

WriteWizard - Collaborative Document Editing Tool: Real-Time Multi-functional Platform

(Hierarchical topic modeling and suggest contributors based on users' IT expertise)

TMP-24-25J-146

Project Proposal Report

Saara M.K.F – IT21361036

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

Faculty of Computing

Sri Lanka Institute of Information Technology
Sri Lanka

August 2024

WriteWizard - Collaborative Document Editing Tool: Real-Time Multi-functional Platform

(Hierarchical topic modeling and suggest contributors based on users' IT expertise)

TMP-24-25J-146

Project Proposal Report

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

Faculty of Computing

Sri Lanka Institute of Information Technology
Sri Lanka

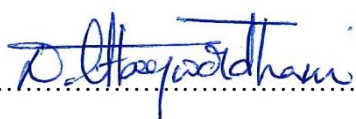
August 2024

Declaration

I declare that this is my own work, and this proposal does not incorporate without acknowledgement 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 publish or written by another person expect where the acknowledgement is made in the text.

| Name | Student ID | Signature |
|--------------|------------|--------------|
| Saara M.K. F | IT21361036 | <i>Saara</i> |

The above candidate is carrying out research for the undergraduate Dissertation under my supervision.



Signature of the Supervisor:

(Dr. Lakmini Abeywardhana)

Date: 23/08/2024



Signature of the Co-supervisor:

(Ms. Karthiga Rajendran)

Date: 22/08/2024

Abstract

This study addresses the challenge of improving topic modeling in collaborative environments by accounting for the varying expertise levels of contributors. The research problem centers on the difficulty of accurately representing the content of collaborative work when participants have different levels of expertise. To tackle this, we propose a system that utilizes Named Entity Recognition (NER) and the Nested Chinese Restaurant Process (nCRP) for expertise assessment and Hierarchical Topic Modeling.

At the outset, users will create an account on the platform by submitting their top three research works, along with personal blogs and GitHub profiles. This information serves as an initial evaluation of their expertise, which is crucial because there is no existing log history to analyze when users first joined the platform. As users continue to contribute to the platform, their activity will be tracked, and their expertise levels will be refined over time through the application of NER and nCRP technologies.

The system diagram will detail how these components interact to achieve the research objectives. The methodology includes several key tasks: collecting data, analyzing text, and implementing the NER and nCRP models. This approach aims to produce a more accurate depiction of topics that aligns with the expertise of each contributor, ultimately leading to more effective collaboration.

The final report will include an analysis of the system's performance, along with insights into the accuracy and utility of the proposed methodology in reflecting the true content of collaborative work based on contributor expertise.

Keywords: Topic modeling, expertise assessment, Named Entity Recognition (NER), Nested Chinese Restaurant Process (nCRP), hierarchical topic modeling, collaborative environments, contributor expertise, text analysis, research platform, and collaborative work.

Table of Contents

| | |
|--|-----------|
| Declaration | 3 |
| Abstract | 4 |
| Table of Contents..... | 5 |
| List of Tables | 7 |
| List of Figures | 8 |
| List of Abbreviations | 9 |
| 1.0 Introduction | 10 |
| 1.1 Background and Literature | 10 |
| 1.2 Research Gap | 12 |
| 1.3 Research Problem..... | 14 |
| 2.0 Objectives..... | 15 |
| 2.1 Main Objective..... | 15 |
| 2.2 Specific Objectives | 15 |
| 3.0 Methodology | 16 |
| 3.1 Requirement Gathering..... | 16 |
| 3.1.1 Past Research Analysis..... | 16 |
| 3.1.2 Identifying Existing System | 17 |
| 3.2 Feasibility Study | 19 |
| 3.2.1 Technical Feasibility | 19 |
| 3.2.1.1 Knowledge on Technologies | 19 |
| 3.2.1.2 Knowledge on Tool..... | 19 |
| 3.2.1.3 Data Collection Knowledge..... | 19 |
| 3.2.2 Schedule Feasibility..... | 20 |
| 3.2.3 Economic Feasibility | 20 |
| 3.3 System Analysis..... | 21 |
| 3.3.1 Software Solution Approach..... | 21 |
| 3.3.2 Tools & Technology | 23 |
| 3.4 Project Requirements | 24 |
| 3.4.1 Functional Requirements..... | 24 |
| 3.4.2 Non-Functional Requirements..... | 25 |
| 3.5 Testing..... | 26 |
| 3.6 Timeline..... | 26 |
| 3.7 Risk Management Plan..... | 26 |
| 3.8 Communication Management Plan..... | 28 |
| 3.8.1 Communication Objectives..... | 28 |
| 3.8.2 Communication Media..... | 28 |

| | |
|-----------------------------------|-----------|
| 4.0 Commercialization..... | 33 |
| 5.0 Budget | 33 |
| 6.0 Summary | 34 |
| References..... | 35 |

List of Tables

| | |
|---|----|
| Table 1: Research Gap Among Existing Topic Modeling methods with comparison to hLCTA..... | 13 |
| Table 2:Risk Management Plan | 28 |
| Table 3: Communication Media..... | 33 |

List of Figures

| | |
|--|----|
| Figure 1: High-Level System Diagram for the Research Project | 17 |
| Figure 2: TimeLine of the Project | 26 |

List of Abbreviations

| | |
|--------------|--|
| hLCTA | Hierarchical Latent Community Topic Analysis |
| LDA | Latent Dirichlet Allocation |
| LCTA | Latent Community Topic Analysis |
| pLSA | Probabilistic Latent Semantic Analysis |
| NER | Named Entity Recognition |
| nCRP | Nested Chinese Restaurant Process |
| EM | Expectation-Maximization |

1.0 Introduction

1.1 Background and Literature

Hierarchical Topic Modeling and Suggest contributor based on their IT expertise.

