MSc Business Analytics Consultancy Project/Dissertation 2024-25

CANDIDATE #	**MXLJ4**
DATE	**01/08/2025**
WORD COUNT	**11156/12000**
TITLE OF PROJECT	
Building a Cost-Efficient Social Listening Tool for Strategic Consultancy: Cloud-Native Implementation and Automated Topic Discovery.	

Confidentiality Statement ☐

Where a project partner designates that a project should remain confidential (i.e. only made available to the supervisor & second marker), please check the box above. By doing so, this submission will remain as a record of academic work completed and will not be copied, reproduced, transferred, distributed, leased, licensed or shared with any other individual(s) and/or organisations, including web-based organisations other than the designated markers at any point in time.

Student Disclaimer:

I hereby declare that this dissertation is my individual work and to the best of my knowledge and confidence, it has not already been accepted in substance for the award of any other degree and is not concurrently submitted in candidature for any degree. It is the end product of my own independent study except where other acknowledgement has been stated in the text.

Marking Sheet – MSc Business Analytics Consultancy Project/Dissertation 2024-25

Criteria/Weight	Supervisor's comments
Topic, theoretical framework, literature & methodology (30%): • Topic is clearly identified, with well-defined boundaries. • Demonstrates knowledge of relevant theories and their limitations. • Uses current and relevant literature from reliable sources. • Develops an appropriate and adequate methodology for the topic. • Ensures methodology facilitates replication and reproducibility of results.	
Analysis and conclusions & recommendations (30%): • Uses primary and/or secondary data effectively. • Conducts rigorous analysis and interpretations. • Considers alternative interpretations/arguments. • Identifies and justifies limitations with reasonable arguments. • Draws conclusions/recommendations that are fully consistent with the evidence presented. • Demonstrates an understanding of the business context.	
GenAI Use and Critical Evaluation (10%): • Explores GenAI's potential for the project. • Tests and compares multiple GenAI tools/methods. • Evaluates AI-generated code, insights, literature reviews ...etc. • Documents when, where, and why GenAI is used. • Addresses ethical concerns such as data privacy, AI dependency, and academic integrity	
Structure, originality & presentation (10%): • Provides a concise and coherent summary. • Maintains a well-structured and logical presentation. • Demonstrates proper language use, style, graphs, tables, and referencing. • Uses appropriate and effective visualisations. • Presents meaningful business recommendations.	
Complexity of project scope & progress made towards business goals (10%): • Demonstrates progress in overcoming technical and operational challenges encountered during the project. • Shows advancement in addressing problem framing and data-related challenges.	
Project Management (10%): • Demonstrates structured project planning and management • Effectively engages with the supervisor throughout the dissertation process.	

General marking guidelines			
85 +	Outstanding work of publishable standard.	**50 - 59**	Good work which only covers a basic analysis. Some problems but no major omissions.
70 - 84	Excellent work showing mastery of the subject matter & excellent analytical skills.	**40 - 49**	Inadequate work. Not sufficiently analytical. Some major omissions.
60 - 69	Very good work. Interesting analysis with original insights. Some minor errors.	**0 - 39**	Work seriously flawed. Lack of clarity & argumentation. Too descriptive.

FINAL MARK: _____

Building a Cost-Efficient Social Listening Tool for Strategic Consultancy: Cloud-Native Implementation and Automated Topic Discovery.

1 ABSTRACT

In an era of rapidly expanding social media data, organizations increasingly seek tools to extract strategic insights efficiently and cost-effectively. This dissertation presents the design, implementation, and evaluation of a cloud-native social listening platform, developed to support exploratory research and strategic hypothesis generation within a consultancy context. The system integrates scalable services from Google Cloud Platform including BigQuery, Vertex AI, and Looker Studio to construct an end-to-end pipeline capable of ingesting, processing, and analyzing large volumes of social media content.

Key innovations include the use of embeddings-based text representations combined with K-Means clustering for topic modeling, as well as Large Language Models (LLMs) for generating human-readable topic labels. A low-code user interface built with Google Sheets enables non-technical consultants to execute complex analytical tasks without specialized programming skills. Evaluation demonstrates substantial efficiency and cost benefits, with a typical analysis of 1,000 social media items processed in under nine minutes at a cost below $0.40, compared to several hours of manual research or high-cost commercial platforms. Qualitative feedback from consultants further confirms the tool's utility for rapid hypothesis formation, while highlighting areas for future enhancement such as user interface improvements, data export capabilities, and customizable LLM prompts.

The research contributes an applied example of combining cloud-native infrastructure, modern NLP techniques, and user-centric design to enable scalable, low-cost social listening, offering a practical framework that can be extended for broader industry applications.

2 ACKNOWLEDGEMENT

I express my sincere gratitude to the remarkable individuals who have provided invaluable support and guidance throughout the various phases of this dissertation project. Without their contributions, achieving this milestone would not have been possible.

First and foremost, I extend my deepest appreciation to my UCL supervisor, Tjun Hoh. His unwavering support, insightful feedback, and open discussions were instrumental in shaping the direction and success of this research. His guidance was truly invaluable.

I am profoundly grateful to Sense Worldwide for the incredible opportunity to collaborate on this project. Special thanks are due to Leila Boushehri, my company mentor, for her insightful guidance and continuous support throughout my time with the organization. Her mentorship was crucial in navigating the practical aspects of this project. I also wish to specifically thank Freddie Gibbons, Senior Consultant, for his invaluable feedback, and for imparting essential knowledge about the intricacies of the consulting world, which significantly enriched the practical application of this work.

Lastly, but certainly not least, I wish to convey my heartfelt thanks to my family and friends for their incredible love, encouragement, and unwavering support during my Master's journey at UCL.

3 TABLE OF CONTENTS

1 ABSTRACT 3

2 ACKNOWLEDGEMENT 4

4 LIST OF TABLES 7

5 LIST OF FIGURES 7

6 INTRODUCTION 9

6.1 Background and Context 9

6.2 Sense Worldwide: A Case Context 9

6.3 Research Questions: 10

6.3.1 Central Research Question: 10

6.3.2 Sub-questions: 10

7 LITERATURE REVIEW 11

7.1 Social Listening and Strategic Insight Generation 11

7.2 Analytical Architectures for Social Media Data 11

7.3 NLP Models for Topic Detection and Sentiment Analysis 12

7.4 Comparative Review of Existing Social Listening Platforms 12

7.5 Research Gap and Contribution 13

8 METHODOLOGY AND SYSTEM DESIGN 14

8.1 Research Approach 14

8.1.1 Linking Methodology to Research Questions 14

8.2 System Architecture Design 15

8.2.1 High-Level System Overview 15

8.2.2 System Layers 17

8.3 Data Model Design 19

8.4 Data Collection Strategy 22

8.4.1 Selection of Data Sources 22

8.4.2 Data Types Collected 23

8.4.3 Web Scraping Methods 24

8.4.4 Data Volume 25

8.4.5 Ethical and Legal Considerations 25

8.4.6 Data Quality and Preprocessing 26

8.5 Data Analysis Methods 27

8.5.1 Embeddings 29

8.5.2 Topic Modeling (K-Means) 29

8.5.3 Sentiment Analysis 31

8.5.4 LLM-based Topic Interpretation (Gemini) 31

8.5.5 Dimensionality Reduction (UMAP) 32

8.6 Tools and Technologies 32

8.7 Implementation Approach and Phases 34

8.7.1 Implementation Phases 34

8.8 Evaluation Strategy 34

8.8.1 Utility for Exploratory Research and Strategic Insight Generation ... 35
8.8.2 Ease of Use and Cognitive Load ... 35
8.8.3 Efficiency (Time Savings) ... 35
8.8.4 Cost-Effectiveness ... 35

9 IMPLEMENTATION ... 36

9.1 User Interface and Visualization ... 36

9.1.1 Google Sheets Control Panel ... 36
9.1.2 Looker Studio Dashboard ... 40

10 RESULTS AND EVALUATION ... 45

10.1 Cost and Efficiency Gains ... 45

10.1.1 Cost Analysis ... 46
10.1.2 Time Efficiency ... 46
10.2 User Experience and Thematic Feedback ... 47

10.3 Overall Discussion ... 48

10.3.1 Cost-Efficiency vs. Analytical Depth ... 48
10.3.2 Speed vs. Interpretive Nuance ... 48
10.3.3 Accessibility and Cognitive Load ... 48
10.3.4 Strategic Use Cases ... 49
10.4 Limitations ... 49

10.4.1 Narrow Consultant Sample ... 49
10.4.2 Partial Usability Evaluation ... 49
10.4.3 Limited Platform Scope ... 49
10.4.4 Limited Clustering Methodology ... 50
10.4.5 Incomplete Data from Selective Scraping ... 50
10.4.6 Low-Code Design Choices ... 50
10.5 Future Work ... 51

10.5.1 Data Expansion and Completeness ... 51
10.5.2 Interface and Usability Enhancements ... 51
10.5.3 Advanced Analysis Features ... 51

11 GENAI USAGE ... 52

11.1 Writing Support: Multi-LLM Approach ... 52

11.2 Code Generation Analysis ... 52

11.3 Topic Labels Generation and evaluation ... 53

11.4 Ethical Considerations and Research Integrity ... 54

12 REFERENCES ... 54

13 APPENDIX ... 57

13.1 unified_social_content_items view ... 57

13.2 Data Infrastructure and Workflow Snapshots ... 60

13.3 Project Management ... 65

13.3.1 Critical Reflections ... 65
13.3.2 Project timeline ... 65
13.3.3 Meeting Notes ... 69
13.4 Source Code Repository ... 74

4 LIST OF TABLES

TABLE 7.1 COMPARATIVE ANALYSIS OF SOCIAL LISTENING PLATFORMS 13
TABLE 8.1 MAPPING OF RESEARCH QUESTIONS TO METHODOLOGY COMPONENTS. 14
TABLE 8.2 COMPONENTS OF THE PRESENTATION AND CONTROL LAYER 17
TABLE 8.3 COMPONENTS OF THE ORCHESTRATION AND INGESTION LAYER 17
TABLE 8.4 COMPONENTS OF THE DATA STORAGE LAYER 18
TABLE 8.5 CORE ANALYTICAL COMPONENTS AND TASKS 18
TABLE 8.6 VISUALIZATION TOOLS AND FEATURES 18
TABLE 8.7 SEARCH ENGINE RESULT TABLES 19
TABLE 8.8 SCRAPING JOB METADATA TABLE 19
TABLE 8.9 RAW PLATFORM DATA TABLES 20
TABLE 8.10 SCHEMA FOR UNIFIED_SOCIAL_CONTENT_ITEMS VIEW 21
TABLE 8.11 ML & VISUALIZATION TABLES 21
TABLE 10.1 ESTIMATED COST BREAKDOWN FOR TYPICAL ANALYTICAL REPORT 46
TABLE 10.2 ESTIMATED AUTOMATED PROCESSING TIME PER REPORT 46
TABLE 10.3 SUMMARY OF USER TESTING FEEDBACK (CONSULTANT PERSPECTIVE) 47
TABLE 11.1 EVALUATION OF GEMINI-GENERATED TOPIC LABELS ACROSS CLUSTERS. 53

5 LIST OF FIGURES

FIGURE 8.1 HIGH LEVEL OVERVIEW 15
FIGURE 8.2 DETAILED SYSTEM ARCHITECTURE 16
FIGURE 8.3 THE DATA ANALYSIS PIPELINE. 28
FIGURE 9.1 SEARCH LINKS SHEET - USER-DEFINED QUERIES INITIATE AUTOMATED LINK RETRIEVAL. 37
FIGURE 9.2 SEARCH RESULTS SHEET - CONSULTANTS VERIFY AND SELECT LINKS. 38
FIGURE 9.3 SOCIAL SCRAPER SHEET - INITIATING CONTENT SCRAPING FOR SELECTED URLS. 38
FIGURE 9.4 SCRAPED CONTENT SHEET - REVIEWING INGESTED SOCIAL CONTENT. 39
FIGURE 9.5 TOPIC MODELER SHEET - TRIGGERING THE END-TO-END ANALYSIS PROCESS. 39
FIGURE 9.6 RUN SUMMARY - OVERVIEW OF CONTENT SCOPE AND SOURCES. 40
FIGURE 9.7 TOPICS OVERVIEW - LLM-ENHANCED CLUSTER LABELING SUPPORTS RAPID INTERPRETATION. 41
FIGURE 9.8 TOPIC TREND - VISUALIZING TOPIC EVOLUTION AND EMERGENCE. 42
FIGURE 9.9 SENTIMENT BREAKDOWN - HIGH-LEVEL AND TOPIC-SPECIFIC SENTIMENT VIEWS. 42
FIGURE 9.10 TOPIC MAP - UMAP-BASED CLUSTER VISUALIZATION FOR SEMANTIC EXPLORATION. 43
FIGURE 9.11 WORD CLOUD – A VISUAL SUMMARY OF DOMINANT LEXICAL PATTERNS 44
FIGURE 9.12 CONTENT EXPLORER – REVIEWING ORIGINAL TEXT WITH ANNOTATIONS. 45
FIGURE 13.1 BIGQUERY SERP_SEARCH TABLE – THIS TABLE LOGS ALL SEARCH ENGINE QUERIES MADE THROUGH THE BRIGHTDATA SERP API, INCLUDING KEYWORDS, TIMESTAMPS, ESTIMATED RESULT COUNTS, AND SEARCH METADATA. 61
FIGURE 13.2 BIGQUERY SERP_RESULT TABLE – STORES INDIVIDUAL SEARCH RESULTS (URLS, TITLES, METADATA) RETRIEVED PER SERP QUERY, ENABLING LINK TRACKING AND SELECTION FOR SCRAPING. 61
FIGURE 13.3 BRIGHTDATA WEB SCRAPER DASHBOARD – OVERVIEW OF REDDIT AND QUORA SCRAPING TASKS, SHOWING THE NUMBER OF URLS PROCESSED, SUCCESS RATE, AND DATA DELIVERY METHOD. 62
FIGURE 13.4 BIGQUERY SCRAPE_JOB TABLE – MAINTAINS RECORDS OF SCRAPING JOBS TRIGGERED FROM SELECTED LINKS, INCLUDING DATASET AND SNAPSHOT IDENTIFIERS FOR TRACEABILITY. 62
FIGURE 13.5 BIGQUERY REDDIT_DATA TABLE – CAPTURES RAW SCRAPED REDDIT CONTENT INCLUDING POST BODIES, COMMENTS, ENGAGEMENT METRICS, AND AUTHOR METADATA. 63

