THE FUTURE OF AYURVEDA:
HARNESSING THE POWER OF ARTIFICIAL INTELLIGENCE FOR PERSONALIZED TREATMENT AND DIAGNOSIS

Jayasinghe Arachchige Sunera Chamoda Jayasinghe

(IT20216078)

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

Department of Computer Science and Software Engineering

Sri Lanka Institute of Information Technology
Sri Lanka

September 2023

DECLARATION

I declare that this is my own work, and this dissertation 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 my 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 non-exclusive right to reproduce and distribute my dissertation in whole or part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as article or books).

Name	Student Identity Number	Signature
Jayasinghe J A S C	IT20216078	Sunera

Signature of the Supervisor
(Dr. Darshana Kasthurirathna)

Date

……... ……...

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 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.

ACKNOWLEDGEMENT

First and foremost, I sincerely thank my mentor Dr. Darshana Kasthurirathna, for his unwavering support and direction, which enabled me to successfully complete my undergraduate research. My supervisor, Dr. Samantha Rajapaksha, the co-supervisor of this research project deserves my deepest gratitude for always being available to assist. Due to the fact that this research project combines technology and Ayurveda, both technology specialists and Ayurvedic professionals were needed for advice and support. It is very appreciated that Dr. Mrs. Wasantha Janaki Wickramaarchchi who provided such extensive help throughout the project to close the knowledge gap in those fields. My sincere gratitude also goes out to the other doctors of Gampaha Wikramarachchi Ayurvedic University. Last but not least, I would like to convey my thanks to everyone who has helped with this project in some way, whether directly or indirectly, including my teammates, family, and friends.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

Background Literature

Information quality

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 factor. 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, we can better assess the credibility of health information on social media and make informed decisions about them.

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 we'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, it is important to 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, we can ensure that the health information consumed 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, it is important to 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, we 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, it is important to 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

them. By considering these factors, we 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, it is important to 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, we can better assess the relevance and credibility of information in social media health content that is salient to our information 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 to understand who, what, when, where, why, and how referencing salience. On the other hand, salience also leads to privacy concerns.

Information extraction

Information extraction aims to transform unstructured data into structured, machine-readable data. Unstructured data refers to data that doesn't have a predefined data model or structure, such as text documents, social media posts, and multimedia content. Such data may be in text, image, video, or audio formats. This structured data can be used for various applications, including knowledge graph construction, question-answering systems, and semantic search engines. Information extraction techniques have evolved significantly thanks to advancements in various machine learning and deep learning techniques.

Early information extraction approaches relied on rule-based methods and hand-crafted patterns to extract information. These systems were limited in scalability and adaptability, as they required manual intervention to create rules for each new information pattern.

The shift towards machine learning and deep learning-based approaches brought scalability and adaptability to information extraction. Natural language processing approaches, such as named entity recognition and relation extraction, became popular in text format data. These approaches utilized labeled data for training classifiers, which could then be generalized to extract information from new texts. Named Entity Recognition is a fundamental component of information extraction, involving the identification and classification of entities like names of people, organizations, locations, and more. Conditional Random Fields (CRFs) and Recurrent Neural Networks (RNNs) were widely used for named entity recognition tasks. Additionally, large-scale pre-trained language models like Bidirectional Encoder Representations from Transformers (BERT), introduced in 2018, have significantly improved named entity recognition task performance by leveraging contextual information [2]. Relation extraction involves identifying and classifying relationships between entities in text. Early methods used hand-crafted features and rule-based systems, but deep learning approaches, such as Convolutional Neural Networks (CNNs) and Transformer models, have shown remarkable performance improvements in recent years. These models can capture complex semantic relationships between entities. Distant supervision is a technique used to generate training data for relation extraction automatically. It aligns

existing knowledge bases with text to identify entity pairs and their relationships. However, noisy labeling and data imbalance issues are challenges in this approach.

Part-of-speech tagging is a fundamental task in natural language processing that involves assigning each word in a sentence a grammatical category, such as noun, verb, adjective, etc. Part-of-speech tagging is crucial for various downstream natural language processing tasks, including information extraction. Early part-of-speech tagging systems relied on hand-crafted rules and linguistic knowledge. However, machine learning approaches, particularly Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs) have been widely adopted. More recently, neural network-based models, such as recurrent neural networks (RNNs) and transformers, have achieved state-of-the-art results due to their ability to capture contextual information effectively. [3]

Dependency parsing is the process of analyzing the grammatical structure of a sentence by identifying the syntactic relationships (dependencies) between words. These relationships are typically represented as a tree structure known as a dependency tree. Dependency parsing is crucial for understanding the grammatical structure of a sentence, which is essential for many natural language processing tasks, including information extraction. Early dependency parsers used hand-crafted rules and linguistic principles. Transition-based and graph-based parsing algorithms, such as the Shift-Reduce and Arc-Standard algorithms, gained popularity. Dependency parsing also benefited from neural network-based models, including graph convolutional networks (GCNs) and transformer-based models like Bidirectional Encoder Representations from Transformers (BERT), which improved parsing accuracy and efficiency.

Sentiment analysis, also known as opinion mining, involves determining the sentiment or emotion expressed in a piece of text, such as positive, negative, or neutral. This task has significant applications in understanding public opinion, customer feedback analysis, and social media monitoring. Early sentiment analysis methods relied on rule-based approaches and sentiment lexicons. However, machine learning techniques, including supervised learning and deep learning, have become dominant in sentiment

analysis. Sentiment analysis models are typically trained on labeled datasets and can range from simple models like logistic regression to complex neural architectures like recurrent neural networks (RNNs) and convolutional neural networks (CNNs). Sentiment analysis research has also extended to aspect-based sentiment analysis (ABSA), where the goal is to identify sentiment towards specific aspects or entities within a text, providing more fine-grained sentiment insights. [4]

Coreference resolution is the task of determining when two or more expressions in a text refer to the same entity. For example, in the sentence "John said he would come," coreference resolution identifies that "he" refers to "John." This task is crucial in information extraction because it helps establish relationships between different mentions of the same entity, which is essential for constructing accurate knowledge graphs and understanding the context of a text. Early approaches to coreference resolution relied on hand-crafted rules and linguistic knowledge, but these methods had limitations in handling complex and ambiguous cases. Machine learning-based approaches, particularly mention-pair models and mention-ranking models have gained prominence. The advent of neural network-based models, such as recurrent neural networks (RNNs) and transformer-based models, has significantly improved coreference resolution accuracy. These models can learn complex patterns in text and capture long-range dependencies. One of the notable architectures for coreference resolution is the mention-ranking neural network (RNNG), which combines both mention-pair and mention-ranking strategies. Coreference resolution remains a challenging problem, especially in cases involving indefinite or ambiguous references, complex syntactic structures, and cross-document coreference. Addressing gender and bias-related challenges is also a concern in coreference resolution, as models must make accurate and fair decisions. [5]