Topic modeling is a classic text mining task aimed at discovering hidden topics that occur within a document collection. This field, which falls under the broader domains of Machine Learning (ML) and Natural Language Processing (NLP), has been extensively studied, leading to the development of several key methodologies. Among the most prominent are Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (pLSA), Latent Dirichlet Allocation (LDA) [1], and Latent Community Topic Analysis (LCTA) [1]. These methods scan entire collections of words in documents, detect frequently used words and themes, and group them into summarized structures that best represent the underlying content. This enables a deeper understanding of large textual datasets by uncovering abstract "topics" or "hidden themes" embedded within the text.

However, each of these methods has its limitations. A critical issue, which has not been sufficiently addressed, is the assumption that all contributors to a document or set of documents possess the same level of knowledge or expertise. This assumption is often unrealistic in real-world scenarios, where individuals within a group have varying levels of skill and knowledge. Ignoring this variation can lead to inaccurate summaries and misrepresentations of the content.

For instance, consider a four-member research team contributing to a paper. Each member writes the introduction based on their understanding of the research area. At the end of the day, one member's contribution may be more insightful due to their superior knowledge, while others might lack the depth of understanding. When these contributions are summarized without considering the varying expertise levels, the final summary may not accurately reflect the most valuable insights. Prioritizing the expertise level of contributors when summarizing text can significantly improve the accuracy of topic modeling.

Recognizing the need to address this gap, I propose an extension to the existing Latent Community Topic Analysis (LCTA) called **Hierarchical Topic Modeling in Latent Community Topic Analysis (hLCTA)**. The term "Hierarchical" reflects the incorporation of a structured understanding of contributor expertise, acknowledging that people within a community possess varying levels of skills. Unlike previous models that assume uniformity in knowledge levels, hLCTA explicitly accounts for these differences, thereby enhancing the accuracy of the summarization process. This model not only eliminates redundancy and prioritizes words based on expertise but also retains all the beneficial features of existing topic modeling methods.

The introduction of hLCTA represents a significant advancement in the field of topic modeling. By incorporating a hierarchical structure that considers contributor expertise, this model addresses a critical limitation of existing methods. It improves the accuracy of topic detection and summarization in collaborative environments, making it particularly valuable for group projects where members have diverse skill levels.

The significance of this research lies in its potential to identify how we approach collaborative content creation and analysis. By improving the accuracy of topic modeling, hLCTA [10] can lead to better decision-making in fields such as education, research, and business. For fund providers, the value of this research is clear: investments in collaborative projects will yield more accurate insights, leading to more informed decisions and better allocation of resources.

On a societal level, the reduction of redundancy and enhancement of efficiency in knowledge generation can have widespread benefits. More accurate topic modeling can streamline the analysis of large textual datasets, making it easier to derive meaningful insights and drive innovation. This research, therefore, holds the potential to make significant contributions not only to the field of machine learning but also to the broader goal of improving collaborative efforts in various domains. By improving the accuracy of topic modeling through Hierarchical Topic Modeling in Latent Community Topic Analysis (hLCTA), this research not only contributes to the field of machine learning but also has the potential to enhance collaborative efforts in various domains, from academic research to industry applications. The consideration of expertise levels can lead to more informed

decisions, benefiting fund providers by ensuring that investments in collaborative projects yield more accurate and valuable outcomes. Furthermore, this approach aligns with the broader goal of reducing redundancy and enhancing efficiency in knowledge generation, ultimately benefiting society as a whole

1.2 Research Gap

In the current topic modeling research landscape, there are significant gaps exist in the contributor expertise within collaborative content creation. Most existing methods, including Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (pLSA), Latent Dirichlet Allocation (LDA) [2], and even the more recent Latent Community Topic Analysis (LCTA), operate under the assumption that all contributors to a document or set of documents possess an equal level of knowledge and expertise. This assumption fails to reflect the reality of collaborative environments, where individuals have varying levels of skill and understanding.

This problem can cause the most valuable insights from more knowledgeable contributors to be overlooked or lost when summarizing topics. Current research hasn't fully tackled this issue, which is crucial for making sure topic modeling accurately reflects the true content in collaborative work.

Filling the Gap with Hierarchical Topic Modeling in Latent Community Topic Analysis (hLCTA)

The proposed research introduces **Hierarchical Topic Modeling in Latent Community Topic Analysis (hLCTA)** to fill this gap. Unlike traditional methods, hLCTA acknowledges the diversity in expertise levels among contributors and incorporates this understanding into the topic modeling process. By doing so, it ensures that the contributions of more knowledgeable individuals are given appropriate weight, leading to more accurate and meaningful topic detection and summarization. This research aims to achieve several key objectives with the introduction of hLCTA [3]:

By considering the hierarchical structure of expertise, hLCTA improves the accuracy of topic modeling, particularly in collaborative environments where contributor knowledge levels vary.

The model reduces redundancy by prioritizing words and phrases based on the expertise of the contributors, ensuring that the most valuable insights are highlighted.

hLCTA retains all the beneficial features of existing topic modeling methods, while addressing their limitations.

The research seeks to broaden the applicability of topic modeling by providing a method that can be effectively used in various domains where collaborative content creation is common.

With the above proposed approach, the research seeks to develop a hierarchical extension to LCTA that integrates the contributor's expertise levels in the topic modeling process. Also aiming to validate the effectiveness of hLCTA through experiments and case studies [2].

| Feature | Latent Semantic Analysis (LSA) | Probabilistic Latent Semantic Analysis (pLSA) | Latent Dirichlet Allocation (LDA) | Modeling in Latent Community Topic Analysis (hLCTA) |
|-----------------------------------|---|--|---|---|
| Handling of Contributor Expertise | Treats all words equally, irrespective of the contributor's expertise. | Same as LSA; does not differentiate based on contributor expertise. | Same as LSA; assumes all contributors have equal importance. | Explicitly considers varying expertise levels of contributors, giving more weight to inputs from knowledgeable individuals. |
| Methodology | Linear algebra-based, uses Singular Value Decomposition (SVD) to find relationships | Probabilistic model derived from a latent class model, focuses on likelihoods of | Bayesian probabilistic model, assumes documents are mixtures of topics and topics | Combines hierarchical topic modeling with community detection, accounting for |

| | | | | |
|-----------------------------|--|---|---|---|
| | between documents and terms. | word distributions. | are mixtures of words. | contributors' expertise levels. |
| Accuracy of Topic Detection | Moderate, term frequency-based | Moderate, term frequency-based | Generally accurate but complex | High, reflects true content with expertise weighting |
| Scalability | Scales well, may struggle with very large data | Scales with increased computational needs | Scales well but computationally intensive | Scales effectively, higher computational demand due to added complexity |

