from using this app.
- Prior Requirements Expected from Target Audience:
    - General literacy to engage with a mobile application.
    - Users are expected to behave responsibly.
    - Users are expected to respect other parties' privacy.
    - When posting content, it is expected to have qualifications in the relevant scopes.

- Identified strengths of the solution:
    - There are no existing social networks that can share only health-related content. This means this solution is a specialized approach that can target only interested individuals. (There is a community question and answers-based social network which is Stack Exchange Medical Sciences which is currently in Beta release)
    - Elimination of posting duplicating content.
    - Improved queries using a specialized database (use of an Online Analytical Processing database such as Apache Pinot).
- Known weaknesses of the solution:
    - There is no mechanism to refrain from posting content that intends to market products. Several scenarios have been identified in existing social media platforms; some people post Ayurvedic-related content to market their products.
    - There is no mechanism that collects users' qualifications and lets them post articles only if they have relevant qualifications. Hence any person can even post something, even if it may be in assumptions.
- Possible threats can occur from the solution:
    - There are several cases that have been reported in even qualified, well-reputed personalities spam medical-related information for their benefit. For example, during the Covid-19 pandemic, it can be seen on many existing social media platforms regarding vaccination programs. Such activities can threaten this solution since it may decrease the confidentiality of the users and may give up using the social network.
- Opportunities can open up from the solution:
    - Since many people can share their experiences relating to health-related matters inside the social network, collecting and organizing such information can be beneficial.
- As future improvements, the weaknesses mentioned above can be addressed.

Forecasted Budget

This can be changed by using other services during the project's development.

- Azure Container Apps – Free Tier
- AWS VM – Free Tier (12 Months)
- GitHub Actions – Free Tier
- Docker Hub – 5 USD/month or Free Tier
- Firebase Messaging – Free
- Google Play Registration Fee – 25 USD

Work Breakdown Structure

- Initiation, Planning, and Design
    - Discussions with supervisors
    - Decide technologies and additional functionalities.
    - Create project charter and proposal documents.
- Implementation
    - Develop a social network.
    - Develop enhanced search functionality service.
    - Develop scheduled service to update the knowledge base using newly created content in the social network.
    - Deployment of the solution.
    - Testing.
- Conclusion
    - Prepare the research paper.
    - Prepare the final report and other required submissions.

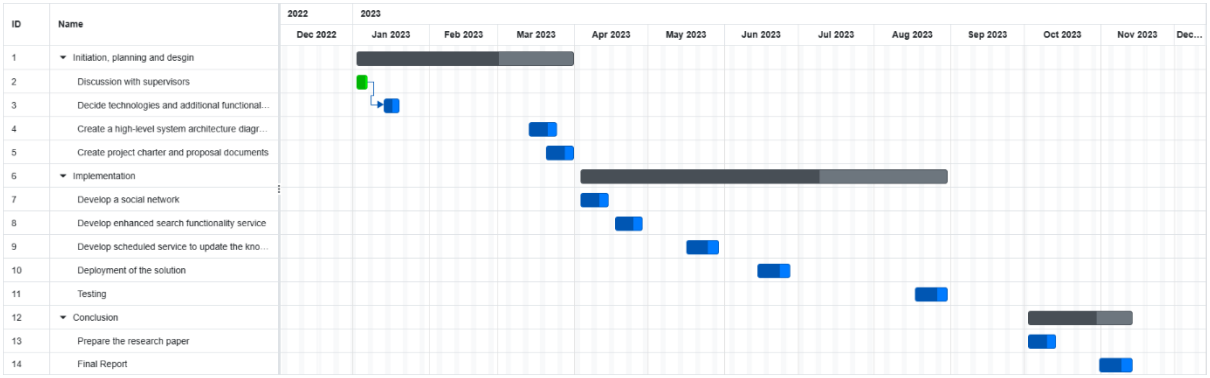

Figure 11: Gantt Chart

References

[1] A. Abbasi, D. Adjeroh, M. Dredze, M. J. Paul, F. M. Zahedi, H. Zhao, N. Walia, H. Jain, P. Sanvanson, R. Shaker, M. D. Huesch, R. Beal, W. Zheng, M. Abate and A. Ross, "Social Media Analytics for Smart Health," *IEEE Intelligent Systems,* vol. 29, no. 2, pp. 60-80, 2014.

[2] L. Wang, L. Zhang, and J. Jiang, "Duplicate Question Detection With Deep Learning in Stack Overflow," *IEEE Access,* vol. 8, pp. 25964 - 25975, 2020.

[3] K. Denecke, "Extracting Medical Concepts from Medical Social Media with Clinical NLP Tools: A Qualitative Study," 2014.

[4] S. Luca and N. Goharian, "QuickUMLS: a fast, unsupervised approach for medical concept extraction," 2014.