Duplication information detection, also known as text similarity or duplicate content detection, is the task of identifying and measuring the similarity between pieces of text. It is a crucial component in various applications, including plagiarism detection, content deduplication, and document clustering. Early duplication information detection methods often relied on simple techniques such as exact string matching or text hashing. These methods were effective in identifying direct, verbatim copies of

text but struggled with more subtle forms of duplication, including paraphrasing and document structure variations. Vector space models, including Term Frequency-Inverse Document Frequency (TF-IDF) and Latent Semantic Analysis (LSA), represented documents as vectors in high-dimensional spaces. These models allowed for measuring document similarity by calculating cosine similarity between document vectors. While effective for some tasks, they still had limitations in capturing semantic similarity. The advent of word embeddings, such as Word2Vec and GloVe, enabled a more nuanced understanding of word semantics. These embeddings could be used to calculate sentence or document embeddings by averaging or pooling word embeddings. This allowed for better measuring semantic similarity between texts. Siamese networks and other neural architectures, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have become increasingly popular for duplication information detection. Siamese networks, in particular, are designed to learn the similarity between two input texts and are trained on pairs of similar and dissimilar text samples. They have the advantage of capturing complex relationships between words and phrases in texts. Pretrained language models, like BERT (Bidirectional Encoder Representations from Transformers), have revolutionized duplication information detection. These models can encode entire sentences or documents and are fine-tuned for various tasks, including text similarity. BERT-based models have achieved state-of-the-art results in many natural language processing benchmarks due to their contextual understanding of language. [6]

Near duplication detection, part-of-speech tagging, coreference resolution, dependency parsing, and sentiment analysis can be used as integral components in the information extraction pipeline. They help in identifying entities and relationships, understanding the grammatical context, and extracting sentiment-related information from text. For instance, sentiment analysis can be used to assess public sentiment toward specific entities or products mentioned in news articles. In this project, it is expected to experiment with an integral solution of these concepts to extract information from health-related social networks.

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.

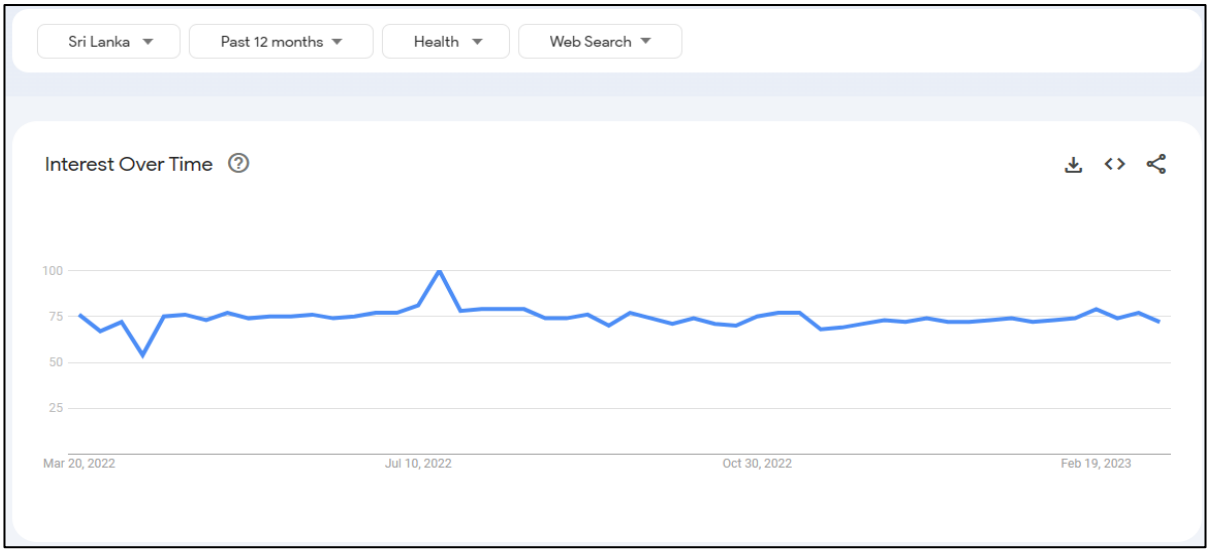

Figure 1: Google Search Trends

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.

These networks host vast amounts of textual data, which can be invaluable for healthcare professionals, researchers, and users seeking information and support. To harness the full potential of this data, there is a growing need for robust and integrated natural language processing solutions that can extract meaningful insights from unstructured text. It is expected to address the research gap in developing an integrated solution for various natural language processing tasks, including near duplication detection, part-of-speech tagging, dependency parsing, coreference resolution, relation extraction, and sentiment analysis within the context of a health-based social network.

Currently, there is a lack of comprehensive solutions that seamlessly integrate multiple natural language processing tasks into a single pipeline for health-based social networks. While individual natural language processing tasks are well-studied, their integration into a coherent workflow tailored to the specific needs of this domain remains underexplored.

Health-related texts are often complex, diverse, and domain-specific. Existing natural language processing tools may not perform optimally in this context, especially relating to Ayurvedic medical practices. Developing and fine-tuning natural language processing models for health-based social network data, considering the nuances of medical terminology, abbreviations, and user-generated content, is a significant challenge. Also, it may require generalized approaches while adhering to domain-specialized approaches since, in a social media context, it may used by users who are not specialized in such areas. Hence, it is important to focus on a combination of such approaches.

Ensuring the quality and reliability of information is crucial in healthcare. The identification and management of duplicate or near-duplicate content in user-generated posts are essential but often neglected tasks in information extraction pipelines for health-based social networks. Existing social network platforms such as Stackoverflow are experimenting with such features in the present, with its latest Medical Sciences Stach Exchange platform currently in beta release.

Extracting meaningful relationships between medical concepts, users, and their discussions is vital for knowledge discovery and recommendation systems. However, relation extraction models tailored to the unique needs of a health-based social network are lacking. There are existing studies but more specialized for medical practices since those trained model datasets depend on clinical notes and existing research papers relating to health context. Understanding the sentiment and emotional tone of user-generated content can provide valuable insights into users' mental states, opinions, and experiences. Developing sentiment analysis models that can capture the nuances of health-related sentiment in social network discussions is an unaddressed issue in existing research. Also, deploying an integrated natural language processing pipeline

in a real-world health-based social network requires solutions that are scalable, efficient, and can handle the vast volume of incoming data without significant latency.

Extracting information from health-related social networks raises privacy and ethical concerns. Addressing these issues while designing the integrated solution is essential for user trust and compliance with data protection regulations.

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.