Table 1: Research Gap Among Existing Topic Modeling methods with comparison to hLCTA

1.3 Research Problem

In collaborative projects, multiple people contribute to creating documents or content. However, not everyone in the group has the same level of knowledge or expertise on the subject matter. Some people may have a deep understanding, while others may only know the basics. When existing topic modeling methods like Latent Semantic Analysis (LSA) [9] or Latent Dirichlet Allocation (LDA) [6] are used to summarize or identify key topics in these documents, they typically assume that all contributors are equally knowledgeable.

These assumptions lead to two main problems:

1. When summarizing topics, contents from documents contributed by individuals with different expertise level, the ideas from those who have a deeper understanding of the subject can easily get lost when mixed with contributions from less knowledgeable team members. As a result, the finalized summary might not reflect the most accurate and important information.
2. When every contribution is treated as equally important, regardless of the contributor's expertise, the finalized summary might give an inaccurate picture of the documents' content. This can lead to misunderstanding/misinterpretations of the main theme of the content.

These issues are significant because they reduce the accuracy and reliability of topic modeling, especially in collaborative settings where people have different levels of expertise. The existing research methods haven't adequately addressed this problem, which is essential for ensuring that the true content and key insights in collaborative work are accurately captured.

The proposed research aims to solve these problems by developing a new approach called Hierarchical Topic Modeling in Latent Community Topic Analysis (hLCTA). This approach will consider the varying expertise levels of contributors, allowing the most knowledgeable inputs to be given appropriate weight in the summarization process. This way, the research will ensure that the topics identified are not only more accurate but also better represent the true content of the documents.

2.0 Objectives

2.1 Main Objective

To develop and validate Hierarchical Topic Modeling in Latent Community Topic Analysis (hLCTA) to improve topic detection accuracy in collaborative environments by integrating contributors' expertise levels into the topic modeling process. [4]

2.2 Specific Objectives

Develop a hierarchical model that integrates contributors' expertise level into the topic modeling framework.

Evaluate the performance of the hLCTA in comparison to traditional topic modeling methods. (LSA, pLSA, LDA)

Measure the improvement in topic detection accuracy by incorporating contributor expertise.

SMART Criteria:

- **Specific:** Focuses on integrating expertise levels into topic modeling for improved accuracy in collaborative settings.
- **Measurable:** Performance will be measured through comparison with existing methods and evaluation metrics for accuracy and redundancy.
- **Achievable:** Leverages existing text data and computational resources to develop and test the model.
- **Realistic:** The objectives are achievable with current technology and methodologies in topic modeling and machine learning.
- **Time-bound:** The study aims to complete development, testing, and evaluation phases within the specified project timeline.

3.0 Methodology

3.1 Requirement Gathering

Requirement gathering was through performing an extensive analysis of past research conducted throughout recent years, identification and analysis of the existing systems, as well as reading through a variety of online resources. Also, some real-world scenarios were used to figure out this research problem and requirement gathering. When go through the Problems related to existing Topic Modeling methods mostly went through the research papers related to them.

3.1.1 Past Research Analysis

The first step in our methodology involves conducting a thorough analysis of existing research papers for our study. This process is crucial for several reasons: this provides a deep foundation of existing knowledge, helps identify gaps in the current literature.

The primary objective of this analysis is to gather insights into the methodologies, techniques, and technologies that have been previously used in collaborative topic modeling and expertise detection. When it comes to the research papers, for topic modeling related research papers mainly focused on and assumed collaborators in the same domain have the same level of knowledge. But when we do more research, we realize among the topic modeling methods such as: LDA, pLSA, are mainly focusing on the above points but they had lack of accuracy in each level of the research project. By focusing on collaborators' expertise level, we are planning to achieve our main objective throughout our research project. We aim to understand the strength and limitations of these approaches to refine our own methodology and avoid potential pitfalls. Therefore, while identifying these research topics related to Hierarchical topic modeling, we have come up with the relevant solutions for the problems, the existing techniques currently facing.