Figure 13.6 Query Result View – Unified content extracted from Reddit and Quora, including text, timestamps, and community-level metadata to support downstream NLP analysis. 63
Figure 13.7 BigQuery embeddings_cache Table – Stores generated text embeddings and sentiment scores for each content item, avoiding redundant computation. 64
Figure 13.8 BigQuery kmeans_run Table – Metadata repository tracking topic modeling runs, cluster counts, and associated parameters. 64
Figure 13.9 BigQuery document_topic_assignments Table – Links each content item to its assigned topic cluster, supporting both analytics and visualization 65
Figure 13.10 Color-coded task categories 66
Figure 13.11 Miro Timeline May 5 - May 9 66
Figure 13.12 Miro Timeline May 12 - May 16 66
Figure 13.13 Miro Timeline May 19 - May 23 66
Figure 13.14 Miro Timeline May 26 - May 30 67
Figure 13.15 Miro Timeline June 2 - June 6 67
Figure 13.16 Miro Timeline June 9 - June 13 67
Figure 13.17 Miro Timeline June 16 - June 20 67
Figure 13.18 Miro Timeline June 23 - June 27 67
Figure 13.19 Miro Timeline June 30 - July 4 68
Figure 13.20 Miro Timeline July 7 - July 11 68
Figure 13.21 Miro Timeline July 14 - July 18 68
Figure 13.22 Miro Timeline July 21 - July 25 68
Figure 13.23 Miro Timeline July 28 - August 1 68

6 INTRODUCTION

6.1 BACKGROUND AND CONTEXT

In today's data-driven business landscape, understanding consumer sentiment, market trends, and emerging opportunities is essential for strategic decision-making. Social media platforms have become a vast, dynamic repository of public opinion and real-time discourse, offering consultancies a unique lens into consumer perspectives. However, this data is inherently unstructured, voluminous, and ephemeral posing significant challenges for extraction and interpretation.

Traditional market research methods such as surveys and interviews offer depth but are often time-consuming, expensive, and unable to capture the spontaneous nature of online conversations. Manual social media monitoring, while more agile, is typically inefficient, lacks scalability, and is labor-intensive. For strategy consultancies operating in fast-paced environments, these limitations can delay critical insights and hinder responsiveness.

Consequently, there is a growing need for more scalable, systematic, and cost-effective methods of harnessing social media intelligence tools that can rapidly synthesize large volumes of text into actionable insights without sacrificing interpretability or precision.

6.2 SENSE WORLDWIDE: A CASE CONTEXT

Sense Worldwide, a London-based strategic innovation consultancy established in 1999, is known for its co-creation methodologies and commitment to integrating unconventional perspectives into the innovation process. The company specializes in understanding consumer behavior to help clients develop forward-thinking products, services, and experiences.

A core strength of Sense Worldwide lies in The Sense Network a global community of over 6,000 diverse creators from 1,300+ cities (Donkin, 2010). This network serves as a source of cognitive diversity, fueling the firm's ability to identify future trends and deliver tailored strategic insights. The company's approach blends survey-based research and qualitative analysis to generate meaningful client recommendations, and its roster includes leading brands such as Nike, PepsiCo, and Samsung.

6.3 RESEARCH QUESTIONS:

Against this backdrop, this dissertation explores how an intelligent and cost-efficient platform can be developed to address the analytical bottlenecks associated with traditional and manual social media analysis. Specifically, the project investigates the design and deployment of a Google Cloud-based social listening system tailored to the needs of a strategy consultancy.

6.3.1 Central Research Question:

How can a Google Cloud-based social listening platform, designed for cost-effective and on-demand execution, be developed to effectively gather and analyze social media data, thereby augmenting traditional market research and accelerating the generation of strategic consumer insights for a consultancy?

6.3.2 Sub-questions:

- What architectural design and data pipeline components are optimal for a scalable and cost-effective social media data ingestion and processing system built on Google Cloud Platform?
- How can Natural Language Processing (NLP) techniques particularly embeddings, topic modeling, and Large Language Models be leveraged to identify key themes, assess sentiment, and derive interpretable insights from unstructured social media data?
- How does the implementation of such a tool enhance efficiency, depth of insight generation, and cost-effectiveness when integrated into existing market research workflows?

This dissertation aims to demonstrate the design, implementation, and evaluation of such a solution, contributing both to academic discourse on cloud-native analytical systems and to the practical toolkit of modern consultancies seeking actionable insights from digital conversations.

7 LITERATURE REVIEW

7.1 SOCIAL LISTENING AND STRATEGIC INSIGHT GENERATION

Social media platforms have become dominant arenas for the expression of public opinion, consumer sentiment, and cultural discourse. The academic field of social media analytics has emerged to process this unstructured data into usable information (Fan & Gordon, 2014). Within this field, social listening refers not merely to monitoring brand mentions, but to extracting thematic and emotional patterns that inform strategic decisions (Culnan et al., 2010; Chaffey & Ellis-Chadwick, 2019).

Whereas traditional market research methods (e.g., surveys, interviews) remain important for hypothesis testing and structured feedback (Malhotra, 2019), they often lag behind in speed and breadth. Social listening can capture real-time, unsolicited perspectives from diverse consumer groups, enabling firms to surface emergent needs and behaviors (Kaplan & Haenlein, 2010). However, the volume, velocity, and variety of social data (Laney, 2001) introduce analytical challenges that require scalable technical architectures and advanced computational methods.

7.2 ANALYTICAL ARCHITECTURES FOR SOCIAL MEDIA DATA

The processing of social media data demands robust data engineering pipelines capable of handling semi-structured text, diverse schemas, and fluctuating data quality (Kamburugamuve et al., 2018). Engineering solutions typically consist of ingestion (via APIs or scraping), transformation (ETL/ELT), storage, and analysis.

Recent research emphasizes the role of cloud-native services in overcoming infrastructure complexity. Tools like Google Cloud Platform's BigQuery, Pub/Sub, and Cloud Functions enable modular and scalable data operations (Lakshmanan, 2022). Unlike monolithic systems, cloud-native pipelines allow asynchronous ingestion, serverless compute execution, and on-demand scalability ideal for time-sensitive consultancy workflows.

Yet beyond engineering, the ability to derive actionable insights depends on the sophistication of the NLP methods used particularly for identifying topics and sentiment in unstructured conversations.

7.3 NLP Models for Topic Detection and Sentiment Analysis

A growing body of research has explored the use of machine learning and NLP to analyze user-generated content. Historically, Latent Dirichlet Allocation (LDA) was the standard for topic modeling (Blei et al., 2003), providing interpretable but often coarse-grained clusters. Recent advances favor the use of text embeddings (Mikolov et al., 2013; Devlin et al., 2019) combined with clustering algorithms such as K-Means, which allow richer semantic modeling.

Moreover, tools like BERTopic (Grootendorst, 2022) integrate transformer embeddings with class-based TF-IDF and HDBSCAN to produce high-quality, dynamically labeled topics. Although more interpretable than vanilla K-Means, BERTopic is resource-intensive and less suitable for low-latency cloud environments.

Sentiment analysis, too, has evolved from simple polarity classification to multi-dimensional models capturing emotional nuance. Pre-trained APIs like Google Cloud Natural Language or open frameworks like VADER offer off-the-shelf solutions but may lack adaptability to niche domains (Liu, 2012).

To bridge the interpretability gap, recent work has turned to Large Language Models (LLMs) such as GPT-4 and Gemini, which can assign human-readable labels to cluster outputs (Brown et al., 2020). These models can interpret embeddings and topic clusters at scale, enabling faster insight generation with minimal manual intervention.

7.4 Comparative Review of Existing Social Listening Platforms

The commercial landscape for social listening is saturated with feature-rich but often rigid tools. Platforms such as Brandwatch, Sprinklr, and Talkwalker offer real-time monitoring, influencer tracking, and sentiment analytics tailored for brand and PR teams (Rapp & Schlerf, 2019). However, these tools exhibit major limitations for strategy consultants who require deeper, more customizable insight generation.

Table 7.1 Comparative Analysis of Social Listening Platforms

Feature	*Brandwatch / Sprinklr*	*Open-source (e.g., Twint)*	*This Platform*
NLP Flexibility	Low (pre-set models)	High	High
Cost Structure	High (SaaS licensing)	Free	Pay-per-use
Custom Topic Modeling	No	Yes	Yes
Cloud-Native Integration	Minimal	Complex to set up	Full (GCP)
End-User Control	Low	High (but technical)	High (low-code)

These enterprise platforms are often closed-box systems with limited support for advanced NLP customization. Consultants are unable to modify topic models, adjust clustering algorithms, or interact with raw text embeddings. Additionally, they are prohibitively expensive for many small firms or freelance analysts.

In contrast, open-source frameworks such as Twint, Scrapy, spaCy, and Gensim provide flexibility but require significant development resources and infrastructure management unsuitable for non-technical teams or rapid-use cases (Gensim, n.d.).

7.5 Research Gap and Contribution

Despite the evolution of commercial and open-source solutions, a gap remains for cloud-native, LLM-integrated, and consultancy-oriented social intelligence platforms. There is a lack of:

- Cost-efficient systems that use serverless architectures to minimize overhead.
- Pipelines that embed modern NLP models (embeddings, K-Means, LLMs) directly into data workflows.
- Tools that combine low-code UIs with customizable AI for non-technical consultants.
- Solutions tailored to episodic, project-based analysis, as opposed to continuous brand monitoring.

This dissertation responds to that gap by designing and evaluating a modular platform that integrates cloud infrastructure (Google Cloud), advanced analytics (BigQuery ML, Gemini LLMs), and a user-friendly interface (Google Sheets + Looker Studio). The

resulting system offers a scalable and replicable blueprint for consultants to generate strategic insights from social media conversations without high technical or financial barriers.

8 METHODOLOGY AND SYSTEM DESIGN

8.1 RESEARCH APPROACH

This project uses a Design Science Research (DSR) methodology, which focuses on solving real-world problems by developing and evaluating functional artifacts (Hevner et al., 2004; Peffers et al., 2007). This approach aligns with the goal of building a practical social listening tool for Sense Worldwide, enabling both applied insight and scholarly contribution.

8.1.1 Linking Methodology to Research Questions

This dissertation is guided by three main research questions. Table 8.1 below summarizes how each methodological component and system feature directly addresses these questions.

Table 8.1 Mapping of research questions to methodology components.