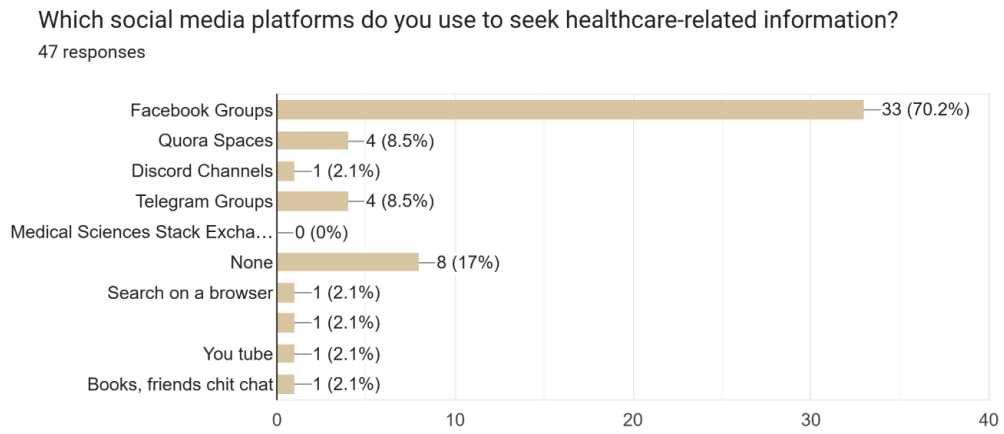

Figure 2: User Research Response

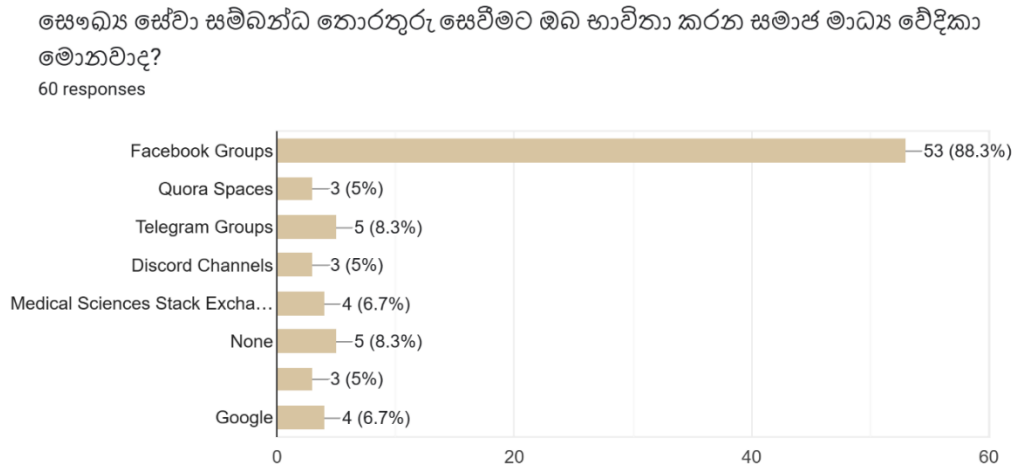

Figure 3: User Research Response

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.

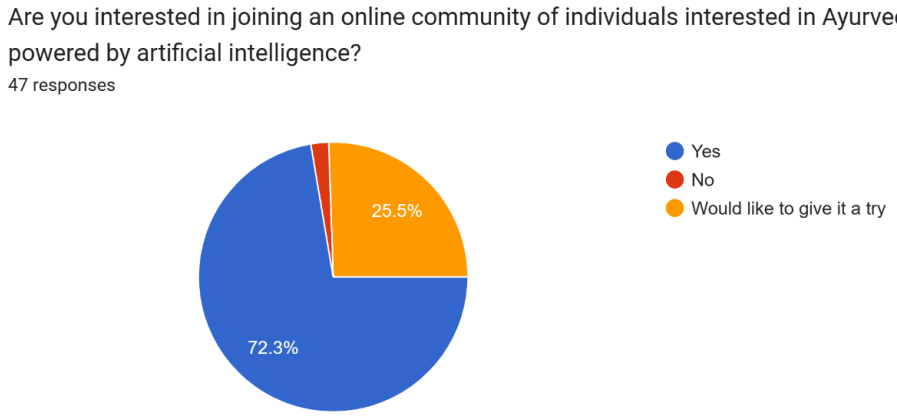

Figure 4: User Research Response

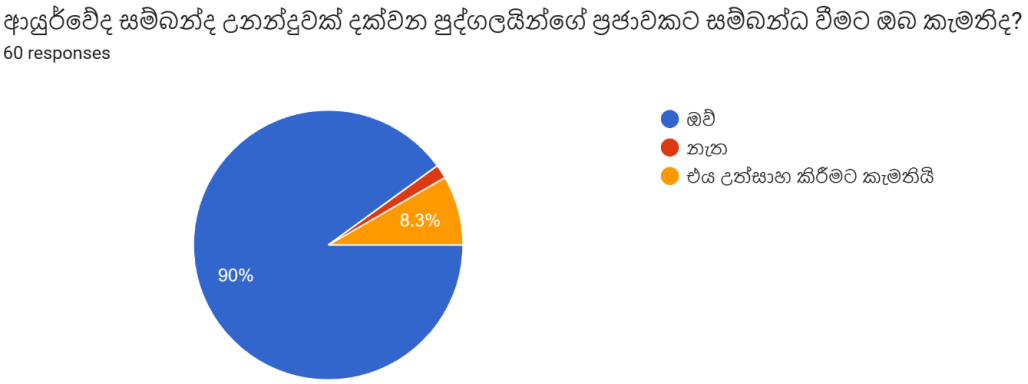

Figure 5: User Research Response

As problems that participants are facing when seeking health-related information using current approaches, it is 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 6: User Research Response

නිවැරැදි තොරතුරු ලැබෙන්නෙ නැති ගැටලුව

not accurate

නිවරදි තොරතුරු නොමැතිකම

Fake links

හැකිය

පිළිතුර වල නිවැරදි බව

Deta slowly

නියමින ප්‍රමිතිය පිළිබද ,

No problems

Fake Details

අවශ්‍ය දේ නොතිබීම

Niwaradi thorathiru labenne natha

Correct details

තොරතුරු සොයා ගැනීමට අපහසු වීම

සමහරවිට අනාවශ්‍ය දේ නොතිබීම

Privacy issues, trust issues

තොරතුරු වල නිරවද්‍යතාව.

ගැටළුවක් නෑ

විශ්වාසනීයත්වය පිලිබඳ ගැටලු.

Figure 7: 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, enhanced searching, and additionally, it is expected to implement an information extraction pipeline to extract trending health trends circulating inside the social network. The extracted information will be saved in a knowledge base and will be used as the data source for the chatbot service, which is a part of the overall project.

Research Objectives

Main objective

Implement a social network that will be limited to sharing health-related content. Users have the ability to both ask questions and post articles and can respond to them as well.

Sub objectives

- Detect near duplications of the content against existing public content in the social network to ensure that users do not repeatedly post highly similar or identical content, reducing redundancy and maintaining content quality.
- Implement an enhanced semantic search functionality to provide users with more accurate and comprehensive results for health-related topics while understanding the context and intent behind user queries. Moreover, search results will be improved by considering user preferences and historical interactions.
- Implement an information extraction pipeline to automatically extract and analyze health-related trends from the social network's content and save in a knowledge base.

METHODOLOGY

Overall System Architecture

High-level system architecture

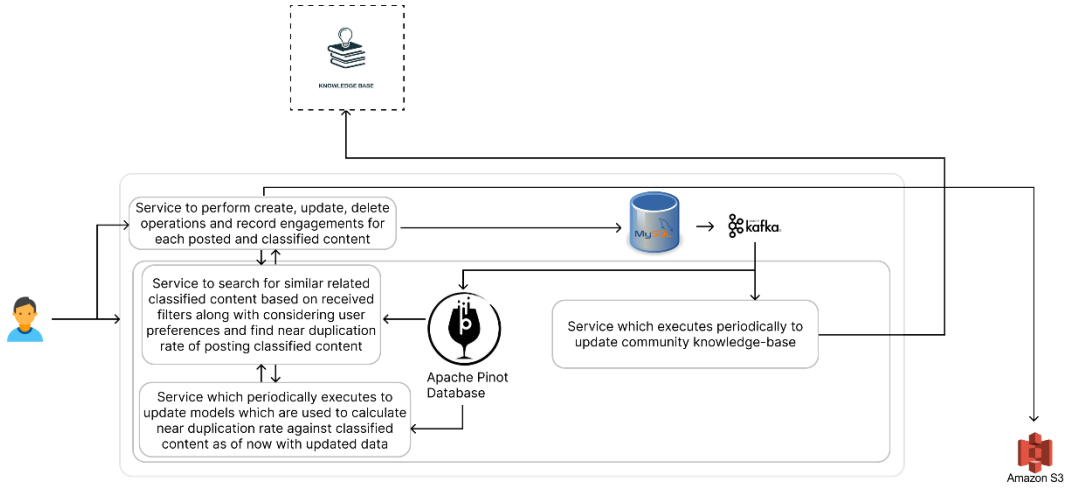

Figure 8: High-Level System Architecture Diagram

- Overall service implementation architecture is based on Microservices architecture. Each implementing service has been identified as follows.
    - Service which handles CREATE, UPDATE, and DELETE operations against the users' content.
    - Service which contains the enhanced semantic search functionality.
    - Service that updates enhanced search functionality model based on existing data in the social network.
    - Service that periodically executes to update the existing knowledge base based on periodically added data in the social network.
- MySQL database is used as a persistent relational schema storage while Apache Pinot database is used for low-latency query purposes. Apache Pinot database was selected for improved query performance purposes due to the following advantages.