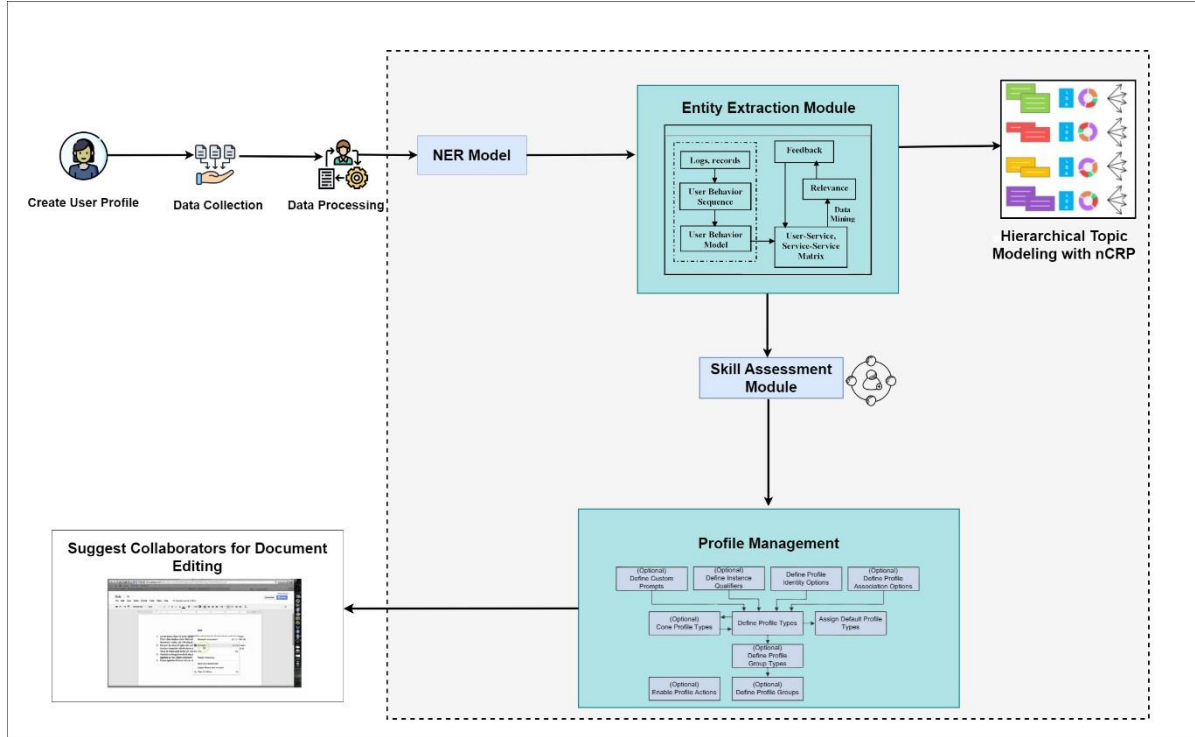


Figure 1: High-Level System Diagram for the Research Project

3.1.2 Identifying Existing System

To identify existing systems relevant to Hierarchical Topic Modeling in Latent Community Topic Analysis (hLCTA), we begin with a thorough literature review. This involves examining academic papers, articles, and case studies on key methodologies such as Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (pLSA), Latent Dirichlet Allocation (LDA), and Latent Community Topic Analysis (LCTA) [3]. The aim is to understand their approaches to topic detection and the limitations related to handling contributor expertise. A comparative analysis will follow, focusing on how these methods operate, specifically evaluating LSA's linear algebra-based approach, pLSA's probabilistic modeling, LDA's Bayesian approach, and LCTA's community detection features. This assessment will highlight the strengths and weaknesses of each method.

The next step involves assessing how these systems integrate or overlook varying levels of contributor expertise. This includes identifying gaps in how current methods address the differential importance of contributions based on expertise. We will also review the system architecture and functionality of these existing systems, such as their data preprocessing capabilities, topic extraction algorithms, and evaluation metrics, to understand their integration and deployment challenges.

Performance evaluation is critical to understand the effectiveness of existing systems. We will review performance metrics and benchmarks used for traditional topic modeling methods and gather user feedback to identify the systems' strengths and weaknesses in terms of accuracy and scalability. Additionally, understanding the user and application context through case studies will provide insights into practical challenges and unmet needs of current systems.

To stay current, we will review recent technological advancements in machine learning and

natural language processing that could enhance topic modeling. This will help assess the applicability of new technologies to existing systems. Finally, all findings will be meticulously documented, summarizing the functionalities, limitations, and gaps of existing systems, and providing recommendations for improvements based on identified deficiencies.

3.2 Feasibility Study

3.2.1 Technical Feasibility

3.2.1.1 Knowledge on Technologies

This subsection assesses our team's proficiency with essential technologies integral to the project. We are focusing on:

Named Entity Recognition (NER): NER is pivotal for extracting relevant entities from text, such as names, organizations, and locations. Our team must understand how to apply NER algorithms to identify and categorize these entities within user-generated content and research papers. [6]

Nested Chinese Restaurant Process (nCRP): The nCRP is used for topic modeling, allowing us to discover the hierarchical structure of topics in user contributions. Mastery of nCRP involves understanding its probabilistic framework and how to implement it for dynamic topic assignment.

Latent Dirichlet Allocation (LDA): LDA helps in uncovering latent topics in large collections of text. Our team needs to be adept at configuring LDA parameters and interpreting the topic distributions it generates.

Latent Community Topic Analysis (LCTA): LCTA is employed to analyze the thematic content of documents while considering the influence of community structure. Our familiarity with LCTA will enable us to model and interpret complex relationships within user data.

Probabilistic Latent Semantic Analysis (pLSA): pLSA is used for identifying underlying semantic structures in text data. Knowledge of pLSA will assist in modeling user expertise and content relevance.

Our team's ability to effectively utilize these technologies is essential for building a robust system that can accurately analyze user data, detect expertise levels, and provide meaningful insights. This competence ensures that we can implement and integrate these technologies seamlessly into our research project.

3.2.1.2 Knowledge on Tool

Here, we assess the team's expertise with the tools needed for the project, such as Apache Tomcat, data visualization software, and machine learning frameworks. Proficiency in these tools ensures that we can efficiently develop, test, and deploy our system.

3.2.1.3 Data Collection Knowledge

This section evaluates our understanding of data collection methodologies, including designing surveys and conducting interviews. Proper data collection is essential for gathering the necessary inputs to train our models and validate our system's effectiveness.

3.2.2 Schedule Feasibility

A detailed project schedule will be developed, outlining the time frame for each task and sub-task. We will use standard project management tools like Gantt charts to visualize timelines and ensure that the project stays on track. If working in a team, roles will be clearly defined, and each team member will be assigned specific sections of the project.

3.2.3 Economic Feasibility

Economic feasibility determines whether the project can be completed within the allocated budget and if it is a worthwhile investment.

Cost Estimation: Identify initial costs (e.g., hardware, software, licenses) and ongoing expenses (e.g., maintenance, salaries). Include a contingency fund for unexpected costs.

Budget Allocation: Distribute the budget among different project phases, such as research, development, and testing. Perform a cost-benefit analysis to ensure the project's benefits justify the expenses.

Funding Sources: Explore available internal funding options and seek external sources like grants or sponsorships if needed.