Research Question	Methodological Approach / System Component
RQ1: What architectural design and data pipeline components are optimal for a scalable and cost-effective social media data ingestion and processing system built on Google Cloud Platform?	• System Architecture Design (Section 8.2) • Detailed descriptions of orchestration, ingestion, storage, and processing layers • Cloud-native design choices explained for scalability and cost-efficiency
RQ2: How can Natural Language Processing (NLP) techniques particularly embeddings, topic modeling, and Large Language Models be leveraged to identify key themes, assess sentiment, and derive interpretable insights from unstructured social media data?	• Data Analysis Methods (Section 8.5) • Embeddings generation using Vertex AI • K-Means topic modelling • LLM-based topic interpretation (Gemini) • Sentiment analysis pipeline
RQ3: How does the implementation of such a tool enhance efficiency, depth of insight generation, and cost-effectiveness when integrated into existing market research workflows?	• Evaluation Strategy (Section 8.8) • Cost and time savings analysis • Usability testing with consultants • Comparison to manual methods and commercial tools

8.2 SYSTEM ARCHITECTURE DESIGN

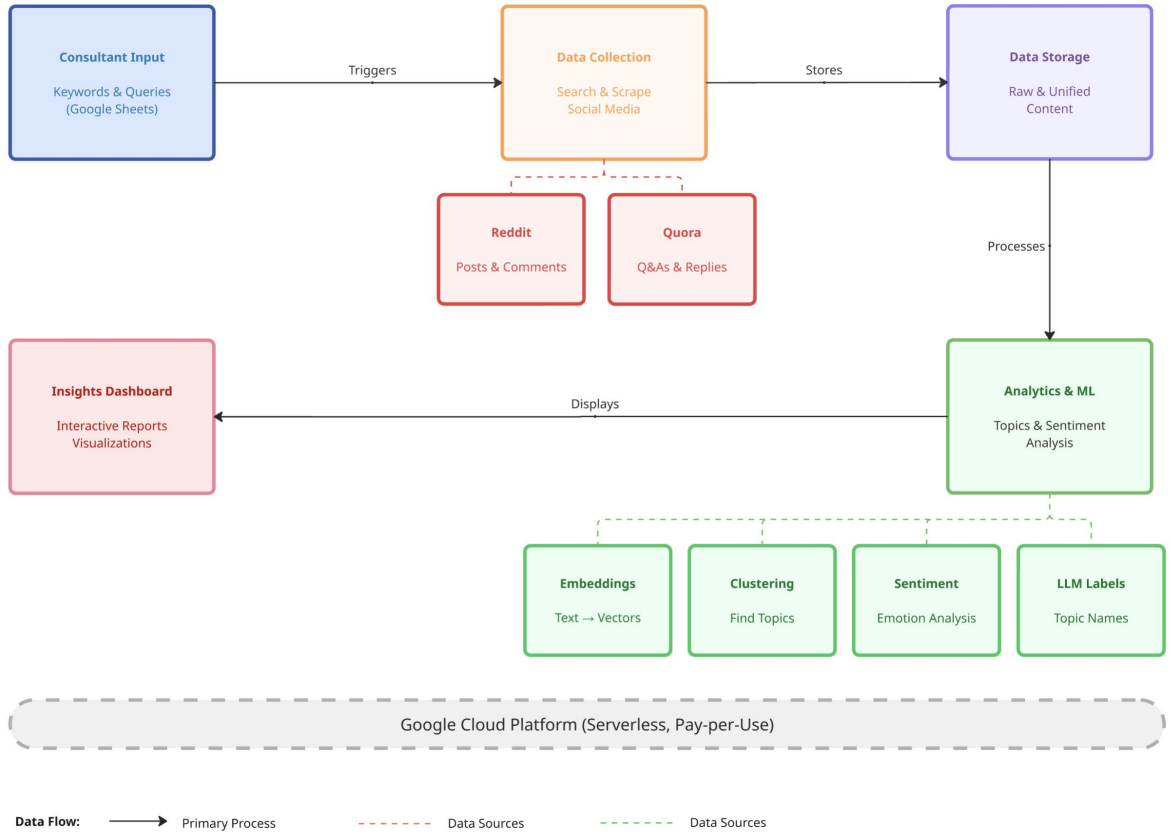


Figure 8.1 High Level Overview

8.2.1 High-Level System Overview

The social listening platform is a data processing pipeline that takes search queries from consultants and delivers topic insights from social media discussions. The system operates entirely in the cloud and follows a straightforward flow: collect → store → analyze → visualize.

What the system does: Consultants input keywords through a Google Sheets interface. The system searches for relevant social media posts, scrapes the content and comments, analyzes the text to identify discussion topics and sentiment, then presents the findings in interactive dashboards.

How it works: The platform uses web scraping to gather social media content, stores everything in a cloud database, applies machine learning to find patterns and topics in the text, and creates visualizations that consultants can explore. Each component is designed to handle workloads on-demand, scaling up or down based on usage.

Key design principle: The system prioritizes cost-efficiency through serverless, pay-per-use cloud services. Rather than maintaining always-on infrastructure, components activate only when needed, making it practical for consulting projects of varying sizes.

Figure 8.1 depicts the high-level system overview. To better understand how these components interact, the next Figure 8.2 illustrates the detailed system architecture, showing how the layers and services integrate.

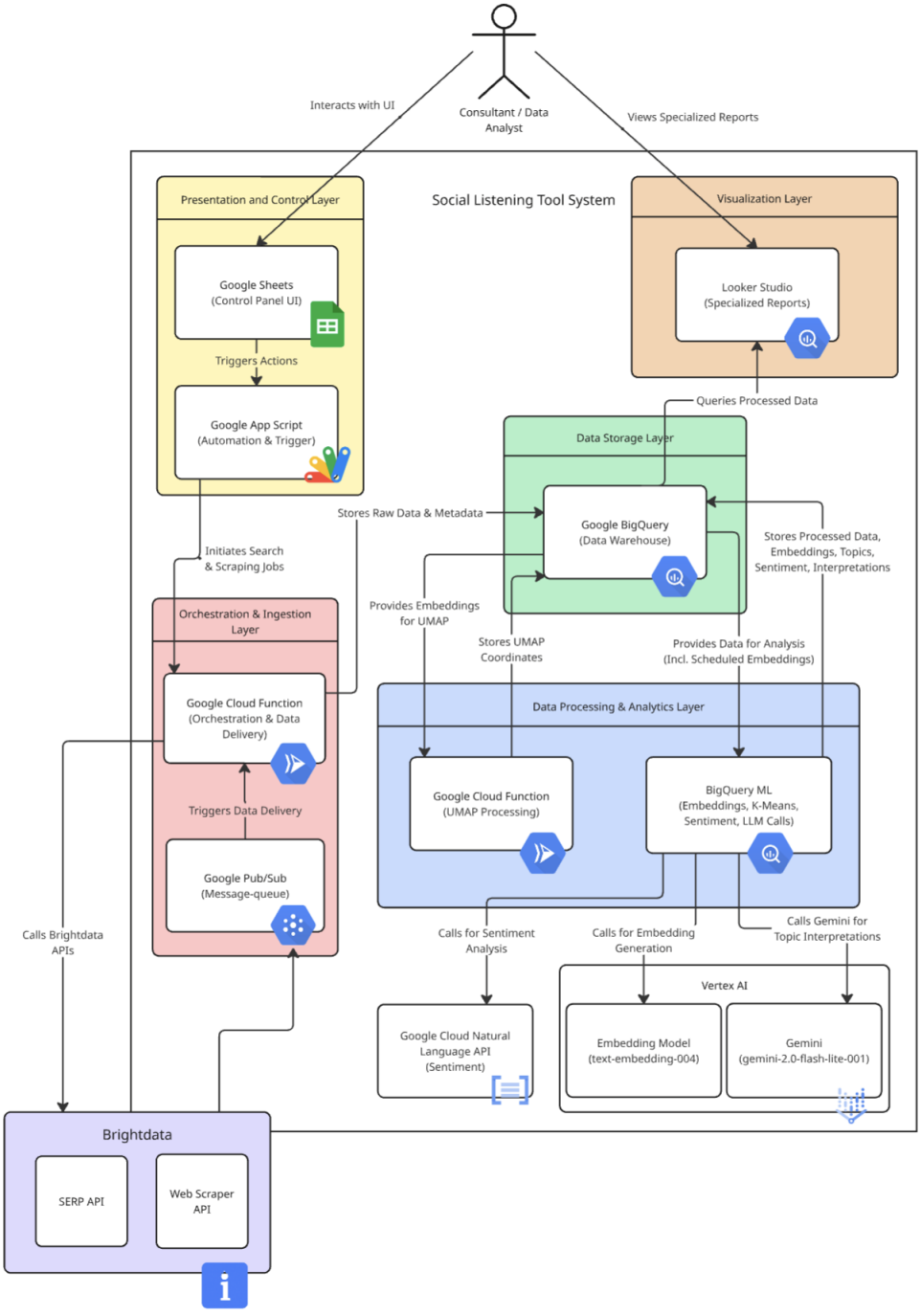


Figure 8.2 Detailed System Architecture

8.2.2 System Layers

The system is structured into five layers, each responsible for a distinct function in data collection, transformation, analysis, and visualization.

8.2.2.1 Presentation and Control Layer

Table 8.2 Components of the Presentation and Control Layer

Component	*Purpose*	*Key Benefits*
Google Sheets	Control Panel UI	• Intuitive, low-code interface • Familiar to end users • Reduces frontend development costs
Google Apps Script	Automation & Triggers	• Translates user actions into API calls • Orchestrates backend workflow

Consultants interact with the control panel built in Google Sheets to input search criteria and initiate data processing. Apps Script provides automation logic. It translates user actions (e.g., button clicks) into programmatic triggers, initiating Cloud Function API calls and orchestrating workflow across system layers.

8.2.2.2 Orchestration and Ingestion Layer

Table 8.3 Components of the Orchestration and Ingestion Layer

Component	*Purpose*	*Key Benefits*
Google Cloud Functions	Orchestration & Data Delivery	• Serverless model • Executes Brightdata calls • Efficient, pay-per-use model
Google Pub/Sub	Messaging Queue	• Reliable asynchronous messaging • Decoupled, scalable architecture
Brightdata	SERP & Scraper APIs	• Managed web scraping • Reduces need for in-house infrastructure

Brightdata performs external web searches and scrapes content from social media sources. The scraped data is delivered directly to a Pub/Sub topic, which acts as a message queue and ingestion buffer. Upon message arrival, a Cloud Function is automatically triggered. This function parses the content and inserts it into the appropriate BigQuery raw data tables using associated job metadata (e.g., *snapshot_id*, *dataset_id*).

8.2.2.3 Data Storage Layer

This layer functions as the central repository for raw and processed data, supporting both dynamic querying and scalable storage.

Table 8.4 Components of the Data Storage Layer

Component	*Purpose*	*Key Benefits*
Google BigQuery	Data Warehouse	• Serverless and scalable • Pay-per-query • Manages raw and analytical tables

8.2.2.4 Data Processing and Analytics Layer

Table 8.5 Core Analytical Components and Tasks

Task	*Technology*	*Description*
Embeddings Generation	BigQuery ML + Vertex AI (text-embedding-004)	Converts text into dense vectors for clustering and semantic search
Clustering	BigQuery ML (K-Means, Cosine Distance)	Organizes embeddings into topical groups
Sentiment Analysis	BigQuery ML + Cloud NLP API	Assigns sentiment scores to content
Topic Labeling	Gemini LLM (Vertex AI)	Interprets and names clusters using representative samples
Dimensionality Reduction	Cloud Function (Python + UMAP)	Projects high-dimensional embeddings into 2D for visualization

The embeddings generation only runs on newly ingested data using a scheduled query, preventing redundant computations and optimizing cost.

8.2.2.5 Visualization Layer

The final layer renders analytical outputs into interactive dashboards that empower consultants to explore, interpret, and act on social media insights.

Table 8.6 Visualization Tools and Features

Component	*Purpose*	*Key Features*
Looker Studio	Interactive Dashboards	• Real-time BigQuery connection • Topic visualizations • Sentiment trends • UMAP clusters • Gemini-generated summaries

Each dashboard corresponds to a specific analysis run (identified by a *run_id*) and presents outputs including topic distributions, sentiment distribution, and LLM-generated topic labels and summaries all in a user-friendly, no-code format.

8.3 DATA MODEL DESIGN

The data architecture consists of three core layers:

1. **Raw Data Tables:** preserving the original structure and lineage of ingested data.
2. **Intermediate Transformation Layer:** consolidating and standardizing platform-specific data into a unified structure.
3. **Analytics Tables:** optimized for machine learning and visualization pipelines.

8.3.1.1 Raw Data Tables

Raw data tables store unprocessed information exactly as ingested from external sources, maintaining data lineage and enabling reprocessing.

8.3.1.1.1 Search Engine Results

Captures metadata and outcomes from automated web searches to systematically discover relevant social media content.

Table 8.7 Search Engine Result Tables

Table	*Purpose*	*Key Fields*
serp_search	SERP query metadata	request_id, search_query, search_engine, timestamp
serp_result	Individual search results	serp_request_id (FK), link, title, rank, global_rank

The *serp_search* table maintains an audit trail of all search queries, including search terms, target engine, and execution timestamp. The *serp_result* table stores individual links returned by queries with their original ranking and metadata.

8.3.1.1.2 Scraping Management

Orchestrates the collection of social media content from discovered URLs through batch processing workflows.

Table 8.8 Scraping Job Metadata Table

Table	*Purpose*	*Key Fields*
scrape_job	Batch scraping requests	snapshot_id, dataset_id, urls_in_batch, total_urls_count

The *scrape_job* table manages content extraction operations. The *dataset_id* identifies platform scraper configuration, while *snapshot_id* provides unique identification for each scraping session.

8.3.1.1.3 Platform Specific Content

Stores raw social media content in native hierarchical structure, preserving original context and relationships.

Table 8.9 Raw Platform Data Tables

Table	*Source*	*Content Structure*
reddit_data	Reddit	post_id, title, user_posted, post_karma, comments (nested), community_name
quora_data	Quora	post_id, header, author_name, views, upvotes, top_comments (nested)

The *reddit_data* table captures complete Reddit post structure including hierarchical comment threads, engagement metrics, and community metadata. The *quora_data* table stores questions and answers with engagement signals like views and upvotes, maintaining the nested comment structure.