    - 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.
    - 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.
    - Low latency: Pinot is optimized for low query latency, so it can quickly process and return query results.
    - 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 uses Apache Kafka as an event stream processing technology to update the knowledge base in real-time and make enhanced semantic search functionality prepared to search for newly added content.
    - Apache Kafka is an open-source distributed event streaming platform that is widely used for building real-time data pipelines and event-driven applications. It is designed to handle high-throughput, fault-tolerant, and scalable event streaming and can retain event data for a specified retention period (configurable), allowing consumers to replay events and perform real-time analytics or processing.
    - It consists of two messaging model architectures.
        - Publish-Subscribe Messaging Model: Producers publish messages to topics, and consumers subscribe to topics to receive those messages. Topics act as channels or categories where data is organized and stored.
        - Message Queue Messaging Model: Messages are typically consumed by one consumer within a consumer group. However, multiple consumer groups can exist, each with its own set of consumers, and each group receives a copy of the same data.
- It uses the 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 semantic search functionality and information extraction pipeline functionality.

Implementation designs

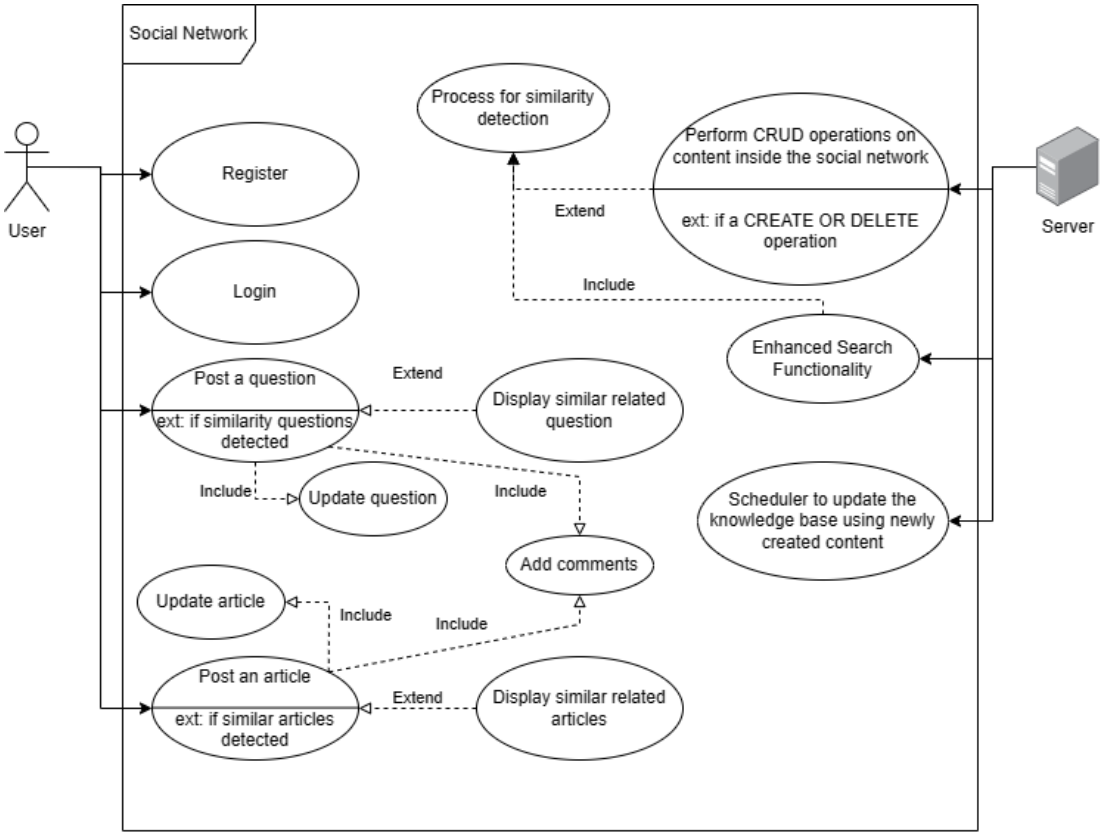

Figure 9: Use Case Diagram

- 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.