Financial Projections: Estimate the break-even point and return on investment (ROI) to gauge financial viability. Assess potential risks and develop strategies to manage them.

Ensuring economic feasibility helps confirm that the project is financially sustainable and provides value relative to its costs.

3.3 System Analysis

3.3.1 Software Solution Approach

The software solution approach for implementing Hierarchical Topic Modeling in Latent Community Topic Analysis (hLCTA) involves several critical steps and components. This approach aims to integrate contributor expertise levels into the topic modeling process to improve the accuracy and relevance of topic detection. Below are the detailed aspects of the software solution approach:

1. **Data Collection Module:** Gathers and preprocesses documents from various sources.
2. **Preprocessing Module:** Cleans and prepares text data for modeling.
3. **Expertise Integration Module:** Incorporates contributor expertise levels into the topic modeling process.
4. **Topic Modeling Engine:** Executes hierarchical topic modeling algorithms.
5. **Evaluation and Validation Module:** Assesses the performance and accuracy of the topic modeling.
6. **Visualization and Reporting Module:** Presents results and insights in an accessible format.

System Components

1.1 Data Collection and Ingestion

- **Objective:** Collect raw text data from various sources relevant to the collaborative documents.
- **Approach:** Implement data ingestion pipelines that extract and load documents into a centralized repository. The pipeline should handle different formats (e.g., PDFs, DOCX) and sources (e.g., web scraping, API integration).

1.2 Data Preprocessing

- **Objective:** Prepare raw text for topic modeling by cleaning and normalizing data.
- **Approach:** Develop preprocessing modules that perform text cleaning (e.g., removing stop words, special characters), tokenization, and stemming or lemmatization. The preprocessing should be customizable based on the specific requirements of the dataset.

1.3 Expertise Integration

- **Objective:** Incorporate contributor expertise levels into the topic modeling process to enhance accuracy.
- **Approach:**
 - **Expertise Assessment:** Develop a mechanism to evaluate and categorize the expertise of contributors based on predefined criteria or metadata.
 - **Weight Adjustment:** Adjust the weights of contributions in the topic modeling process according to the assessed expertise levels. This involves modifying the input data or the topic modeling algorithm to account for varying levels of expertise.

1.4 Hierarchical Topic Modeling Engine

- **Objective:** Apply hierarchical topic modeling that integrates expertise levels to detect and summarize topics accurately.
- **Approach:**
 - **Hierarchical Clustering:** Implement hierarchical clustering to organize topics into a structured hierarchy.
 - **Community Detection:** Use community detection algorithms to identify clusters of related topics and reflect contributor expertise in the hierarchical structure.
 - **Integration:** Incorporate expertise weighting into the topic detection algorithms to ensure that contributions from more knowledgeable individuals are appropriately

emphasized.

1.5 Evaluation and Validation

- **Objective:** Assess the performance of the hLCTA model in comparison to traditional methods.
- **Approach:**
 - **Performance Metrics:** Define and use metrics such as accuracy, precision, recall, and F1-score to evaluate the effectiveness of hLCTA.
 - **Validation:** Compare the results of hLCTA with those from existing methods like LSA, pLSA, and LDA. Use case studies and real-world data to validate the model's performance and reliability.

1.6 Visualization and Reporting

- **Objective:** Provide clear and actionable insights through visualization and reporting.
- **Approach:**
 - **Visualization:** Develop interactive dashboards and visualizations to present the hierarchical structure of topics, expertise impact, and other relevant metrics.
 - **Reporting:** Generate detailed reports summarizing the findings, including insights into how contributor expertise influenced the topic modeling results.

Integration and Workflow

2.1 Data Flow

1. **Data Collection:** Raw data is collected and ingested into the system.
2. **Preprocessing:** Data is cleaned and prepared for modeling.
3. **Expertise Integration:** Expertise levels are assessed and integrated into the data.
4. **Topic Modeling:** The hierarchical topic modeling engine processes the data, incorporating expertise levels.
5. **Evaluation:** The results are evaluated and validated.
6. **Visualization and Reporting:** Insights are visualized and reported to stakeholders.

2.2 Key Considerations

- **Scalability:** Ensure the system can handle increasing volumes of data and contributors.
- **Customization:** Allow for customization of preprocessing and modeling parameters based on specific use cases.
- **User Interaction:** Provide user-friendly interfaces for interacting with the system, especially for visualization and reporting.

3. Implementation Steps

1. **Design and Development:** Create detailed designs for each component and develop the necessary software modules.
2. **Integration:** Integrate the components into a cohesive system, ensuring smooth data flow and interaction.
3. **Testing:** Conduct rigorous testing to validate each component and the overall system.
4. **Deployment:** Deploy the system in the chosen environment, ensuring proper configuration and performance.
5. **Training:** Provide training and support to users to ensure effective use of the system.

This approach ensures that the Hierarchical Topic Modeling in Latent Community Topic Analysis (hLCTA) solution is well-defined, focusing on integrating expertise levels into topic modeling to enhance accuracy and relevance.

3.3.2 Tools & Technology

1. Data Collection Tools

- Survey Platforms: For gathering user feedback and requirements (e.g., Google Forms, SurveyMonkey).
- Web Scraping Tools: To collect data from web sources (e.g., BeautifulSoup, Scrapy).

2. Data Analysis Tools

- Named Entity Recognition (NER): For identifying and classifying entities in text (e.g., SpaCy, NLTK).
- Nested Chinese Restaurant Process (nCRP): For hierarchical topic modeling (e.g., custom implementations in Python/R).
- Latent Dirichlet Allocation (LDA): For topic modeling and discovering themes (e.g., Gensim library).
- Latent Community Topic Analysis (LCTA): For analyzing community-specific topics (e.g., custom algorithms or libraries).
- Probabilistic Latent Semantic Analysis (pLSA): For dimensionality reduction and topic modeling (e.g., custom implementations or libraries).