8.3.1.2 Unified Content View: Intermediate Transformation Layer

Between the raw ingestion layer and analytics layer lies the *unified_social_content_items* view a logical construct that standardizes, flattens, and integrates content across all platforms.

This BigQuery view dynamically compiles social media content from Reddit and Quora into a clean, consistent schema for downstream analytical tasks such as embeddings generation, sentiment scoring, and topic modeling.

Key Design Features

- **Hierarchical Flattening:** Nested comment/reply structures are normalized into discrete rows, each representing a standalone content item.
- **Schema Standardization:** Data from Reddit and Quora is unified under a shared schema, agnostic of the originating platform.
- **Stable ID Generation:** Unique identifiers are generated using deterministic hash functions (e.g., *FARM_FINGERPRINT*) based on content features, ensuring consistency even when re-scraped.
- **Version Control & Deduplication:** Uses *ROW_NUMBER()* windowing to retain the most recent instance of each content item, filtering out stale duplicates.

Table 8.10 Schema for unified_social_content_items View

Field	*Type*	*Description*
source	STRING	Platform name (e.g., Reddit, Quora)
content_type	STRING	Type (post, comment, reply)
content_item_id	STRING	Unique ID across platforms/types
parent_content_item_id	STRING	Parent content's ID (e.g., post ID for comment)
top_level_post_id	STRING	ID of root post/question
content_item_url	STRING	Direct URL
author_username	STRING	Content creator username
primary_text	STRING	Core text content
full_text_context	STRING	Extended contextual text
engagement_score	INT64	Standardized metric (upvotes, likes)
content_timestamp	TIMESTAMP	Content creation time
community_or_channel_name	STRING	Subreddit or Quora topic
hashtags	ARRAY<STRING>	Extracted hashtags
mentions	ARRAY<STRING>	Extracted mentions
media_urls	ARRAY<STRING>	Associated image URLs
video_urls	ARRAY<STRING>	Associated video URLs
record_load_timestamp	TIMESTAMP	Ingestion timestamp
snapshot_id	STRING	Scraping batch ID (links to scrape_job)

The full SQL definition is provided in Appendix 13.1 unified_social_content_items view.

8.3.1.3 Analytics Tables

Contains processed data optimized for machine learning operations and dashboard visualization, storing outputs from embeddings generation, clustering, and dimensionality reduction.

Table 8.11 ML & Visualization Tables

Table	*Purpose*	*Key Fields*
embeddings_cache	Pre-computed ML features	unified_id, embeddings, sentiment_score, embedding_model_name
document_topic_assignments	Clustering results	content_item_id, topic_id, run_id
kmeans_run	Clustering metadata	run_id, num_clusters, parameters
topic_labels	Human-readable topics	topic_id, label, description, run_id
document_umap_coordinates	Visualization data	content_item_id, x_coordinate, y_coordinate, run_id

The *embeddings_cache* table stores pre-computed text embeddings and sentiment scores, eliminating redundant ML service calls. The *document_topic_assignments* table links content to topic clusters from K-means operations.

The *kmeans_run* table provides metadata for each clustering execution, enabling reproducibility and comparison of different configurations. The *topic_labels* table transforms cluster identifiers into human-interpretable descriptions using large language models.

The *document_umap_coordinates* table stores dimensionally-reduced embeddings optimized for interactive dashboard visualization, enabling exploration of topic relationships and clustering patterns.

For detailed schema previews and sample records from key tables, refer to Appendix 13.2 Data Infrastructure and Workflow Snapshots.

8.4 Data Collection Strategy

This section outlines the strategy employed for collecting the raw social media data necessary for the platform's development and analytical capabilities. The primary objective was to acquire publicly accessible, text-rich social media content that is suitable for Natural Language Processing (NLP) and capable of yielding deep consumer insights for strategic consulting purposes. The subsequent subsections detail the rationale behind the selection of specific data sources, the types of data collected, the methods utilized for web scraping, and the ethical and legal considerations observed.

8.4.1 Selection of Data Sources

The primary social media data sources selected for this project are Reddit and Quora. This selection was guided by several key criteria aligned with the project's objectives and the specific needs of Sense Worldwide:

- **Rich Textual Content:** Both platforms are highly text-centric, providing extensive user-generated content (posts, comments, answers) that is ideal for textual analysis using Natural Language Processing (NLP) techniques like embeddings, topic modeling, and sentiment analysis.

- **Diverse Perspectives and Niche Communities:** Reddit, with its vast array of subreddits, offers access to highly engaged communities discussing diverse and often niche topics, providing granular insights into specific interests, pain points, and emerging trends. Quora, as a question-and-answer platform, provides direct insights into user questions, problems, and collective knowledge, often with more structured, opinion-based answers.
- **Public Accessibility:** Content on both platforms is largely publicly available, facilitating collection through legitimate scraping methods.
- **Relevance to Consultancy Insights:** The nature of discussions on Reddit (e.g., product feedback, brand sentiment, grassroots community trends) and Quora (e.g., common problems, solutions, expert opinions) directly lends itself to the type of consumer insights and "points of view" that strategy consultancies seek.

8.4.2 Data Types Collected

The data collection strategy focused on extracting comprehensive and granular information from selected social media platforms to facilitate rich textual analysis and insight generation. For each content item (posts, questions, comments, and replies), the system aimed to capture fields critical for both identification and analytical depth.

The specific types of data collected from both Reddit and Quora include:

- **Content Identification:** Unique identifiers for each post/question and its nested comments/replies (*post_id, comment_id*), along with the direct URL of the content item.
- **Core Textual Content:** The main body of the post/question (*text_content, post_text*), titles (*title*), and the full text of all comments and replies. This forms the primary input for Natural Language Processing.
- **Author Information:** Usernames of the content creators (*author_name, user_posted, commenter_name*).
- **Temporal Data:** Timestamps indicating when the content was posted, commented, or replied (*timestamp, post_date, comment_date, date_of_reply*).
- **Engagement Metrics:** Quantitative indicators of popularity or interaction, such as upvotes, scores, views, and shares.

- **Contextual Metadata:** Information providing context to the content, including the subreddit name (*community_name*), question categories, or other platform-specific attributes.
- **Hierarchical Structure:** Critical for understanding conversations, the raw data captures the nested relationships between posts, comments, and replies, preserving the conversational flow for later flattening and analysis.

8.4.3 Web Scraping Methods

The data collection process primarily leverages Brightdata's comprehensive suite of web scraping services to efficiently and reliably acquire social media content. This approach minimizes the need for developing and maintaining complex in-house scraping infrastructure, contributing to the project's development efficiency.

The methodology involves two main phases:

- **Link Acquisition:** The process begins with identifying relevant social media post URLs. The system utilizes Brightdata's SERP API to perform targeted Google searches (e.g., employing the "*site:reddit.com [keyword]*" filter) to retrieve a curated list of social media links relevant to specific keywords or topics. For Quora, a unique challenge arises as its content is not consistently indexed by Google; thus, Quora links are typically provided manually by the user within the system's control panel.
- **Content Scraping:** Once relevant URLs are identified, Brightdata's specialized Web Scrapers (specifically configured for Reddit and Quora) are invoked. These managed scrapers are designed to handle various complexities inherent in web data extraction, including dynamic content loading, anti-bot measures, and IP rotation. They are configured to navigate the specified URLs and extract the detailed content types as outlined in Section Data Types Collected.

Data Delivery Mechanism: Crucially, upon the completion of a scraping job, Brightdata's scrapers are configured to directly publish the collected data as messages to a Google Pub/Sub topic. This asynchronous delivery mechanism ensures reliable and decoupled transfer of raw content to the system's ingestion pipeline for subsequent processing.

8.4.4 Data Volume

The social listening tool is designed for flexible, on-demand data collection. Users can specify the number of search results to retrieve from the SERP API (which defaults to 20 links but is user-adjustable), and directly provide lists of relevant social media URLs for scraping. While there is no strict limit imposed by the system itself on the number of links a user can request, the practical constraint is defined by Brightdata's web scraper batch limit of 1,000 links per single batch operation.

A typical, targeted batch scraping request, reflecting the immediate analytical needs of a consultant, often involves retrieving content associated with approximately 50 relevant social media posts. This commonly translates to collecting a combined volume of 300 to 500 individual content items (including posts, comments, and replies) per query execution. This flexible, yet typically moderate-scale, acquisition strategy aligns with the system's design for cost-efficient, on-demand insights. The underlying cloud infrastructure and Brightdata's services are inherently scalable, allowing for the processing of much larger data volumes if required for future expansion or extensive research initiatives, up to the defined batch limits.

8.4.5 Ethical and Legal Considerations

The data collection process for the social listening tool adheres to stringent ethical guidelines and legal considerations, particularly concerning the acquisition and use of publicly available social media content.

- **Terms of Service (ToS) Compliance:** A thorough review of the Terms of Service for all targeted platforms (e.g., Reddit, Quora) and the data acquisition service (Brightdata) was conducted to ensure compliance with their respective policies. The data collected is limited to publicly accessible content, and operations are conducted in a manner that respects platform guidelines.
- **Data Privacy and Anonymity:** The project's focus is on analyzing broad trends, themes, and collective opinions rather than individual user profiling. Data collection is strictly limited to publicly available information. No personally identifiable information (PII) beyond publicly visible usernames (which are treated as pseudonyms for analytical purposes) is actively sought, stored, or processed. Insights generated prioritize aggregate patterns over individual attribution.

- **Responsible Data Use:** The collected social media data is intended exclusively for internal strategic analysis by Sense Worldwide. It will be used to enhance client insights and inform strategic recommendations. There is no intent for public redistribution of the raw data, nor for any use that would unfairly disadvantage individuals or violate their privacy expectations.
- **Legal Compliance:** The data collection methodology operates under general principles of legal compliance regarding the scraping of publicly available web data. By utilizing a managed third-party service like Brightdata, which typically handles compliance aspects for its scraping operations, and focusing on publicly accessible information for analytical purposes, potential legal complexities are mitigated while ensuring responsible data acquisition.

8.4.6 Data Quality and Preprocessing

Ensuring the quality and relevance of collected social media data is critical for producing reliable insights. The system incorporates multiple mechanisms for handling noise, low-quality content, and off-topic material at both automated and manual stages of the pipeline.

At the automated level, the unified_social_content_items view in BigQuery inherently filters out empty records and incomplete content items, ensuring that only documents with substantive text are passed downstream for analysis. Additionally, data transformations standardize field formats and eliminate duplicates through deduplication logic in SQL queries.

Beyond automated cleaning, the platform relies heavily on consultant-driven curation as a key quality assurance step. Within the Google Sheets control panel, several mechanisms enable consultants to filter and refine data prior to analysis:

- Search Results Sheet: Consultants manually review the list of social media URLs retrieved via SERP queries. Irrelevant, spam-like, or off-topic links can be excluded from further processing at this stage.
- Scraped Content Sheet: After data is scraped, consultants inspect the raw textual content directly within the spreadsheet. Low-value posts for example, very short

comments, posts with little engagement, or irrelevant discussions can be manually filtered out or removed prior to initiating the analytical pipeline.

This human-in-the-loop approach leverages consultants' domain expertise to ensure that the final dataset reflects the research context and project objectives. It also mitigates the risk of including noisy or off-topic data that could distort clustering or sentiment results.

While no universal thresholds (e.g. minimum upvote count or text length) are enforced automatically in the current prototype, the system's design provides the flexibility for future enhancements such as:

- Automatic exclusion of posts below configurable engagement thresholds (e.g. upvotes or scores).
- Filtering out extremely short posts under a specified character count.
- Pattern-based removal of known spam or bot-generated content.

Collectively, these measures help maintain the validity and analytical usefulness of the collected data, ensuring that downstream insights are based on high-quality, relevant textual information.

8.5 Data Analysis Methods

This section presents the end-to-end analytical methodology used to process, analyze, and extract insights from unified social media text data. The goal of this pipeline is to convert raw, unstructured textual content into structured, interpretable insights that inform strategic decision-making. The analysis leverages a combination of state-of-the-art machine learning models, natural language processing (NLP) techniques, clustering algorithms, and large language model (LLM) interpretation.

The process begins with generating semantic embeddings from raw social media text using pre-trained transformer models. These embeddings form the foundation for downstream tasks such as topic modeling, sentiment analysis, and dimensionality reduction. Clustering techniques are then applied to identify coherent thematic groupings within the data. Representative documents from each cluster are selected and interpreted using a large language model, providing meaningful topic labels and

summaries. In parallel, sentiment scores are computed to assess the emotional tone of the content. Finally, dimensionality reduction enables the visualization of semantic relationships among posts for qualitative exploration.

Figure 8.3 below illustrates the full data analysis pipeline. It provides a high-level overview of the analytical stages from raw text ingestion to the final topic interpretations highlighting the tools, models, and outputs involved at each step.