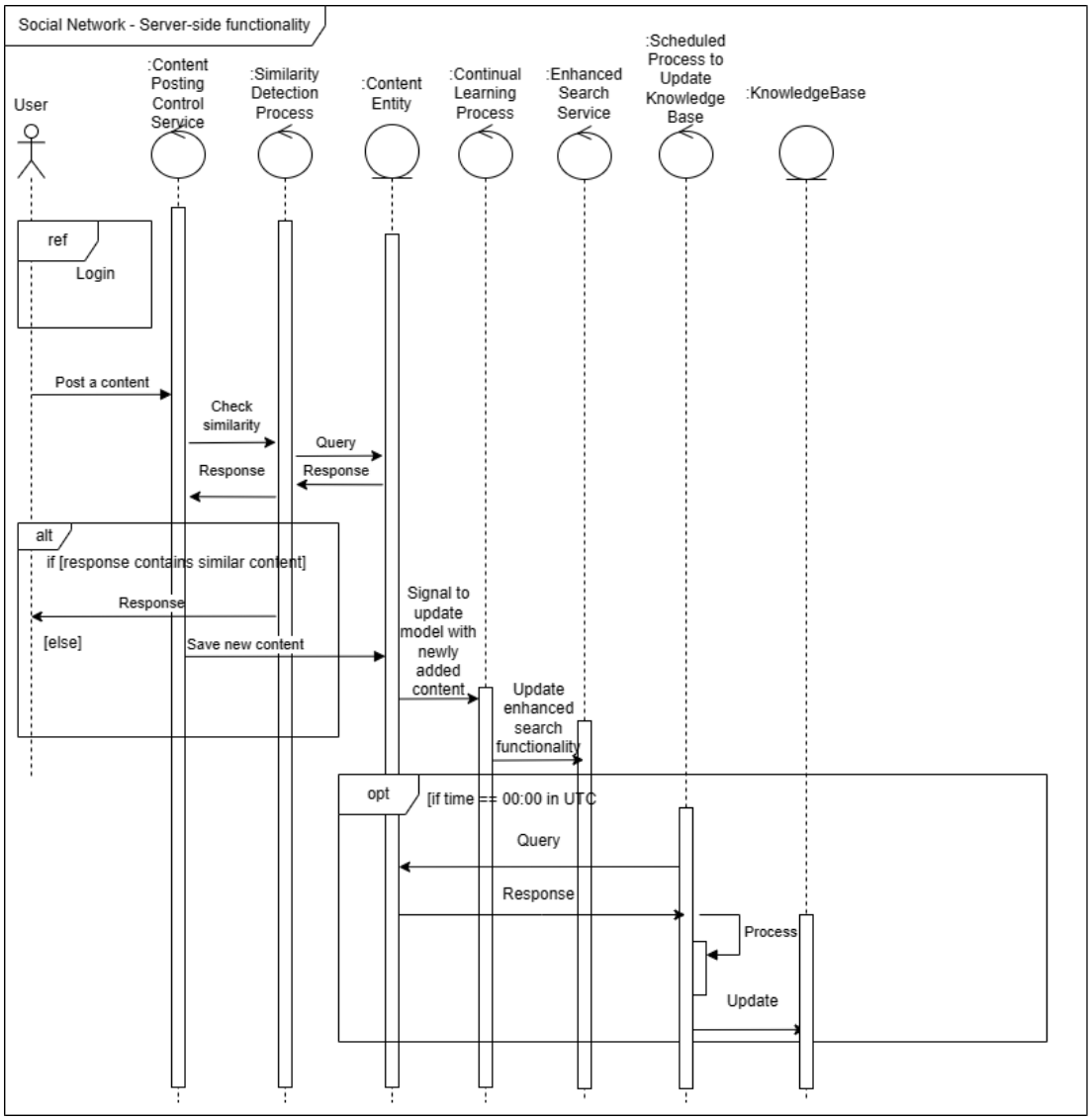

Figure 10: Sequence Diagram

- The system's technical interactions are illustrated in the above sequence diagram.
- A standalone identity service will be used to handle user accounts, user preferences, 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 the Clean Architecture and Command Query Responsibility Segregation (CQRS) Pattern. It enables the creation of

maintainable, scalable, and testable software by separating concerns and promoting the separation of command (write) and query (read) operations.

- Clean Architecture is a software design philosophy introduced by Robert C. Martin, also known as "Uncle Bob." It provides a structured way to organize the codebase, making it more modular and maintainable. The main idea behind Clean Architecture is to separate the code into concentric circles or layers, each with a specific responsibility. These layers typically include:
  - Entities: These are the core domain objects that represent your business logic and data. They should be independent of any external frameworks or databases.
  - Use Cases: Use cases, also known as interactors, encapsulate application-specific business rules and logic. They act as an intermediary between the outer layers (e.g., the UI or controllers) and the inner layers (e.g., entities and data access).
  - Interfaces (Adapters): Interfaces define the boundaries between different parts of your application. They include the input and output ports that allow you to communicate with the external world. Examples of interfaces include web API controllers, database gateways, and external service clients.
  - Frameworks and Drivers: This outermost layer includes the frameworks, libraries, and tools that connect your application to the outside world. It encompasses the UI, databases, external APIs, and other external dependencies.
- Clean Architecture promotes several key principles:
  - Dependency Rule: Dependencies should always point inward, from the outer layers toward the inner layers. This helps maintain a clear separation of concerns and makes it easier to replace or update components without affecting the core business logic.

        - Testability: With the core business logic separated from external dependencies, it becomes easier to write unit tests for the application's critical components.
        - Maintainability: Clean Architecture makes it easier to understand, extend, and modify the codebase over time, as it enforces clear boundaries and responsibilities.
    - CQRS is a pattern that complements Clean Architecture by separating the responsibilities of handling commands (writes) and queries (reads). In a CQRS-based system, it has two primary components:
        - Command Side: This side handles commands that modify the state of the application, such as creating, updating, or deleting data. Commands are typically executed asynchronously and can involve complex business logic.
        - Query Side: This side handles queries for retrieving data. It's optimized for fast and efficient reads. Query-side components may involve caching and denormalization to provide quick access to data.
    - Key concepts and benefits of CQRS:
        - Separation of Concerns: CQRS ensures a clear separation between operations that change data (commands) and operations that retrieve data (queries). This separation makes it easier to optimize and scale each side independently.
        - Scalability: Since read and write operations are separated, you can scale each side independently based on its specific workload and requirements. This is particularly useful in systems with varying read and write loads.
        - Optimized Queries: The query side can be optimized for efficient data retrieval, allowing you to use different storage mechanisms, caching strategies, and denormalization techniques to enhance query performance.
        - Event Sourcing: CQRS often goes hand-in-hand with event sourcing, where changes to the application state are recorded

as a series of immutable events. Event sourcing enables a full audit trail of all changes and supports complex data recovery and analysis.

    - When implementing CQRS in Clean Architecture for the social network backend APIs, it is designed use cases and interfaces to handle commands and queries separately, ensuring a clean separation of responsibilities.

- Internal process communication between the rest of the mentioned services will be facilitated by Apache Kafka when content is added, updated, or deleted in the social network.

Enhanced Semantic Search Functionality

The enhanced search functionality of Sentence-Transformers semantic search is utilized to make social network searches more convenient. There are two approaches to semantic search.

- Symmetric search:
    - In symmetric semantic search tasks, both the query and the entries within your dataset are typically of similar length and contain a comparable amount of content. To illustrate this, consider searching for similar questions. For instance, your query might be "How to learn Python online?" with the goal of finding an entry such as "How to learn Python on the web?" In these scenarios, it's often feasible to interchange the query and the entries within your dataset to achieve the desired results.
- Asymmetric search:
    - Asymmetric semantic search tasks, the query is usually short, such as a question or a set of keywords, and the objective is to locate a longer paragraph or document that provides an answer or detailed information related to the query. For example, a query like "What is Python?" would aim to retrieve a paragraph like "Python is an interpreted, high-level, and general-purpose programming language. Python's design philosophy..." In asymmetric tasks, reversing the

positions of the query and the entries in your dataset typically does not yield meaningful results.

- It used asymmetric search implementation to deliver better search results for the users.

The semantic search functionality implemented for the proposed social network consists of several steps.

- The initial step in the data collection process involves obtaining unprocessed text data from social media platforms, which may include various forms of user-generated content such as questions, articles, and responses. This data is then queried from the Apache Pinot database.
- To begin the process of encoding the user query, a pre-trained "all-MiniLM-L6-v2" sentence-transformer model is initialized. This model is known for its fast processing speed, high accuracy and performance.
- Using the loaded Sentence Transformer model, the received user query is encoded to obtain its embedding. Similarly, the collected data from the database is encoded using the same model to obtain embeddings for each content.
- The next step is to search for closer content from the collected data for the received user query, based on cosine similarity. The results are returned in a paginated format, providing an efficient and effective way to access relevant content.

Information Extraction Pipeline to Extract Trending Health Information

The implementation of the information extraction pipeline involves the utilization of several natural language processing techniques and previously conducted research to efficiently detect current health trends within the proposed social network.

If we consider a sentence such as "Sugar has a negative impact on diabetes, but it is beneficial for low blood sugar," a series of steps can be proposed to extract the required information from a health-based social network.

