Summiva: An Enterprise-Scale NLP System for Content Summarization, Tagging, Grouping, and Search

Project Proposal & Statement of Work

Project Manager (PM):

Saikumar Bollam

Team Members:

Saikumar Bollam

POTENTIAL ADVISORS:

- *Dr. Liangming Pan (Mentor)*
- *Dr. Greg Chism (Faculty Advisor)*

Date: *January 31, 2025*

History Table

Version	***Summary of Changes***	***Date***
0.1	*Initial draft; section owners assigned and outline established.*	*01/17/2025*

0.5	*Template information removed; first draft of all sections completed.*	*01/31/2025*
0.9	*Revised sections based on mentor/advisor feedback; updated deliverables.*	*02/03/2025*

Table of Contents

1. *Executive Summary …… 1*
2. *User/Market Research …… 2*
3. *Product Features …… 3*
    - *3.1 Feature 1: Summarization Engine*
    - *3.2 Feature 2: Advanced Tagging & Grouping*
    - *3.3 Feature 3: Enterprise Search Interface*
4. *Project Timeline & Gantt Chart …… 5*
5. *Ethics …… 7*
6. *Approvals …… 9*
7. *Appendix …… 10*
    - *A. Advisor Engagement & Project Team Responsibilities*
    - *B. Ground Rules for Project Management*

---

1. Executive Summary

Written by Saikumar Bollam.

Summiva is an innovative enterprise-scale NLP system designed to extract, summarize, tag, group, and search unstructured text data from any online source. The system automatically retrieves webpage content via web scraping and applies state-of-the-art summarization algorithms—both extractive (e.g., TextRank, BERT-based methods) and abstractive (e.g., T5, PEGASUS)—to generate concise, accurate summaries. These summaries are enriched with advanced tagging through named entity recognition and topic modeling, organized using clustering algorithms, and made easily searchable via an integrated dual search engine that combines Elasticsearch (for keyword queries) with FAISS (for semantic, vector-based retrieval).

The product addresses a clear market need. Current summarization tools often focus on consumer use and lack the structured, enterprise-grade features such as robust tagging and scalable search. Summiva is intended for corporate data teams, researchers, and content marketers who require secure, efficient access to actionable insights from large volumes of unstructured data. By integrating modern NLP techniques with a rigorous ethical framework, Summiva offers a unique, innovative solution that is both practical and responsible.

Development will be carried out entirely by the student, with guidance from the mentor and faculty advisor. The work will span content ingestion, algorithm integration, UI design (if time permits), and extensive testing and ethical review. At project completion, Summiva will be delivered as a deployable web application accompanied by a detailed final report and an interactive demonstration at iShowcase.

Preliminary Subsystem Responsibilities
(Although this is a single-person project, the following table outlines the functional breakdown for clarity.)

Team Member	*Feature Responsibility*
Saikumar Bollam	*Full System Development: Summarization, Tagging, Grouping, and Search Integration*

Table 1: Preliminary Subsystem Responsibilities

---

2. User/Market Research

Written by Saikumar Bollam.

Overall Market:

The market for enterprise-grade data processing solutions is expanding rapidly. Corporations and research institutions increasingly rely on automated tools to manage unstructured data, with projections showing multi-billion-dollar growth in the global analytics and NLP market.

Existing Competitors:

Current solutions—such as consumer-focused summarizers and basic web scraping tools—do not offer the structured tagging, advanced grouping, and high-speed semantic search required by enterprise users. Summiva's integration of transformer-based models and hybrid search methods sets it apart from these tools.

User Insights:

Empathy interviews with potential users (corporate data analysts, research scientists, and content marketers) reveal key pain points:

- *Difficulty in quickly extracting key insights from large volumes of text.*

- *A need for secure, privacy-compliant tools that prevent data leakage.*
- *Frustration with existing tools that lack structured tagging and comprehensive search features. Summiva is designed to address these issues by providing an end-to-end platform that not only summarizes content but also organizes and makes it searchable in a meaningful, secure way.*

---

3. Product Features

Written by Saikumar Bollam.

3.1 Feature 1: Summarization Engine

Description:

The Summarization Engine automatically extracts and condenses content from input URLs using both extractive and abstractive techniques. It ensures that key information is retained while reducing redundancy.