3. Development Tools

- Integrated Development Environments (IDEs): For coding and debugging (e.g., PyCharm, VS Code).
- Version Control Systems: To manage code changes and collaboration (e.g., Git, GitHub).

4. Computing Resources

- Servers: For data processing and storage (e.g., AWS, Azure).
- Local Machines: Development and testing environments.

5. Visualization Tools

- Data Visualization Libraries: For creating charts and graphs (e.g., Matplotlib, Seaborn, Plotly).
- Dashboard Tools: For presenting analysis results (e.g., Tableau, Power BI).

6. Project Management Tools

- Task Management: To track progress and assign tasks (e.g., Jira, Trello).
- Collaboration Platforms: For team communication and file sharing (e.g., Slack, Microsoft Teams).

7. Documentation Tools

- Documentation Software: For creating and maintaining project documentation (e.g., Confluence, Google Docs).

These tools and technologies will be used to ensure effective data collection, analysis, and project management throughout the research process.

3.4 Project Requirements

3.4.1 Functional Requirements

User Account Management

- **Account Creation:** Users must be able to create an account by providing their name, email, and password.
- **Profile Setup:** Users must input their top 3 research works, personal blogs, and GitHub profiles. A questionnaire based on these inputs will be provided.
- **Profile Update:** Users should be able to update their profile information, including adding new research works or blogs.

Expertise Assessment

- **Initial Expertise Evaluation:** Upon account creation, the system uses NER to extract key entities from users' submitted documents to estimate their initial expertise level.
- **Ongoing Expertise Tracking:** As users continue to work on the platform, their activity and contributions will be analyzed using NER and nCRP to update and refine their expertise levels.

Topic Modeling

- **Hierarchical Topic Modeling:** The system will apply Hierarchical Topic Modeling using nCRP to analyze and categorize documents into a hierarchical structure of topics and subtopics.
- **Dynamic Updates:** The topic models should be dynamically updated based on new contributions and expertise levels.

Data Collection and Analysis

- **Document Collection:** Users' documents and contributions need to be collected and stored securely.
- **Text Analysis:** Implement NER to extract relevant entities and nCRP for hierarchical topic modeling.
- **Reporting:** Generate reports and visualizations based on topic modeling results and expertise levels.

User Interface

- **Dashboard:** Provide a user-friendly dashboard for users to view their contributions, expertise levels, and topic models.
- **Search and Filter:** Allow users to search and filter documents based on topics, keywords, and expertise levels.

Integration

- **System Integration:** Ensure seamless integration between NER, nCRP, and Hierarchical Topic Modeling components.
- **External Data Sources:** Integrate with external data sources like GitHub for data collection and analysis.

3.4.2 Non-Functional Requirements

Performance

- **Scalability:** The system should handle a growing number of users and documents efficiently, maintaining performance.
- **Response Time:** Ensure fast response times for document processing, expertise evaluation, and topic modeling.

Security

- **Data Privacy:** Implement strong security measures to protect user data, including encryption and secure access controls.
- **Authentication:** Use secure authentication methods for user login and account management.

Usability

- **User Experience:** The interface should be intuitive and easy to navigate for users with varying technical backgrounds.
- **Accessibility:** Ensure the platform is accessible to users with disabilities, following relevant accessibility standards.

Reliability

- **System Availability:** Ensure high availability of the platform with minimal downtime.
- **Error Handling:** Implement robust error handling to manage and report any issues during data processing or user interactions.

Maintainability

- **Code Quality:** Follow best practices for code quality and documentation to facilitate future maintenance and updates.
- **Modularity:** Design the system in a modular fashion to simplify updates and integration of new features.

Compliance

- **Legal Compliance:** Ensure that the system complies with relevant data protection regulations, such as GDPR.
- **Ethical Standards:** Adhere to ethical standards in data collection, analysis, and user interaction.

3.5 Testing

Testing will involve multiple phases to ensure the reliability and accuracy of the Hierarchical Topic Modeling in Latent Community Topic Analysis (hLCTA) system. Initial testing will include unit tests on individual components, such as the Nested Chinese Restaurant Process (nCRP) and Named Entity Recognition (NER) algorithms. Integration tests will follow, focusing on how these components interact within the system, ensuring they accurately assess contributor expertise levels. Finally, system-level testing will simulate real-world scenarios to validate overall system performance and the accuracy of the topic modeling results. Feedback from these tests will be used to refine the system before deployment.

3.6 Timeline

The project will follow a structured timeline, beginning with the initial research and development of the hLCTA model. Key milestones include the completion of the feasibility study, system architecture design, and the development of key algorithms such as nCRP and NER. Subsequent milestones will cover the integration of these components, rigorous testing phases, and final system deployment. Each phase will be carefully scheduled to allow for iterative development, ensuring that adjustments can be made based on testing outcomes and peer feedback.



Figure 2: TimeLine of the Project

3.7 Risk Management Plan

The risk management plan addresses potential challenges that may arise during the project, including technical difficulties in implementing the hLCTA model, integration issues with nCRP and NER algorithms, and potential delays in the project timeline. Mitigation strategies include setting up contingency plans for technical hurdles, conducting regular reviews to

catch integration issues early, and allocating buffer time in the project schedule to accommodate unforeseen delays. The plan will also include strategies to manage the risks associated with accurately modeling contributor expertise levels in collaborative environments.