- Data preprocessing
    - Each queried content or response will be considered separately in this step.
    - First, each queried text will be processed to remove punctuation using regular expressions.
    - Next, the processed texts in the previous step will undergo a process to get the singular form of each word. It uses used NLTK WordNetLemmatizer module for this functionality.

```
def get_singular_form(word):
    # Get the singular form of a word using WordNet lemmatizer
    lemmatizer = WordNetLemmatizer()
    singular_form = lemmatizer.lemmatize(
        word, pos="n"
    )  # 'n' specifies that it is a noun
    return singular_form
```

Figure 11: Process Singular Form of a Word

    - Next, the processed texts in the previous step will undergo a coreference resolution process. It is used spaCy "neuralcoref" coreference resolution module for this functionality.

```
def coref_resolution(text):
    doc = nlp(text)
    # fetches tokens with whitespaces from spacy document
    tok_list = list(token.text_with_ws for token in doc)
    for cluster in doc._.coref_clusters:
        # get tokens from representative cluster name
        cluster_main_words = set(cluster.main.text.split(" "))
        for coref in cluster:
            if (
                coref != cluster.main
            ):  # if coreference element is not the representative element of that cluster
                if (
                    coref.text != cluster.main.text
                    and bool(
                        set(coref.text.split(" ")).intersection(cluster_main_words)
                    )
                    == False
                ):
                    # if coreference element text and representative element text are not equal and none of the coreference element words
                    # are in representative element. This was done to handle nested coreference scenarios
                    tok_list[coref.start] = (
                        cluster.main.text + doc[coref.end - 1].whitespace_
                    )
                    for i in range(coref.start + 1, coref.end):
                        tok_list[i] = ""

    return "".join(tok_list)
```

Figure 12: Process Co-Reference Resolution

- Tokenization
    - This step in processing a sentence is to tokenize it into individual words or tokens. This process, also known as tokenization, can be achieved using Python programming language's built-in methods, as outlined in the proposed solution.
    - Tokenize the sentence: ["Sugar", "is", "negatively", "affecting", "diabetes", "while", "it", "is", "good", "for", "low-blood-sugar", "."]
- Part-of-Speech (POS) Tagging
    - Apply part-of-speech tagging to identify the grammatical roles of each token. This helps distinguish between entities and sentiment-related modifiers.
    - Example POS tags:
        - "Sugar" is a noun (entity).
        - "is" is a verb.
        - "negatively" is an adverb (sentiment modifier).
        - "affecting" is a verb.
        - "diabetes" is a noun (entity).
        - "while" is a conjunction.
        - "it" is a pronoun.
        - "good" is an adjective (sentiment modifier).
        - "for" is a preposition.
        - "low-blood-sugar" is a noun phrase (entity).
        - "." is a punctuation mark.
- Dependency Parsing
    - Use dependency parsing to understand the grammatical relationships between words. This step helps establish connections between entities and sentiment modifiers.
    - Example relationships:
        - In the first clause, "Sugar" is the subject, "is" is the verb, "negatively" modifies the verb, and "diabetes" is the object.
        - The conjunction "while" connects two clauses.

- Named Entity Recognition
    - Identify and extract entities mentioned in the sentence. Use named entity recognition (NER) to recognize entities.
    - Example entities:
        - "Sugar"
        - "diabetes"
        - "low-blood-sugar"
- The above three steps, which include Part-of-Speech (POS) tagging, dependency parsing, and named entity recognition functionalities, are achieved using a third-party service called Wikifier API.
- Sentiment Analysis
    - Analyze the sentiment of each clause. Identify words or modifiers that convey sentiment.
    - Example sentiments:
        - In the first clause, "negatively" suggests a negative sentiment.
        - In the second clause, "good" suggests a positive sentiment.
- Relationship Extraction
    - Based on the dependency parsing and the understanding of the sentence, extract relationships between entities and sentiment.
    - This step clarifies how entities are connected and the sentiment associated with those connections.
    - It is used OpenNRE project [7] to achieve this functionality.
    - Example relationships:
        - In the first clause, the relationship is the negative effect of "Sugar" on "diabetes."
        - In the second clause, the relationship is the positive effect of "Sugar" on "low-blood-sugar."
- Overall Analysis
    - Combine the extracted information to create a structured representation of the sentence.
    - Consider the context and structure to provide a holistic interpretation.

    - Example structured representation:
        - Entity 1: Sugar
        - Entity 2 (Clause 1): Diabetes
        - Sentiment (Clause 1): Negative
        - Relationship (Clause 1): Affecting
        - Entity 1: Sugar
        - Entity 2 (Clause 2): Low-blood-sugar
        - Sentiment (Clause 2): Positive
        - Relationship (Clause 2): Beneficial for
- Each analyzed entity, relationship, and sentiment will be saved in the Neo4j database.

Business Evaluation

- Target Audience – Any interested user can join the social network without any age frame since it is limited to sharing only health-related content. However, 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 search results from queries using a specialized database and semantic search functionality (use of an Online Analytical Processing

database such as Apache Pinot and libraries such as SentenceTransformers).

- 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, anyone can even post something, even if it may be in assumptions.
    - The proposed information extraction pipeline does not detect some specialized keywords relating to Ayurvedic medical practices.
- 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.

Testing and Implementation

Implementation

The AyurMinds mobile application is an all-encompassing digital solution that provides a wide range of health-related services. Its features include a chatbot, a plant identification tool, a social network, and a doctor recommendation service. The application was created using React Native and is supported by various database services such as MongoDB, MySQL, Apache Pinot, and Neo4j. The system employs several machine learning, deep learning, and natural language processing solutions that were developed using the Python programming language. In addition, it has various backend APIs that were developed using Node.js and .NET. Microsoft Azure services are utilized as a cloud hosting provider to ensure optimal service delivery.

Testing

- In order to ensure optimal accuracy and performance of each sub-component within the system, a thorough test planning has been established. The testing phase will specifically evaluate the efficacy of the implemented enhanced semantic search functionality and information extraction pipeline.
- To achieve this, a custom dataset sheet has been annotated with ground truth labels for entities, sentiment, and relationships. Additionally, multiple test case scenarios have been defined, including edge cases designed to test corner cases within each sub-component. Rest assured that every effort has been made to ensure comprehensive testing and reliable results. Appropriate evaluation metrics have been used to measure the performance of natural language processing-based solutions, including precision, recall, and f1-score.
- In addition, a few test cases have been conducted to assess the anticipated results of the APIs.
- When it comes to testing system performance, scalability, and resource utilization are crucial factors to consider. It's important to test the system's ability to handle large volumes of data in order to ensure scalability. Additionally, it's important to monitor resource utilization, including CPU and memory, during testing to ensure that the system is performing optimally. It's also helpful to test the system with different input sizes to get a better

understanding of its capabilities. By taking these steps, it can ensure that the system is performing at its best and can handle any demands that may be placed on it.

- When it comes to manual testing, a thorough review of the system's output for a subset of test cases is necessary. During this process, any discrepancies between the system's output and the ground truth should be carefully addressed and resolved. This helps ensure the accuracy and reliability of the system's performance.
- By adhering to this comprehensive test planning, it is anticipated that we will be able to effectively evaluate the precision and dependability of the upgraded semantic search capabilities, information extraction pipeline, and other APIs that have been implemented. This will enable us to make well-informed decisions regarding further enhancement and optimization of these features.

Table 1: Test Case 1

Test Case ID	01
Test Case	Add new content HTTP POST API request
Test Scenario	Add new content as a question or article and make it publicly accessible.
Input	A POST HTTP request in the following format `{ "id": 0, "userId": "string", "header": "string", "body": "string", "contentType": "ARTICLE" }`
Expected Output	1. HTTP status code 201 must be returned.