Key Parameters:

Parameter	***Minimum Requirement***	***Target Maximum***	***Comments***
Summary Length	*10% of original text*	*30% of original text*	*Balances conciseness with information retention.*
ROUGE Score (ROUGE-1)	*≥0.45*	*≥0.60*	*Indicative of quality against reference summaries.*

Processing Time	*n/a*	*≤ 5 seconds per page*	*Ensures responsiveness for interactive use.*

Table 2: Summarization Engine Parameters

3.2 Feature 2: Advanced Tagging & Grouping

Description:

This module uses advanced Named Entity Recognition (NER) and topic modeling (via LDA and BERTopic) to tag and cluster content into meaningful categories. This structured metadata supports both analytics and refined search.

Key Parameters:

Parameter	***Minimum Requirement***	***Target Maximum***	***Comments***
NER Accuracy (F1-score)	*≥0.80*	*≥0.90*	*Ensures high precision and recall in entity detection.*
Clustering Consistency	*Silhouette Coefficient ≥0.3*	*≥0.5*	*Measures coherence of document groupings.*

Table 3: Tagging & Grouping Parameters

3.3 Feature 3: Enterprise Search Interface

Description:

Integrates a dual search system that employs both traditional keyword search (via Elasticsearch) and semantic search (via FAISS). This ensures that users can retrieve

documents quickly and with contextually relevant results.

Key Parameters:

Parameter	***Minimum Requirement***	***Target Maximum***	***Comments***
Search Response Time	*n/a*	*≤ 2 seconds per query*	*Critical for maintaining a responsive interface.*
Search Relevance (NDCG)	*≥0.60*	*≥0.80*	*Reflects the quality of returned search results.*

Table 4: Enterprise Search Parameters

---

4. Project Timeline & Gantt Chart

Written by Saikumar Bollam.

Below is a high-level milestone schedule for the Summiva project. A detailed Gantt chart (maintained separately) will be updated weekly.

Milestone	***Date***
Project Kickoff & Initial Planning	*01/29/2025*

Signed Proposal & SOW Approval	*02/19/2025*
Completion of Content Ingestion & Preprocessing	*03/05/2025*
Summarization Engine Prototype	*03/20/2025*
Tagging & Grouping Module Completion	*04/05/2025*
Enterprise Search & UI Integration	*04/20/2025*
System Testing, Optimization & Ethical Review	*05/05/2025*
Final Deployment & Documentation	*05/20/2025*
iShowcase Presentation	*06/05/2025*

Table 5: Milestone Schedule

Ethics

Written by Saikumar Bollam.

Summiva is built with a strong commitment to ethical data practices and privacy. In addition to the detailed design safeguards and data-handling protocols described

above, the following Ethics Chart outlines potential risks and our assessments based on the questions provided. In this chart, "Y" indicates that a risk is possible and requires mitigation, "N" indicates that the risk does not apply, and "M" (Maybe) signals a potential concern that we will monitor closely.

Ethics Chart

No.	***Question***	***Answer (Y/N/M)***	***Comments***
1	Could a user sell drugs or other illegal items on your platform?	N	Summiva is a summarization/search tool, not a marketplace.
2	Could a user of your platform engage in sex trafficking?	N	The platform does not facilitate commerce or communication channels for illicit activities.
3	Could a user sell class notes or cheat on their homework on your platform?	N	The system processes public textual content and does not support user-to-user commerce or educational cheating.
4	Could a stalker use your project to find someone?	N	Summiva processes publicly available data without creating personal profiles or contact directories.
5	Could your app be used to spy on or track individuals?	N	The system does not interface with personal devices (e.g., cameras, GPS) or provide tracking functionalities.
6	Could your app/software access the camera or microphone and record things without users being aware?	N	Summiva is solely focused on web-based text content and does not include any sensor or device integration.

7	If someone uses your platform, could they be re-traumatized or have their mental health impacted in some way?	N	The application presents neutral, summarised data rather than emotionally charged content.
8	Could your algorithm promote material that would traumatize or upset individuals?	N	The system summarizes and tags existing public content without promoting it; however, continuous monitoring will ensure neutrality.
9	Would your users be upset if the data you collect was given to someone else?	Y	Even though Summiva processes public data, any breach of confidentiality or unintended data sharing would likely upset enterprise users.
10	Could a data leak potentially lead to identity theft?	M	While the system handles primarily public data, inadvertent linking of data could increase re-identification risks; enhanced safeguards will be used.
11	If your site was hacked, would users of that product potentially lose their job, spouse, or family?	N	The platform is not directly tied to critical personal or employment information.
12	Should there be an age limitation on your product?	N	The product is designed for professional enterprise use and does not require an age limitation.
13	Could someone use your product to find, contact, and potentially commit elder abuse?	N	Summiva does not facilitate personal contact or provide sensitive individual identifiers.