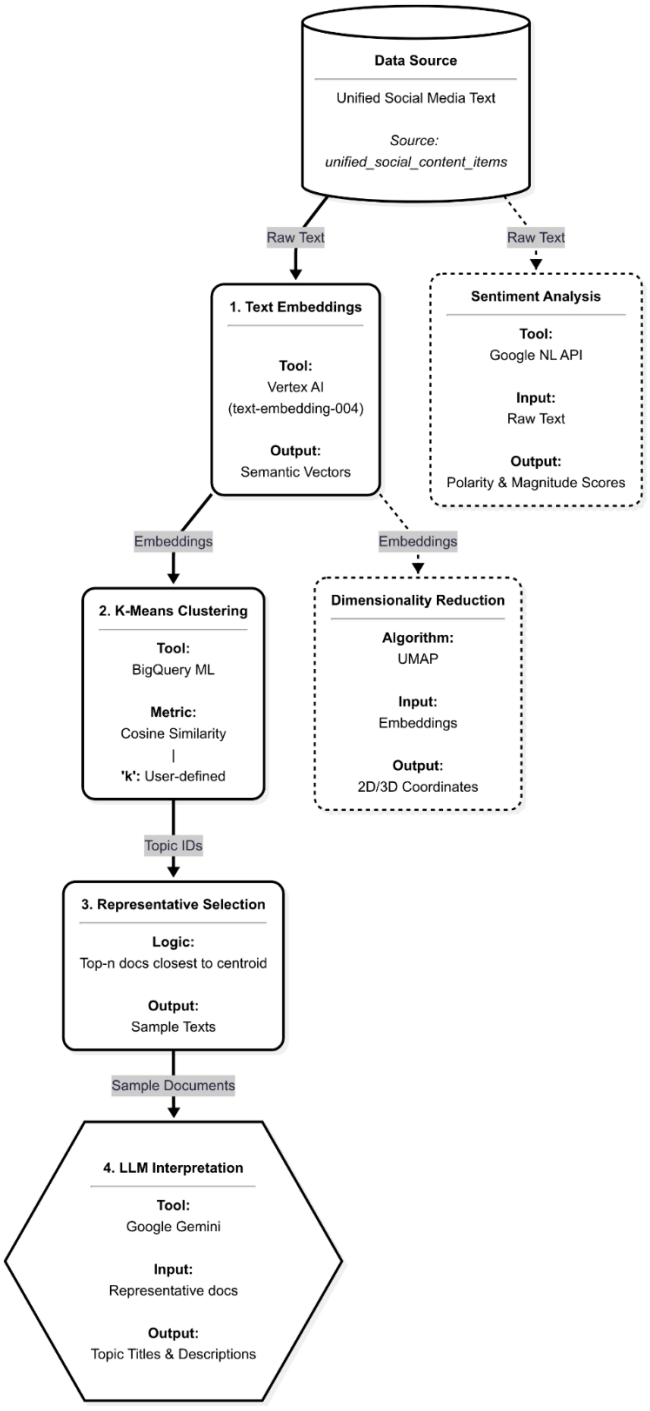


Figure 8.3 The Data Analysis Pipeline.

8.5.1 Embeddings

Following the preparation of unified social media content, the first analytical step involves the generation of text embeddings. Embeddings are dense numerical vector representations of text (words, phrases, or entire documents) that capture their semantic meaning. They are crucial for advanced text analysis because they transform unstructured textual data into a mathematical format, allowing computational models to perform operations like calculating similarity, clustering, and classification.

In this system, embeddings are generated by BigQuery ML, invoking the *text-embedding-004* foundation model served by Google Vertex AI. This process is applied to the *primary_text* field of each content item from the *unified_social_content_items* view. The use of this sophisticated pre-trained model allows for direct processing of raw text, as discussed, and ensures that the generated vectors effectively capture the rich contextual and semantic information necessary for subsequent analytical steps like topic modeling and dimensionality reduction.

8.5.2 Topic Modeling (K-Means)

After generating numerical embeddings that represent the semantic content of each social media item, the system proceeds with topic modeling to identify latent themes within the dataset. K-Means clustering is employed for this purpose, operating directly on the generated text embeddings.

K-Means is an unsupervised clustering algorithm that partitions n observations (in this case, content item embeddings) into k distinct clusters, where each observation belongs to the cluster with the nearest mean (centroid). Each cluster, defined by its centroid, represents a coherent topic or theme derived from the content items it contains.

The decision to use K-Means instead of alternative algorithms such as BERTopic or HDBSCAN was driven primarily by architectural and scalability considerations. K-Means is natively supported within BigQuery ML, enabling clustering to be performed directly inside the cloud data warehouse. This approach minimizes data movement, reduces operational complexity, and supports efficient, near-real-time analysis on large datasets. While methods like BERTopic or HDBSCAN offer advantages in discovering arbitrarily shaped clusters or automatically determining the number of clusters, they require significantly more computational resources and lack seamless integration with

the serverless BigQuery environment. For this dissertation's goal of developing a cost-effective, scalable, and cloud-native solution, K-Means offers a practical balance of performance, interpretability, and architectural simplicity.

The selection of the number of clusters (k) is intentionally flexible in this platform. Rather than imposing a fixed value, the system allows end-users particularly consultants to define k dynamically for each analytical run. This is achieved through the Google Sheets control panel, where users input their desired number of clusters before launching topic modeling. This design choice reflects the varied nature of strategic consulting projects, where different research contexts may require different levels of thematic granularity. For example, a broad market scan might use a lower k to produce high-level themes, while an in-depth investigation could specify a higher k to uncover finer distinctions. While future enhancements could include automated techniques for estimating the optimal k (e.g., Elbow Method, Silhouette Score), the current approach prioritizes user discretion and adaptability to specific research needs.

In summary, K-Means clustering was selected for its:

- Efficiency and Scalability: The algorithm's computational efficiency, coupled with native support in BigQuery ML, makes it highly suitable for processing large datasets directly within the cloud warehouse, minimizing data movement and optimizing performance.
- Simplicity and Interpretability: K-Means provides straightforward cluster assignments that form a clear foundation for further, more sophisticated interpretation.
- Suitability for Embeddings: It performs effectively on dense vector representations like embeddings, particularly when using appropriate distance metrics such as cosine similarity.

The output of the K-Means process is a set of topic IDs, where each content item is assigned to a specific cluster. The interpretation of these numerical clusters into human-understandable topics is performed in subsequent steps, leveraging advanced AI capabilities such as the Gemini Large Language Model for topic labeling and summarization.

8.5.3 Sentiment Analysis

To gain deeper insights into the emotional tone and public perception surrounding specific topics and content, the system performs sentiment analysis on the social media data. Sentiment analysis, or opinion mining, is a Natural Language Processing (NLP) technique used to determine the emotional polarity (positive, negative, or neutral) and strength (magnitude) expressed within a piece of text.

For this project, sentiment analysis is applied to the *primary_text* of each content item. This provides a crucial layer of understanding regarding collective opinions, brand perception, or reactions to specific events or discussions. The analysis leverages a sophisticated, pre-trained NLP model capable of discerning nuances in language to provide accurate sentiment scores, contributing significantly to the actionable insights delivered by the platform.

8.5.4 LLM-based Topic Interpretation (Gemini)

While K-Means clustering effectively groups semantically similar content items into numerical topics, transforming these clusters into human-understandable and actionable insights requires an additional interpretative step. This system leverages the capabilities of a Large Language Model (LLM), specifically Google's Gemini, for automated topic interpretation.

The role of the LLM is to bridge the gap between the purely statistical output of clustering and intuitive human understanding. It operates by analyzing a selection of representative documents (e.g., those mathematically closest to a cluster's centroid) for each identified topic. The LLM then generates concise, human-readable topic titles and brief topic descriptions that summarize the overarching theme of the cluster.

This innovative, LLM-based approach significantly enhances the quality and speed of insight generation compared to traditional manual topic labeling methods. It automates a resource-intensive task, provides contextually rich interpretations, and accelerates the path for consultants to rapidly grasp the essence of newly discovered themes, thereby deriving deeper consumer insights.

8.5.5 Dimensionality Reduction (UMAP)

To facilitate the visual exploration of content relationships and topic clusters, the high-dimensional text embeddings are subjected to dimensionality reduction. Dimensionality reduction is a technique used to reduce the number of features (dimensions) in a dataset while preserving its most important information.

For this project, UMAP (Uniform Manifold Approximation and Projection) is employed. UMAP is a non-linear dimensionality reduction algorithm chosen specifically for its superior ability to preserve both the local (within-cluster proximity) and global (overall arrangement of clusters) structure of the high-dimensional data in a lower-dimensional space (typically 2D or 3D). This is crucial for creating intuitive and accurate scatter plots in the visualization layer.

By projecting the embeddings into a visually navigable space, UMAP enables consultants to intuitively explore the semantic relationships between individual content items, identify the boundaries of topics, and gain a qualitative understanding of the overall data landscape within the interactive dashboard.

8.6 TOOLS AND TECHNOLOGIES

The platform was built using a cloud-native technology stack designed for scalable processing, advanced NLP, and a low-code interface. Technologies were selected to support each stage of the data pipeline, from ingestion to visualization.

Key technologies are grouped below by their functional roles:

Data Warehousing and Compute

- **Google BigQuery:** A serverless, scalable data warehouse used for storing, transforming, and analyzing large volumes of social media data.
- **Google Cloud Functions:** Serverless compute for handling event-driven tasks such as data ingestion, orchestration, and UMAP dimensionality reduction.

AI and Machine Learning Capabilities

- **BigQuery ML:** Enables machine learning tasks directly within BigQuery, including embeddings generation and K-Means clustering.

- **Google Vertex AI:** Provides access to advanced AI models, specifically used for integrating Gemini LLM for topic labeling.
- **Google Cloud Natural Language API:** Performs sentiment analysis on collected textual data.

Orchestration and Messaging

- **Google Pub/Sub:** Manages asynchronous messaging between components, particularly for scraping data delivery.
- **Google Apps Script:** Powers automation and user-triggered actions within Google Sheets for pipeline control.

Visualization and User Experience

- **Looker Studio:** Connects directly to BigQuery for building interactive dashboards and visual reports.
- **Google Sheets:** Acts as a low-code control panel enabling consultants to initiate analyses and review outputs.

External Data Acquisition Services

- **Brightdata:** Provides managed services for web scraping, including the SERP API and pre-built scrapers for Reddit and Quora, enabling reliable data collection without maintaining custom scraping infrastructure.

Programming Libraries and Languages

Development relied primarily on Python for scripting and custom components, leveraging libraries such as:

- **google-cloud-bigquery** for interacting with BigQuery
- **flask** for creating HTTP endpoints in Cloud Functions
- **requests** for external API calls
- **numpy, pandas, scikit-learn** for data manipulation and analysis
- **umap-learn** for dimensionality reduction of embeddings

8.7 Implementation Approach and Phases

The project followed an iterative development process. Different parts of the system were built and tested at the same time, instead of one after another. This made it easier to spot problems early, fix them quickly, and make sure everything worked well together.

8.7.1 Implementation Phases

- **Phase 1: Foundation & Data Ingestion Setup:**
  Set up the Google Cloud environment, configured access, and integrated Brightdata. Built Cloud Functions for search initiation, scraping, and sending raw data to BigQuery.
- **Phase 2: Data Modeling & Transformation:**
  Designed the system's core data structures, including the unified content view and embeddings cache. Added automated MERGE queries to keep embeddings and sentiment data up to date.
- **Phase 3: Analytics & AI Integration:**
  Used BigQuery ML for K-Means clustering. Integrated Google Vertex AI (Gemini) for topic summaries and the Natural Language API for sentiment analysis. UMAP was run via Cloud Functions for dimensionality reduction.
- **Phase 4: Interface & Visualization:**
  Built a Google Sheets-based control panel using Apps Script for pipeline control. Created Looker Studio dashboards for visualizing results.
- **Phase 5: Testing & Refinement:**
  Carried out functional and performance testing, debugging, and early user feedback loops. Refined UI, optimized workflows, and ensured reliable end-to-end execution.

8.8 Evaluation Strategy

The evaluation of the developed social listening platform focused on four key dimensions: utility, ease of use, efficiency, and cost-effectiveness. Given the practical context of the project and the limited number of end-users, the approach was primarily qualitative and exploratory rather than statistically rigorous. No formal usability metrics or quantitative thresholds were applied. Instead, the goal was to gather practical

feedback on whether the tool meets the needs of strategy consultants in real-world use cases.

8.8.1 Utility for Exploratory Research and Strategic Insight Generation

Feedback was gathered through semi-structured interviews and observational walkthroughs with consultants. Participants were asked how effectively the platform supported tasks such as:

- Understanding unfamiliar topics or communities
- Identifying discussion themes or trends
- Generating material for strategic hypotheses or point-of-view reports

8.8.2 Ease of Use and Cognitive Load

Usability was assessed informally as consultants performed core tasks, including:

- Acquiring data via SERP and Reddit scraping
- Interpreting topic clusters
- Accessing and exporting results

Observations were used to identify points of confusion or friction. Feedback was organized into themes such as user interface design, insight interpretability, customization options, and documentation needs.