| Risk | Trigger | Owner | Response | Resource Required |
|--|--|----------------|---|---|
| Technical difficulties in implementing the hLCTA model | Issues reported during implementation phase | My Self | Set up contingency plans including alternative technical approaches and troubleshooting protocols. | Additional technical support, diagnostic tools. |
| Integration issues with nCRP and NER algorithms | Integration tests show incompatibilities or errors | My Self | Conduct regular integration reviews, engage with cross-functional teams to address issues promptly. | Access to integration specialists, testing environment |
| Panel requests changes | Panel is not satisfied with the product/presentation/outcome | Project Leader | Implement necessary changes immediately. - Update all required documents. - Communicate changes to relevant stakeholders. | Project Schedule Plan/Gantt Chart, Product Backlog, Meeting Log |
| Illness or sudden absence of project team member(s) | Illness or other personal emergencies | Project Leader | - Inform supervisor and co-supervisor. - Development team divides functions with equal scope. - Use Project Schedule Plan/Gantt Chart. - Utilize backup resources. | Project Schedule Plan/Gantt Chart, Backup resources |

| | | | | |
|--|---|----------------|---|---|
| Potential delays in the project timeline | Project milestones not met, or progress reports indicate delays | Project Leader | Allocate buffer time in the project schedule, monitor progress closely, and adjust timelines as needed. | Project management tools, additional manpower if required |
|--|---|----------------|---|---|

Table 2: Risk Management Plan

3.8 Communication Management Plan

Effective communication is critical to the success of this project. The communication management plan outlines how information will be exchanged among team members, stakeholders, and other relevant parties throughout the project lifecycle.

3.8.1 Communication Objectives

The primary communication objectives are to ensure that all team members are aligned on project goals, progress, and any issues that arise. This includes regular updates on the development of the hLCTA model, the progress of testing phases, and any adjustments to the project timeline. The plan also aims to facilitate clear and timely communication with stakeholders to ensure their expectations are managed and met.

3.8.2 Communication Media

Communication media will include a mix of formal and informal channels. Regular team meetings will be conducted via video conferencing platforms for detailed discussions and decision-making. Informal communication will occur through messaging apps like WhatsApp, following the user's preference for informal communication styles. Email will be used for official documentation and stakeholder updates, ensuring that a record of communications is maintained.

| Meeting Type | Attendees | Purpose | Frequency | Agenda Items |
|------------------------------|---|--|------------------------|---|
| Planning Kicking-off meeting | Supervisor, Co-supervisor, All Team Members | -- The project's planning phase was formally launched. Following this meeting, the project's scope and governance structure must | Once at Project Level. | --Describe the planning timetable and describe the aims, expectations, and activities of the planning phase. -- State the project scope in |

| | | | | |
|----------------------------|---|--|---|---|
| | | <p>be defined, the expectations of all significant project stakeholders must be established, along with their respective roles and duties, and all current hazards must be identified.</p> <p>--The elements will be finished in their entirety and novelty</p> | | <p>your introduction.</p> <p>-- Go over the key points of the project charter.</p> <p>-- Go over the project's overall schedule.</p> <p>-- Discuss the overall approach of the project.</p> <p>-- Talk about the project's necessary project plans.</p> <p>-- Explain assumptions, limitations, and hazards.</p> <p>-- Talk about or show off any project-supporting tools.</p> <p>-- Recap the conversation (decisions, actions, and risk)</p> |
| Executing Kick-off Meeting | Supervisor, Co-supervisor, All Team Members | <p>--The project's execution phase was formally launched. Following this meeting, the team, supervisor, and co-supervisor are aware of the project's scope, its governance structure, the duties and responsibilities of its participants, and its rules. Once at Project Level or for each major project phase. (Before</p> | Once at Project Level or for each major project phase. (Before Proposal, PP1, PP2, Final) | <p>-- Provide the Project Work Plan and the Meeting Log.</p> <p>-- The Communications Management Plan should be presented</p> <p>Agree on the process for resolving disputes and propose the escalation method.</p> <p>-- Outline the Quality Assurance &</p> |

| | | | | |
|---------------------------------|---|---|--------------|--|
| | | Proposal, PP1, PP2, Final) | | <p>Control procedures, Issue Management, and Project Change Management processes.</p> <p>-- Agree on the team's guiding principles (communication via email, meetings, phone, meeting minutes to be produced, availability, etc.).</p> <p>-- Talk about the upcoming evaluations.</p> <p>-- Recap the conversation (decisions, actions, and risk)</p> |
| Internal Project Status Meeting | All Team Members | <p>-- Go over the project's status. Discuss ongoing projects and assess development.</p> <p>-- Discuss new risks and/or issues and define action points.</p> <p>-- Examine and discuss modification requests, and if necessary, accept or reject them.</p> <p>-- Talk about the upcoming evaluations.</p> | Once a week | <p>Progress status review (presentation of periodic Project Status report).</p> <p>Accomplishments (Current and Planned actions).</p> <p>-- Actual work vs Planned.</p> <p>-- Milestones status.</p> <p>-- Current deliverables status:</p> <p>-Indicators, Existing change requests (current progress), New change requests (input from Research Panel)</p> |
| Actual Project Status Meeting | Supervisor, Co-supervisor, All Team Members | | Twice a week | |

| | | | | |
|--|---|---|---|--|
| | | | | -- Next deliverables status: -Existing change requests (Current progress), New change requests -- Risks & Issues |
| Project Review Meeting | Supervisor, Co-supervisor, All Team Members | -- A meeting discussing the status of the project. -- Major scope adjustments, a significant re-baselining of the project work plan (PWP), ensuring alignment with portfolio goals and objectives, and business strategies are among the subjects that will be covered | Quarterly of the project. (Before Proposal, PP1, PP2, Final) | The completion of required documentation. - Review of significant milestones. - Testing advancement. - Budget, resource, and other risks; issue, and action monitoring. - Panel comments. - Other: People, Resources, and Panel. |
| Project Steering Committee (PSC) Meeting | All Team Members | -- Meeting with the supervisor(s) about the status and follow-up of the project. -- This meeting is also necessary currently because: - Official project permissions are required. - Promises made. | Once a month or at the time a significant project milestone is accomplished, the supervisor must provide their approval (s) | Project debriefing: -- Results during the time period. -- Issues encountered, and solutions found. -- Important issues deserving of management's attention. -- Items that won't be completed until the following milestone or meeting. |