14	If the data on your platform was breached, could it be used to blackmail the users?	M	Although the platform scrapes public data, there is a potential risk if sensitive information is inadvertently included; robust data controls are in place.
15	Does the existence of your project imply that a particular racial group, gender, religion or other protected category is inherently bad, gross, or unwanted?	N	The system is neutral and does not promote any biases against protected groups.
16	Could your product be used to commit hate crimes against a specific group?	N	The platform does not facilitate harmful communication or hate-related actions.
17	Does the primary content of your algorithm focus on something considered deeply unethical?	N	The content is derived from publicly available text and processed in a neutral manner.
18	Does your software contain race, gender, or other stereotypes?	M	There is potential for underlying model bias; ongoing monitoring and adjustments will be made to minimize such risks.
19	Could users of your app scam other individuals?	N	Summiva is a tool for information processing, not for facilitating scams or fraudulent transactions.
20	Is your particular algorithm biased towards predicting correctly only for one race, gender, or other group?	M	The risk of algorithmic bias exists; continuous evaluation and bias mitigation strategies will be employed.
21	Are the users of your project aware of how their data will be used?	Y	Clear consent and transparency mechanisms are integrated into the

			system for all data usage, even for public data.
22	What are the possible misinterpretations of your results? For example, would a white supremacist or misogynist be stoked about your results if they misinterpreted it?	N	The product presents neutral summaries and analytics, making misinterpretation for hate purposes unlikely.
23	Does the use or purchase of your data potentially contribute to a dangerous group or regime?	N	Data is collected only from public sources and is used strictly for summarization and analysis.
24	Could your virtual reality environment cause injury to the user?	N	Summiva does not incorporate virtual reality elements.
25	Are your study participants or users aware that their data will be collected and used?	N	Summiva processes publicly available data; however, transparency regarding data sources is maintained.
26	Does your app contain addictive design elements without benefit to the user?	N	The application is designed as an analytical tool without elements intended to promote addictive behavior.
27	Does your survey contain an aspect of compulsion or unusually large incentive that would command users to take it even if it was to their detriment?	N	No surveys or incentives are part of the platform's operation.
28	Could your research outcomes harm an individual or entity?	M	While the risk is minimal, there is some potential for indirect harm if summaries are misused; precautions will be implemented.

By integrating this Ethics Chart into our overall framework, Summiva not only meets industry best practices for ethical data use but also demonstrates proactive risk management and transparency in handling data and user privacy.

---

6. Approvals

Written by Saikumar Bollam.

The signatures below confirm that all parties approve the proposed project scope, work division, ethical guidelines, and timeline as outlined in this Statement of Work.

Approver Name	***Title***	***Signature***	***Date***
Saikumar Bollam	*Student / Project Manager*		*02/03/2025*
Dr. Liangming Pan	*Mentor*		*02/03/2025*
Dr. Greg Chism	*Faculty Advisor*		*02/03/2025*

Note: The faculty advisor's signature may be applied after grading if required.

---

7. Appendix

A. Advisor Engagement & Project Team Responsibilities

- ***Student Responsibilities:***
    - *Schedule and attend weekly meetings with the mentor and faculty advisor.*
    - *Provide regular status updates, document deliverables, and adjust the project plan as needed.*
- ***Mentor & Faculty Advisor Responsibilities:***
    - *Provide technical guidance and feedback (approximately 1 hour per week).*
    - *Review project documents and deliverables with a minimum lead time of 3 business days for each submission.*

B. Ground Rules for Project Management

- ***Focus:*** *Clearly define objectives at the start of each work session and remain on task.*
- ***Documentation:*** *Maintain thorough records of design decisions, changes, and iterations.*
- ***Transparency:*** *Update the project timeline, Gantt chart, and ethical review documents regularly.*
- ***Accountability:*** *Meet deadlines and communicate any delays or obstacles promptly.*
- ***Feedback Integration:*** *Actively incorporate mentor and advisor feedback to drive continuous improvement.*

This document represents the comprehensive project proposal and Statement of Work for Summiva. As a single-student project, all responsibilities lie with the student (Saikumar Bollam), who will work independently with guidance from the mentor and faculty advisor to ensure the successful execution and delivery of the project.