8.8.3 Efficiency (Time Savings)

Efficiency was explored by comparing approximate durations for manual research workflows, based on consultant estimates, against the measured time taken by the automated platform. Time tracking focused on stages such as data collection, processing, and synthesis. Specific time comparisons are reported in Results and Evaluation Chapter.

8.8.4 Cost-Effectiveness

Cost evaluation was based on actual usage data from Google Cloud and Brightdata services. These costs were compared to estimated consultant labor for equivalent manual work and to licensing costs for commercial social listening tools. Detailed cost results are provided in Results and Evaluation Chapter.

9 IMPLEMENTATION

This chapter presents the technical realization of the system architecture introduced in the Methodology and System Design. It documents the integration of cloud services, user-facing interfaces, and machine learning components, with an emphasis on usability, modularity, and cost-efficiency. The implementation was designed to support rapid data acquisition, thematic analysis, and insight delivery, tailored to the workflow of strategy consultants with minimal coding experience.

9.1 USER INTERFACE AND VISUALIZATION

The user interface consists of two main components:

- Google Sheets: A familiar, low-code interface used as a “control panel” for orchestrating the data pipeline.
- Looker Studio: A dashboarding layer connected to BigQuery for visualizing processed results and enabling exploration.

Together, these components provide a flexible bridge between automated cloud processing and human-centered insight generation.

9.1.1 Google Sheets Control Panel

The Google Sheets interface guides the user through each stage of the data pipeline. Designed for accessibility and simplicity, it abstracts away technical complexity behind spreadsheet-driven triggers.

9.1.1.1 Search Links Sheet

Users begin by entering search queries and target platforms. This initiates a Cloud Function that calls the SERP API to retrieve relevant social media links. This structure minimizes cognitive load by mimicking familiar search behavior.

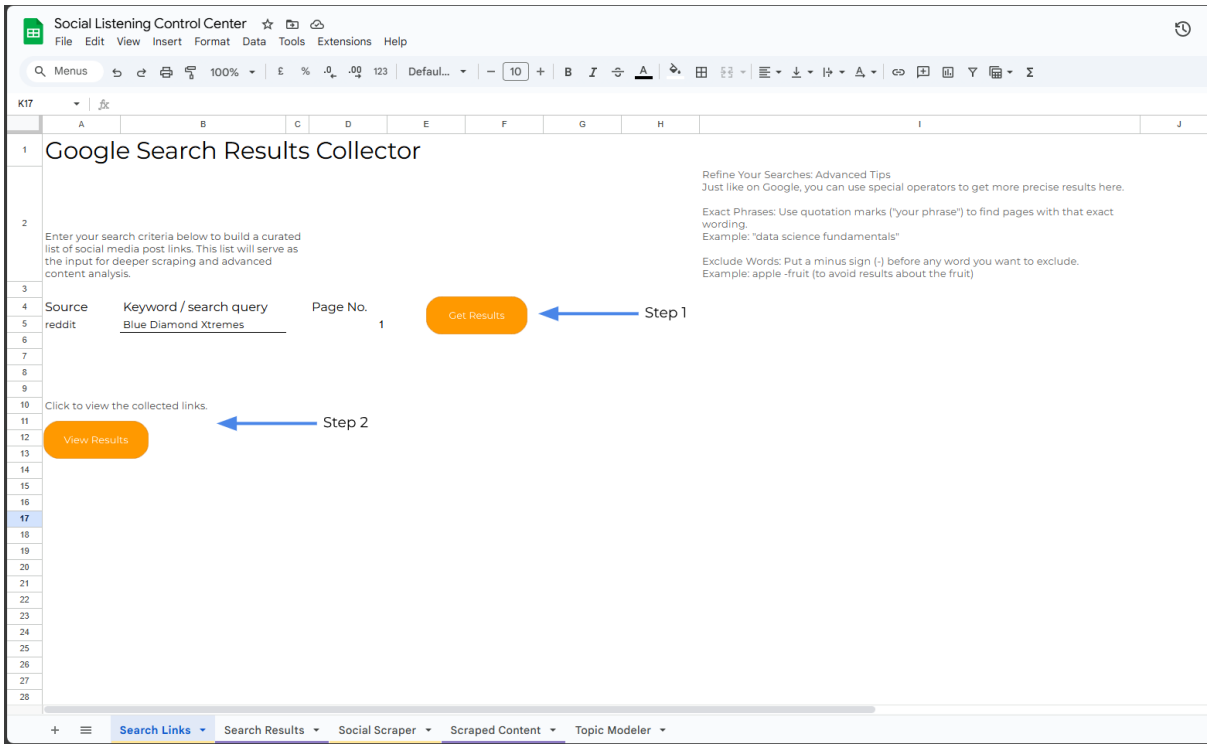


Figure 9.1 Search Links Sheet - User-defined queries initiate automated link retrieval.

9.1.1.2 Search Results Sheet

Search results, including URLs and metadata, are displayed for review. Consultants manually select and curate links for scraping, ensuring data relevance before downstream processing.

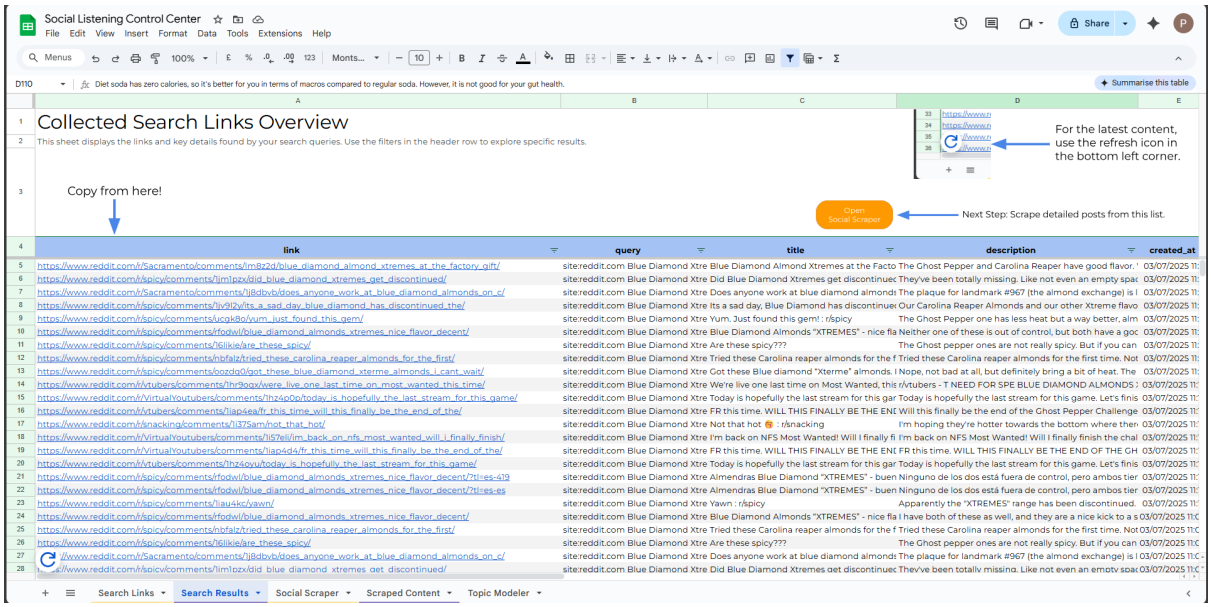


Figure 9.2 Search Results Sheet - Consultants verify and select links.

9.1.1.3 Social Scraper Sheet

Users paste curated URLs, triggering a scraping function via another cloud call. This approach allows flexibility users may mix SERP-acquired links with hand-selected ones (e.g., from Quora , Reddit).

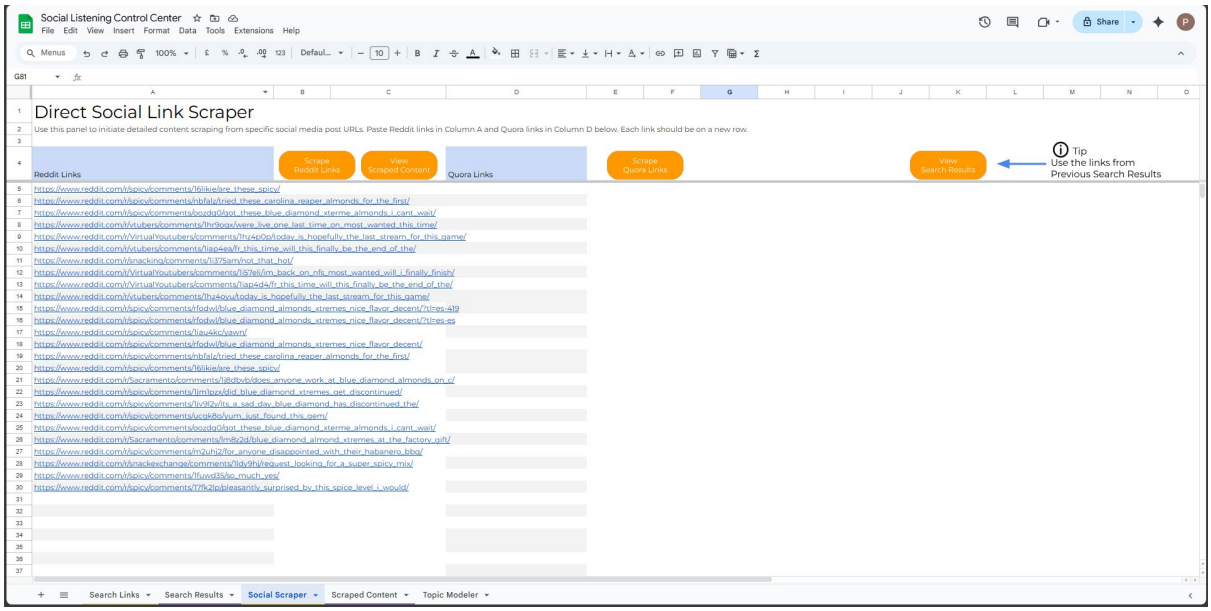


Figure 9.3 Social Scraper Sheet - Initiating content scraping for selected URLs.

9.1.1.4 Scraped Content Sheet

Scraped data is returned and displayed for inspection, offering visibility into the raw dataset prior to analysis. This checkpoint ensures transparency and early quality control.

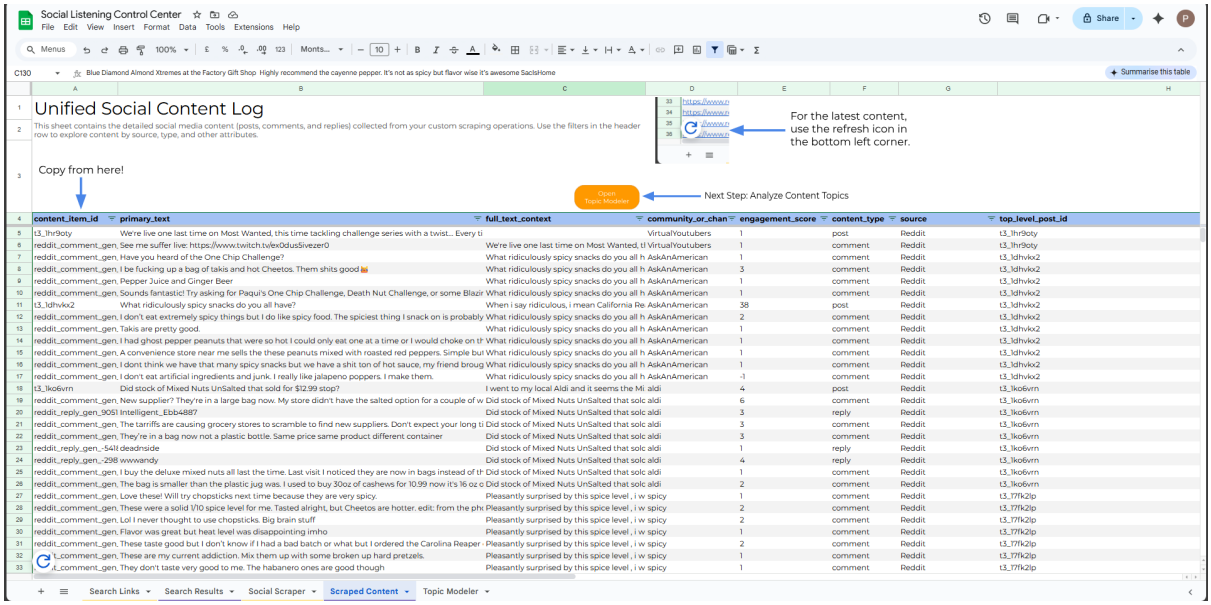


Figure 9.4 Scraped Content Sheet - Reviewing ingested social content.

9.1.1.5 Topic Modeler Sheet

This sheet allows users to launch the full analysis pipeline: embeddings generation, clustering, UMAP, and Gemini-based labeling. Upon completion, a Run ID links directly to the associated Looker Studio dashboard.