	2. Added content should be displayed in the user's private content section and homepage.
Status (Pass/Fail)	Pass

Table 2: Test Case 2

Test Case ID	02
Test Case	Add a response HTTP POST API request
Test Scenario	Add a response to existing publicly available content.
Input	A POST HTTP request in the following format. `{ "contentId": 0, "parentResponseId": 0, "responseId": 0, "userId": "string", "body": "string" }`
Expected Output	1. HTTP status code 201 must be returned. 2. Added response should be displayed in the content.
Status (Pass/Fail)	Pass

Table 3: Test Case 3

Test Case ID	03
Test Case	Delete an existing public content
Test Scenario	Delete an existing public content HTTP DELETE API request.
Input	A DELETE HTTP API request with content identification

Expected Output	1. HTTP status code 201 must be returned. 2. Deleted content must be removed from the homepage but should remain in the user's private content section.
Status (Pass/Fail)	Pass

Table 4: Test Case 4

Test Case ID	04
Test Case	Delete an existing response to public content
Test Scenario	Delete an existing response to a public content HTTP DELETE API request.
Input	A DELETE HTTP API request with response identification
Expected Output	1. HTTP status code 201 must be returned. 2. Deleted response must be removed from the homepage but should remain in the content posted user's private content section.
Status (Pass/Fail)	Pass

Table 5: Test Case 5

Test Case ID	05
Test Case	Search for related public content
Test Scenario	A GET HTTP request with a query parameter to search for related content
Input	A GET HTTP API request with user's query param and expected content type
Expected Output	Related content list along with HTTP status code 200
Status (Pass/Fail)	Pass
Comments	Search functionality takes an extended period of time to return results compared to other APIs

Table 6: Test Case 6

Test Case ID	06
Test Case	Updated information in the knowledge base
Test Scenario	Test the information stored in the knowledge base against the content added
Input	Articles, questions and responses in the social network
Expected Output	Information should be updated as expected
Status (Pass/Fail)	Pass
Feedback	It is noticed a couple of ayurvedic related texts are not updated in the knowledge base

RESULTS AND DISCUSSION

Results

Information extraction pipeline

Within this particular section, we will be delving into the outcome of the methodology that has been employed to extract related entities and sentiments from a given sentence. The sentence in question is "Sugar is negatively affecting diabetes while it is good for low-blood-sugar." We shall be analyzing each clause of this sentence to identify the entities, sentiments, and relationships that have been identified.

- Result 1: Clause 1
    - Entity 1: Sugar
    - Entity 2: Diabetes
    - Sentiment (Clause 1): Negative
    - Relationship (Clause 1): Affecting
    - Discussion: In the first clause of the sentence, we identified two entities: "Sugar" and "Diabetes." "Sugar" is recognized as the subject, and "Diabetes" as the object. The sentiment analysis of this clause indicates a negative sentiment, primarily attributed to the adverb "negatively." The relationship extracted here is that "Sugar" has a negative effect on "Diabetes," implying a detrimental impact.
- Result 2: Clause 2
    - Entity 1: Sugar
    - Entity 2: Low-blood-sugar
    - Sentiment (Clause 2): Positive
    - Relationship (Clause 2): Beneficial for
    - Discussion: In the second clause of the sentence, we also identified two entities: "Sugar" and "Low-blood-sugar." The sentiment analysis of this clause indicates a positive sentiment, primarily attributed to the adjective "good." The relationship extracted here is that "Sugar" is beneficial for "Low-blood-sugar," suggesting a positive impact, possibly in terms of raising low blood sugar levels.

- Overall Discussion: The sentence as a whole presents a contrast in the effects of "Sugar" on different entities. It illustrates the complexity of the relationship between sugar and health. While it is negatively affecting "Diabetes" (likely indicating that excess sugar consumption can worsen diabetes), it is portrayed as beneficial for "Low-blood-sugar" (implying that sugar intake can help raise low blood sugar levels, which can be important for individuals with conditions like hypoglycemia).

The contrasting sentiments in the two clauses emphasize the importance of context when interpreting sentences. This methodology allowed us to systematically identify entities, sentiments, and relationships, providing a structured analysis of the sentence's meaning. However, it's essential to note that the interpretation can also depend on domain-specific knowledge and context beyond the linguistic analysis of the sentence.

Enhanced semantic search functionality

Implementing a search functionality using Sentence Transformers for semantic search is an accurate way to find relevant documents or data based on the underlying semantic meaning of the text rather than just keyword matching. In this results and discussion section, we will break down the process, present the outcomes, and analyze the implications of implementing such a system.

1. Data Preparation:

Began the task by gathering textual data from social networking sites and then proceeded to preprocess it. Then, it was required to ensure that the data was in a format that was compatible with the Sentence Transformers model. This involved a thorough examination and modification of the data to make sure that it was structured appropriately.

2. Sentence Embeddings:

To enhance the effectiveness of your corpus, you can employ a pre-trained Sentence Transformers model, specifically the all-MiniLM-L6-v2 variant. This model allows you to transform each sentence or document in your corpus into highly compact vector embeddings that encapsulate the semantic meaning of the text. The chosen model is

particularly advantageous because it delivers average accurate results and maximum performance compared to other available models.

3. User Query Processing:

When a user enters a search query, apply the same preprocessing steps to convert it into a sentence embedding.

4. Semantic Search:

Perform a semantic search by finding the closest sentence embeddings to the user query in the index. This can be done using cosine similarity or other distance metrics.

One of the most significant advantages of implementing Sentence Transformers for semantic search is the improved search relevance. Traditional keyword-based search may miss relevant documents due to different phrasing or synonyms, but semantic search captures the underlying meaning of the text, leading to more accurate results.

Semantic search reduces noise in search results. Irrelevant documents or sentences with similar keywords but different meanings are less likely to be retrieved because the model focuses on semantic content.

Depending on the choice of Sentence Transformers model and hardware, the system can be highly scalable. You can efficiently index and search through large corpora of text.

The choice of the Sentence Transformers model is critical. Some models are better suited for specific tasks or types of text. Experiment with different models to determine which one performs best for your dataset and use case. Though there are pre-trained models with much accuracy the main reason to select the aforementioned model is its performance.

When conducting a search, the use of semantic search technology can yield more precise and relevant results. However, it's important to note that this method requires a longer processing time compared to traditional keyword-based search. To ensure an acceptable response time, it's recommended to optimize the search process by utilizing tools such as Elasticsearch. This search engine, which is based on Apache Lucene, is

a cost-effective solution suitable for production environments. By implementing this approach, processing costs can be minimized without sacrificing the quality of search results.

Discussion and Finding

Information extraction pipeline

In our research, we developed and applied a comprehensive methodology for extracting related entities and sentiments from complex sentences. Our approach was demonstrated using the sentence, "Sugar is negatively affecting diabetes while it is good for low-blood-sugar." Previously. It is important to take note of the fact that the task of extracting information has been executed with the explicit purpose of gathering the following subsequent data.

- Whether a food/herb or any other consumable item is recommended for a disease.
- Any experiencing symptoms for certain diseases by user experience.

The following research findings summarize our methodology's effectiveness and its implications for natural language understanding.

1. Robust Entity Recognition:

Our methodology successfully identified entities within the sentence. In this specific sentence, "Sugar," "Diabetes," and "Low-blood-sugar" were correctly recognized.

This result suggests that entity recognition can be reliably performed in complex sentences, even when entities are mentioned in different clauses. However, it is lacking in extracting for some specialized entities in relating to Ayurvedic medical practices. Incorporating a custom named entity recognition model may can overcome such weakness.