| | | | | |
|----------------------------|---|--|--|---|
| | | | | <p>-- Assessment of the existing situation in relation to the project's objectives, spending plan, and completion date.</p> <p>-- Official endorsements, commitments, and contractual details.</p> |
| Change Control Meeting | Supervisor, Co-supervisor, All Team Members | Discuss and prioritize change requests or panel inquiries. | There is an important requirement change after the panel discussion. | -- Discuss the panel comments and accept the change requests and start development. |
| Project End Review Meeting | Supervisor, Co-supervisor, All Team Members | <p>The objectives for the Project-End Review meeting are:</p> <p>-- Examine the key accomplishments and project performance.</p> <p>-- Talk about how the project went overall. * Talk about if the goals have been attained and, if not, why.</p> <p>-- Go over the issues and difficulties that were encountered during the project and how they were handled.</p> | Once per project or major project phase. (End of the Project) | <p>-- Evaluate the results and accomplishments of the project.</p> <p>-- Consider project-related information (budget & work history, milestones & timing history, technical & methodological approaches used).</p> <p>-- List the lessons that were learned.</p> <p>-- Plan to implement your business (change management, how to achieve desired outcomes and benefits)</p> |

| | | | | |
|--|--|---|--|--|
| | | -- Talk about best practices and lessons learned that might be used to next initiatives | | |
|--|--|---|--|--|

Table 3: Communication Media

4.0 Commercialization

The commercialization section will explore how the hLCTA system can be brought to market. This will include identifying potential users and applications of the system, such as academic institutions, research organizations, and collaborative platforms. The plan will detail strategies for intellectual property protection, potential partnerships, and avenues for generating revenue from the system. Consideration will also be given to how the system can be scaled and adapted for broader use.

5.0 Budget

The budget section will outline the financial resources required for the project, including costs associated with research, development, testing, and commercialization. This will include expenses for software, hardware, and any necessary licenses, as well as salaries for team members and any consultants. A detailed breakdown of costs will be provided, along with justifications for each expenditure to ensure the project remains financially viable.

1. Cloud Infrastructure & Hosting:
 - Cloud Servers (e.g., AWS, Google Cloud, Azure): LKR 16,000/month
 - Storage Solutions (e.g., AWS S3, Google Cloud Storage): LKR 8,000/month
 - Database (e.g., MongoDB Atlas, Firebase): LKR 7,000/month
2. Development Tools & Licenses:
 - IDE and Code Editors (e.g., JetBrains, VS Code Extensions): LKR 26,000
 - Version Control (e.g., GitHub, GitLab Premium Plans): LKR 8,000/month = LKR 96,000/year
 - API Integrations (e.g., NER API, NLP Toolkits): LKR 54,000/year
3. Security & Compliance:
 - SSL Certificates: LKR 48,000/year
4. Third-Party APIs & Services:
 - NER & NLP Services (e.g., Google NLP, SpaCy Pro, etc.): LKR 44,000/year
5. Miscellaneous Software Costs:

- Project Management Tools (e.g., Jira, Trello Premium): LKR 32,000/year

Total Estimated Cost: LKR 187,000/year

6.0 Summary

The summary will provide a concise overview of the project, reiterating the main objectives, methodologies, and expected outcomes. It will highlight the innovative aspects of the hLCTA system and its potential impact on improving topic modeling accuracy in collaborative environments. The summary will also emphasize the importance of considering contributors' expertise levels and how this project will address challenges in current collaborative work environments.

References

- [1] Grootendorst, M.P. (no date) *Hierarchical topic modeling*, *Hierarchical Topic Modeling - BERTopic*. Available at: https://maartengr.github.io/BERTopic/getting_started/hierarchicaltopics/hierarchicaltopics.html (Accessed: 17 August 2024).
- [2] alvasalvas 2 *et al.* (1959) *Latent dirichlet allocation vs hierarchical Dirichlet process*, *Data Science Stack Exchange*. Available at: <https://datascience.stackexchange.com/questions/128/latent-dirichlet-allocation-vs-hierarchical-dirichlet-process> (Accessed: 17 August 2024).
- [3] Cai, L., Zhou, G., Liu, K., & Zhao, J. (2011). *Learning the Latent Topics for Question Retrieval in Community QA* (pp. 273–281). <https://aclanthology.org/I11-1031.pdf>
- [4] Xu, Y. *et al.* (2018) ‘Hierarchical topic modeling with Automatic Knowledge Mining’, *Expert Systems with Applications*, 103, pp. 106–117. doi:10.1016/j.eswa.2018.03.008.
- [5] Yin, Z., Cao, L., Gu, Q., & Han, J. (2012). Latent community topic analysis. *ACM Transactions on Intelligent Systems and Technology*, 3(4), 1–21. <https://doi.org/10.1145/2337542.2337548>
- [6] Liu, Y., Alexandru Niculescu-Mizil, & Wojciech Gryc. (2009). Topic-link LDA. *International Conference on Machine Learning*. <https://doi.org/10.1145/1553374.1553460>
- [7] Kong, Q., Sun, J., & Xu, Z. (2024). Joint orthogonal symmetric non-negative matrix factorization for community detection in attribute network. *Knowledge-Based Systems*, 283, 111192–111192. <https://doi.org/10.1016/j.knosys.2023.111192>
- [8] J. -D. Zhang and C. -Y. Chow, "CRATS: An LDA-Based Model for Jointly Mining Latent Communities, Regions, Activities, Topics, and Sentiments from Geosocial Network Data," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 11, pp. 2895–2909, 1 Nov. 2016, doi: 10.1109/TKDE.2016.2594772.
- [9] Dao, B., Nguyen, T., Venkatesh, S. *et al.* Latent sentiment topic modelling and nonparametric discovery of online mental health-related communities. *Int J Data Sci Anal* **4**, 209–231 (2017). <https://doi.org/10.1007/s41060-017-0073-y>
- [10] T. He, L. Hu, K. C. C. Chan and P. Hu, "Learning Latent Factors for Community Identification and Summarization," in *IEEE Access*, vol. 6, pp. 30137–30148, 2018, doi: 10.1109/ACCESS.2018.2843726.