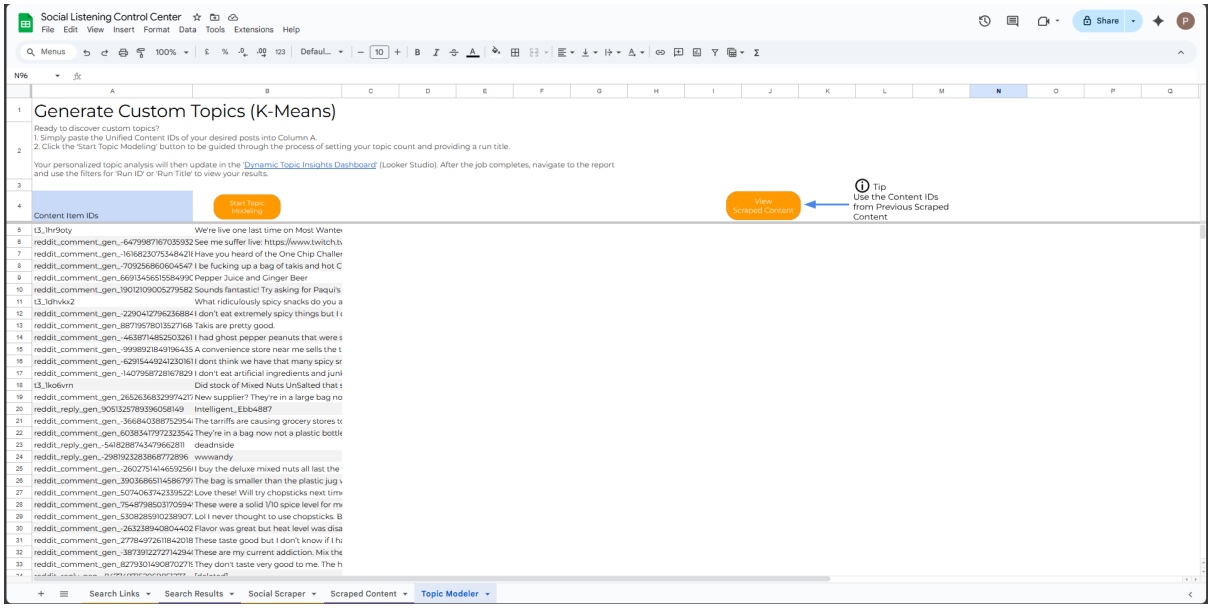


Figure 9.5 Topic Modeler Sheet - Triggering the end-to-end analysis process.

9.1.2 Looker Studio Dashboard

Processed results are visualized in a custom Looker Studio dashboard that connects to BigQuery tables. The dashboard was designed with consultants in mind prioritizing interpretability, interactivity, and modular exploration.

9.1.2.1 Run Summary and Overview

This entry page summarizes metadata for the selected analysis: total records, source breakdown, and content types. It orients users before they dive into deeper analysis.

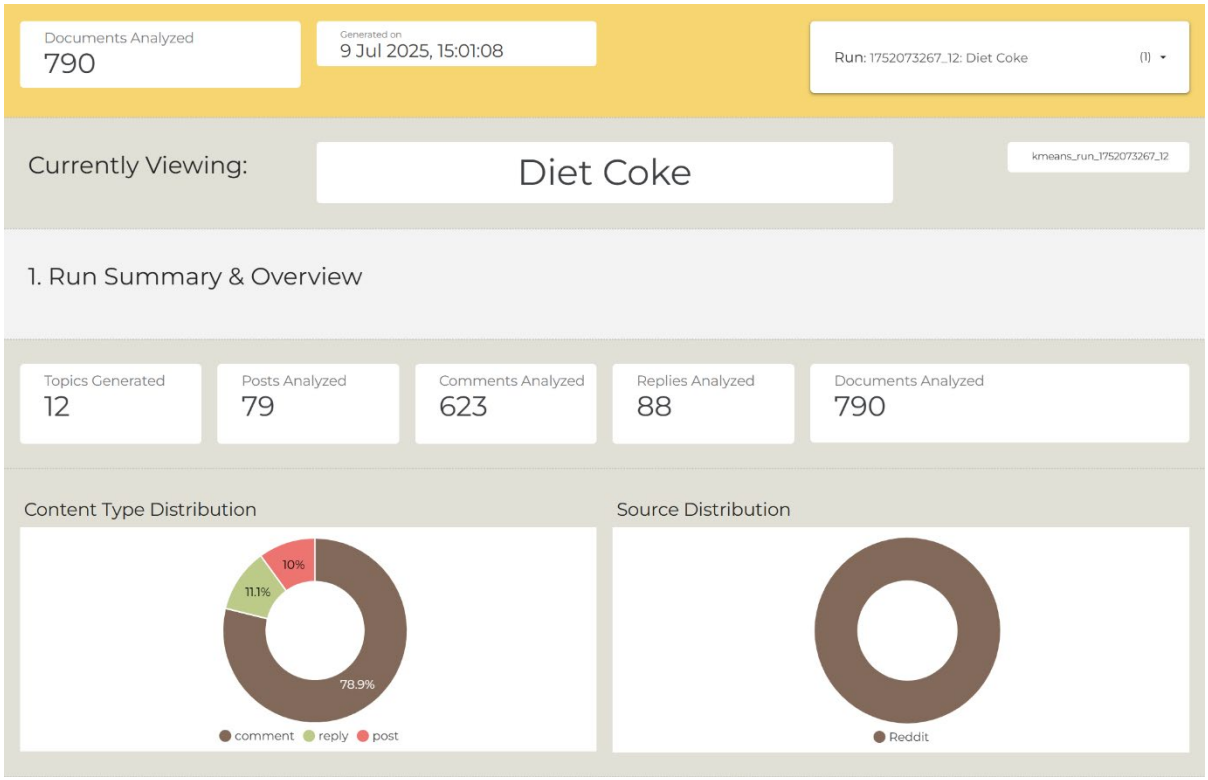


Figure 9.6 Run Summary - Overview of content scope and sources.

9.1.2.2 Discovered Topics Overview

This view presents clustered themes, including LLM-generated labels and descriptions for each topic. Consultants can interpret topics without reviewing every document.

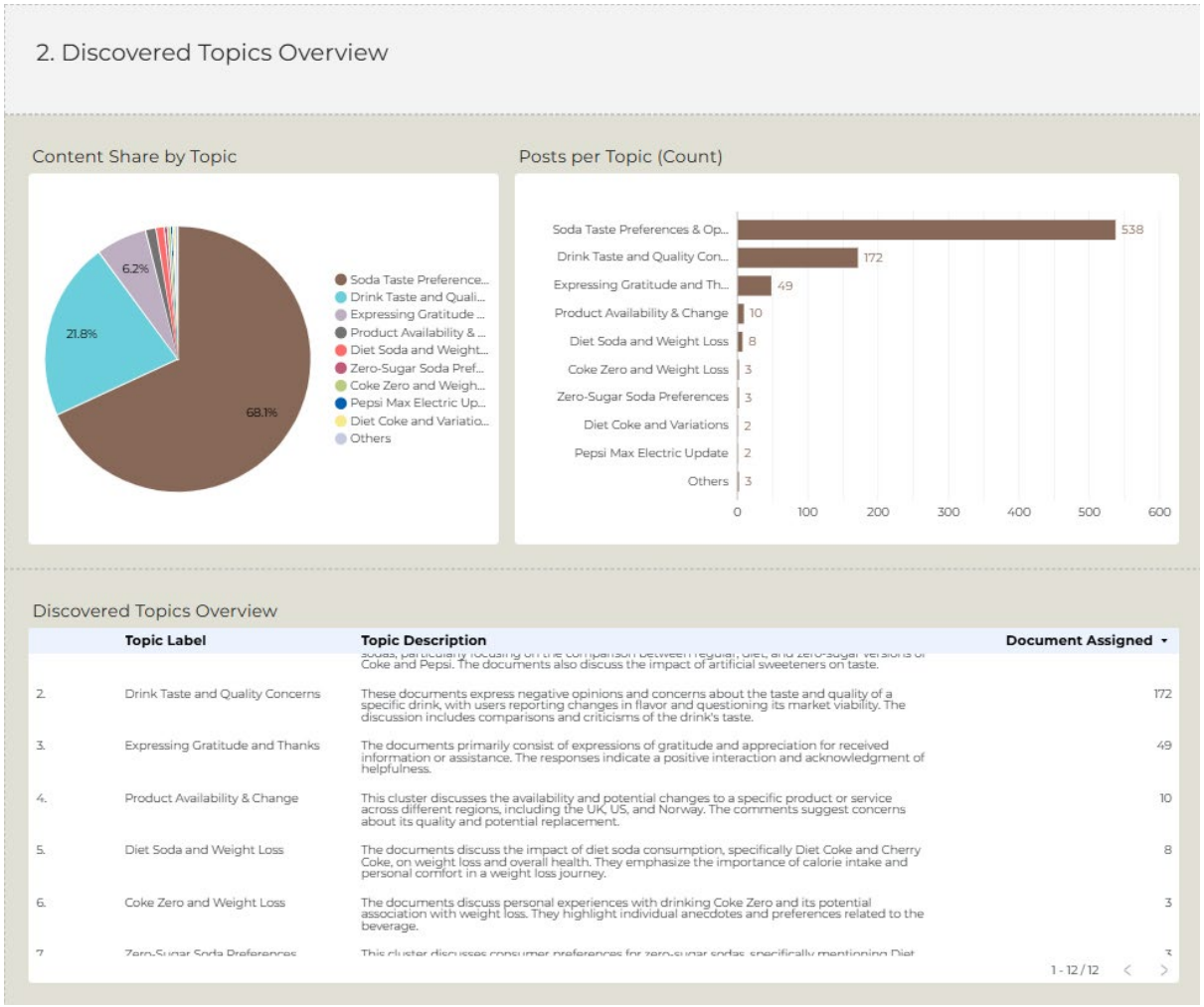


Figure 9.7 Topics Overview - LLM-enhanced cluster labeling supports rapid interpretation.

9.1.2.3 Topic Volume Trend Over Time

Topic mentions are plotted across time to identify spikes, emerging patterns, or declining themes. This temporal framing supports trendspotting and campaign timing.

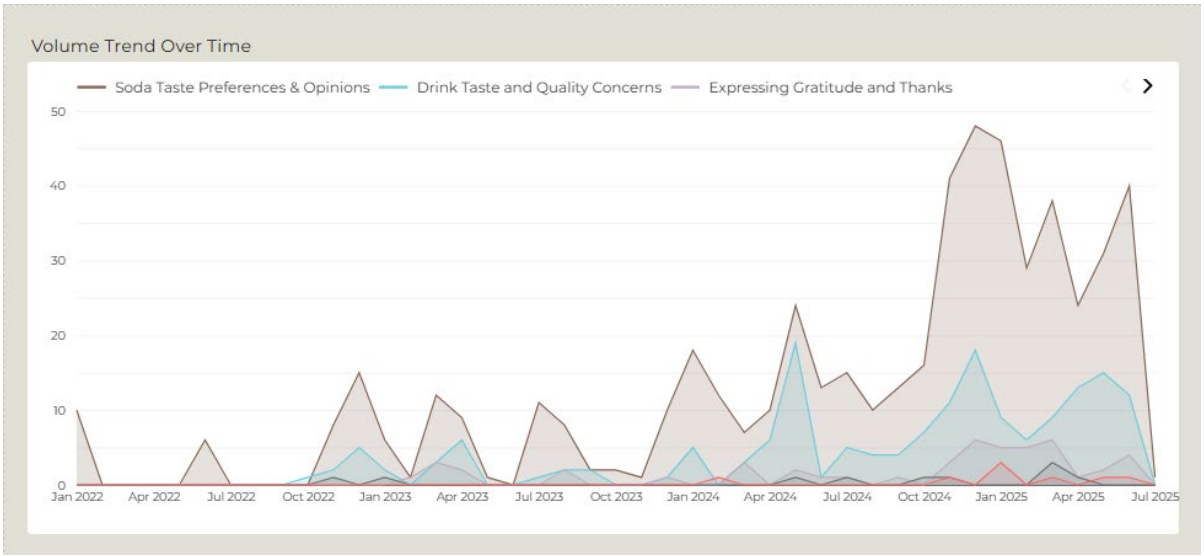


Figure 9.8 Topic Trend - Visualizing topic evolution and emergence.

9.1.2.4 Sentiment Analysis Overview

A dual-level sentiment breakdown is provided: overall and per-topic. This supports quick emotional profiling and segmentation of user attitudes.