2. Accurate Sentiment Analysis:

Sentiment analysis was effectively applied to determine the polarity of each clause. "Negatively" in the first clause and "good" in the second clause were correctly identified as sentiment indicators.

This demonstrates the importance of considering sentiment modifiers in understanding the overall sentiment of a sentence.

3. Relationship Extraction:

Our methodology successfully extracted relationships between entities in the sentence. It captured both the negative effect of "Sugar" on "Diabetes" and the positive effect of "Sugar" on "Low-blood-sugar."

This highlights the ability of the approach to capture nuanced relationships between entities and sentiment, even when they are presented in contrasting clauses.

4. Contextual Analysis:

The methodology places a strong emphasis on considering the context and structure of the sentence. This is crucial for accurately interpreting the relationships and sentiments presented in a complex sentence.

By analyzing the sentence as a whole and in the context of its individual clauses, our approach provides a more nuanced understanding of the information conveyed.

5. Implications for Domain-Specific Analysis:

Our methodology provides a foundation for understanding complex sentences in various domains. In this case, it demonstrated the relevance of sugar's effects on health-related entities (Diabetes and Low-blood-sugar).

The approach can be adapted for domain-specific applications, such as medical or financial texts, by customizing entity recognition and sentiment analysis based on the domain's terminology and sentiment lexicons.

6. Interpretation Beyond Linguistics:

While the methodology excels in linguistic analysis, it is important to acknowledge that the interpretation of entities and sentiment can also depend on domain knowledge and real-world context.

Researchers and analysts should combine the linguistic analysis provided by the methodology with domain expertise to draw accurate and meaningful conclusions.

Enhanced semantic search functionality

- Improved Relevance:

One of the primary findings of our research is that implementing a semantic search approach using Sentence Transformers significantly improves search relevance. Traditional keyword-based search often suffers from false positives and misses relevant documents due to variations in phrasing. However, the semantic search approach captures the underlying meaning of the text, leading to more accurate and relevant results.

- Reduced Noise:

Our experiments revealed a notable reduction in noise in search results when using semantic search. Irrelevant documents or sentences that share similar keywords but different meanings are less likely to be retrieved. This reduction in noise enhances the overall search experience, making it easier for users to find what they are looking for.

- Scalability:

After conducting our analysis, we have discovered that the current system's scalability is not up to par. The primary reason for this is due to the need for more computational power. This could prove to be a costly endeavor, especially in a production environment where expenses can quickly add up.

- Model Sensitivity:

Our research also highlighted that the choice of the Sentence Transformers model is critical to the system's performance. Different models exhibit variations in accuracy and computational efficiency. For instance, fine-tuned models tailored to specific domains or tasks may outperform general-purpose models in certain contexts. Researchers and practitioners should carefully select the model that best suits their dataset and use case.

- Query Processing Time:

It was observed that while semantic search provides more accurate results, it may require slightly more processing time compared to traditional keyword-based search.

The increased computational cost stems from calculating cosine similarities or other distance metrics between embeddings. However, this additional processing time is often acceptable and well worth the improvement in search quality but may be expensive in terms of cost.

- Relevance Threshold:

Our research also explored the concept of relevance thresholds, which can be adjusted to filter search results based on similarity scores. This feature allowed users to customize their search experience, ensuring that they received the most relevant documents according to their preferences. It is set to 0.75, considering the processing power required and the paginated data amount required in the current implementation.

- Practical Applications

Beyond the research findings, we discovered that this semantic search approach has practical applications in various domains. It can be used in information retrieval systems, content recommendation engines, e-commerce product searches, legal document searches, and more, where precision and relevance are critical.

CONCLUSION

The mobile application AyurMinds features a social network component that is specifically designed for the health domain.

In the modern era of technology, a health-based social network has boundless potential to transform the way we approach our well-being. Through fostering a sense of community, offering a plethora of valuable resources and information, and promoting healthy behaviors and lifestyles, such a platform can empower individuals to take control of their health and lead more rewarding lives.

A connected virtual community that offers support, encourages users to share their experiences, and offers expert guidance can effectively dismantle barriers to healthcare access and promote preventative care. Moving forward, the success of a health-based social network will hinge on its ability to maintain strong data privacy and security, provide evidence-based information, and continually adapt to meet the diverse needs of its users.

Ultimately, the aim of a health-based social network is to bridge the gap between technology and well-being, creating a space where individuals can thrive, learn, and support one another on their journey to optimal health. By facilitating communication and collaboration between people from all backgrounds, this platform has the power to make a profound and enduring impact on public health globally.

This social network utilizes two novel approaches that are tailored to meet the needs of individuals seeking to improve their health and wellness.

1. Information extraction pipeline to extract trending health information

Through our research, we have discovered that our methodology excels in extracting entities and sentiment from intricate sentences. This involves a comprehensive breakdown of each sentence, with careful attention paid to linguistic elements, in order to gain valuable insights through a thorough analysis of context. It is important to note that expertise in the relevant domain is essential to enhance interpretation of entities and sentiment in practical scenarios. Our methodology builds upon natural language understanding and text analysis techniques to support specialized social networks in

the healthcare industry. However, we have identified an area for improvement in terms of extracting specialized keywords in Ayurvedic medical practices. We believe that custom-trained models can help address this particular weakness.

2. Enhanced semantic search functionality

Incorporating Sentence Transformers into the process of semantic search can greatly enhance the quality and relevance of search results. However, it is vital to approach this technique with meticulous data preprocessing, careful model selection, and adequate computational power to fully reap its benefits. Our research has shown that leveraging Sentence Transformers for semantic search can bring about significant improvements in search relevance while reducing unwanted noise. Nonetheless, to optimize system performance, it is crucial to pay attention to key factors such as model selection, query processing time, and user feedback. This approach has immense potential for a wide range of applications and represents a promising avenue for future research and development, with a focus on balancing costs and performance metrics.

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] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "Bidirectional Encoder Representations from Transformers," *ArXiv,* 2018.

[3] A. Chiche and B. Yitagesu, "Part of speech tagging: a systematic review of deep learning and machine learning approaches," *Journal of Big Data,* 2022.

[4] M. Wankhade, A. C. S. Rao and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artificial Intelligence Review,* 2022.

[5] R. Liu, R. Mao, A. T. Luu and E. Cambria, "A brief survey on recent advances in coreference resolution," *Artifcial Intelligence Review,* 2023.

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

[7] X. Han, T. Gao, Y. Yao, D. Ye, Z. Liu and M. Sun, "OpenNRE: An Open and Extensible Toolkit for Neural Relation Extraction," in *Proceedings of EMNLP-IJCNLP: System Demonstrations*, 2019.

[8] H. Shen, G. Liu, H. Wang and N. Vithlani, "SocialQ&A: An Online Social Network Based Question and Answer System," *IEEE Transactions on Big Data,* vol. 3, no. 1, pp. 91-106, 2017.

[9] S. Šćepanović, E. Martin-Lopez, D. Quercia and K. Baykaner, "Extracting Medical Entities from Social Media," in *Proceedings of the ACM Conference on Health, Inference, and Learning*, 2020.

[10] K. Adnan and R. Akbar, "An analytical study of information extraction from unstructured and multidimensional big data," *Journal of Big Data,* 2019.

[11] "SentenceTransformers Documentation," [Online]. Available: https://www.sbert.net/.