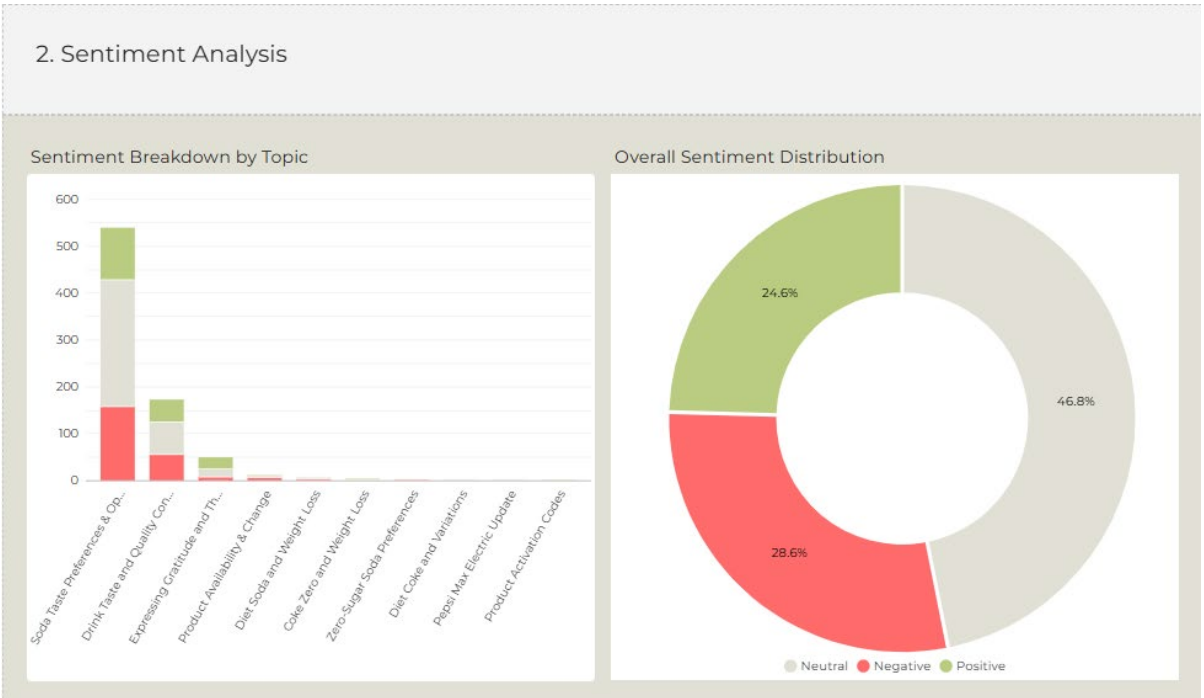


Figure 9.9 Sentiment Breakdown - High-level and topic-specific sentiment views.

9.1.2.5 Topic Map (UMAP Visualization)

UMAP plots embeddings into 2D space, with color-coded clusters. This spatial view reveals topic overlap, separation, and relative semantic distance.

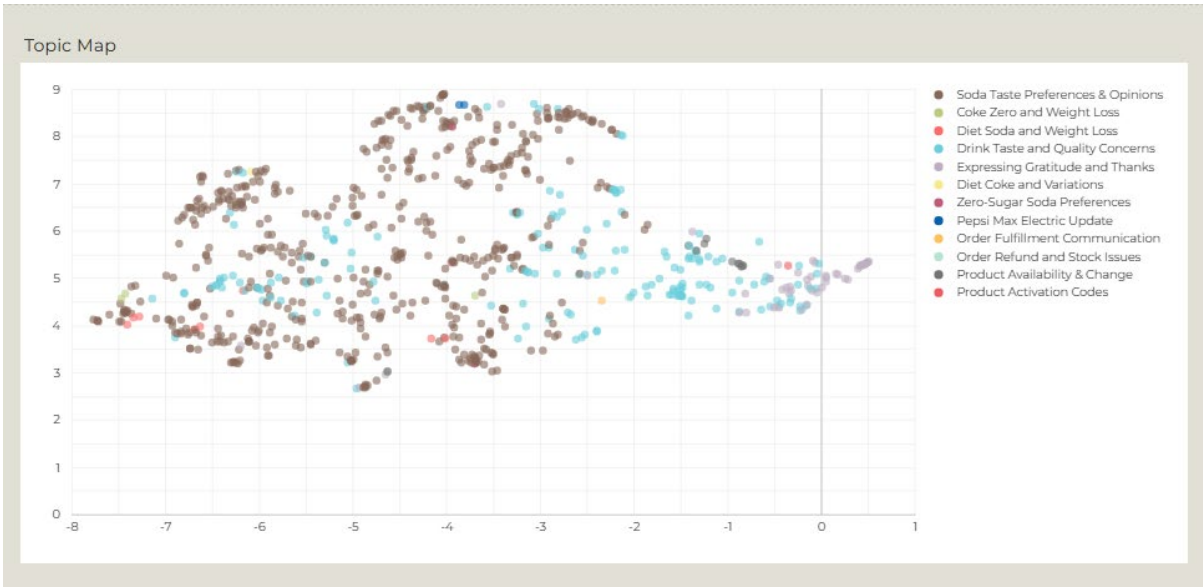


Figure 9.10 Topic Map - UMAP-based cluster visualization for semantic exploration.

9.1.2.6 Word Cloud

This view highlights frequently occurring terms across all records, reinforcing topical themes and aiding qualitative sense-making.

3. Word Cloud

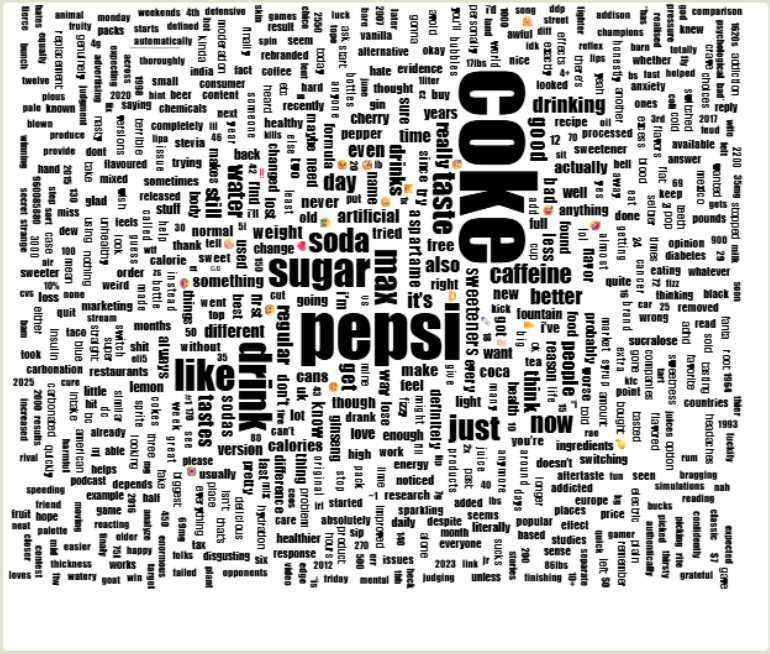
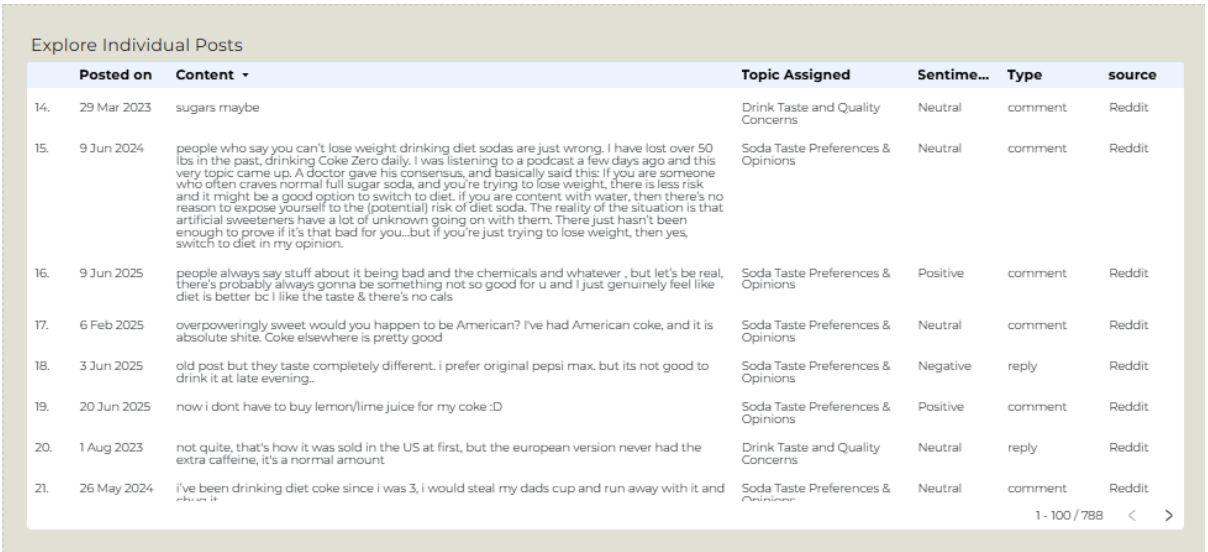


Figure 9.11 Word Cloud – A visual summary of dominant lexical patterns.

9.1.2.7 Individual Content Exploration

For qualitative drilling, this table presents original posts/comments alongside assigned topics, sentiment, and metadata. This enables hypothesis validation and contextual insight extraction.



Explore Individual Posts

	Posted on	Content	Topic Assigned	Sentime...	Type	source
14.	29 Mar 2023	sugars maybe	Drink Taste and Quality Concerns	Neutral	comment	Reddit
15.	9 Jun 2024	people who say you can't lose weight drinking diet sodas are just wrong. I have lost over 50 lbs in the past, drinking Coke Zero daily. I was listening to a podcast a few days ago and this very topic came up. A doctor gave his consensus, and basically said this: If you are someone who often craves normal full sugar soda, and you're trying to lose weight, there is less risk and it might be a good option to switch to diet. if you are content with water, then there's no reason to expose yourself to the (potential) risk of diet soda. The reality of the situation is that artificial sweeteners have a lot of unknown going on with them. There just hasn't been enough to prove if it's that bad for you...but if you're just trying to lose weight, then yes, switch to diet in my opinion.	Soda Taste Preferences & Opinions	Neutral	comment	Reddit
16.	9 Jun 2025	people always say stuff about it being bad and the chemicals and whatever , but let's be real, there's probably always gonna be something not so good for u and I just genuinely feel like diet is better bc I like the taste & there's no cals	Soda Taste Preferences & Opinions	Positive	comment	Reddit
17.	6 Feb 2025	overpoweringly sweet would you happen to be American? I've had American coke, and it is absolute shite. Coke elsewhere is pretty good	Soda Taste Preferences & Opinions	Neutral	comment	Reddit
18.	3 Jun 2025	old post but they taste completely different. i prefer original pepsi max. but its not good to drink it at late evening..	Soda Taste Preferences & Opinions	Negative	reply	Reddit
19.	20 Jun 2025	now i dont have to buy lemon/lime juice for my coke :D	Soda Taste Preferences & Opinions	Positive	comment	Reddit
20.	1 Aug 2023	not quite, that's how it was sold in the US at first, but the european version never had the extra caffeine, it's a normal amount	Drink Taste and Quality Concerns	Neutral	reply	Reddit
21.	26 May 2024	i've been drinking diet coke since i was 3, i would steal my dads cup and run away with it and	Soda Taste Preferences &	Neutral	comment	Reddit

1 - 100 / 788

Figure 9.12 Content Explorer – Reviewing original text with annotations.

The implementation reflects a design philosophy grounded in transparency, usability, and automation. By leveraging familiar tools (Google Sheets, Looker) and cloud-native ML services, the system enables consultants to derive deep insight from unstructured social data without needing to engage with code or infrastructure. Each implementation choice from trigger-based workflows to embedded LLMs supports the broader goal of empowering strategic analysts with accessible AI capabilities.

10 Results and Evaluation

The platform was evaluated for cost, efficiency, and user experience through testing with two consultants from Sense Worldwide. Each run analyzed roughly 1,000 social media records. While the findings are not statistically generalisable, they provide early evidence of the tool's value.

10.1 Cost and Efficiency Gains

The platform was designed to offer cost and time advantages over manual research and commercial software.

10.1.1 Cost Analysis

Generating an analytical report with approximately 1,000 social media items costs about $0.40. This includes expenses for scraping, processing, embeddings, sentiment analysis, and topic labeling.

Table 10.1 Estimated Cost Breakdown for Typical Analytical Report

Service Category	*Specific Operation*	*Estimated Cost (USD)*
Data Acquisition (Brightdata)	SERP API (Google Search Results)	$0.01
	Reddit Web Scraper	$0.15
GCP Data Processing & AI	BigQuery ML - Embeddings Generation	$0.02
	BigQuery ML - Sentiment Analysis	$0.20
	BigQuery ML - K-Means Clustering	$0.01
	BigQuery ML - LLM (Gemini) Calls	$0.00
	Google Cloud Function (UMAP)	$0.00
	Other GCP Services	$0.01
Total Cost per Report		$0.3975

Manual analysis was estimated at around three hours per task. Given UK consultancy rates of £50–£150/hour (Consultancy.uk, 2024), this would cost £150–£450. Tools like Brandwatch and Sprout Social charge $749/month and $299/user/month respectively for access to comparable features (Brandwatch, 2025; Sprout Social, 2025).

10.1.2 Time Efficiency

Processing time for a typical run of ~1,000 documents required just over eight minutes from data acquisition through to insight generation.

Table 10.2 Estimated Automated Processing Time per Report

Process/Service	*Specific Operation*	*Estimated Time (Minutes)*
Data Acquisition	SERP API Call (Link Acquisition)	0:06
	Reddit Web Scraper (Data Collection)	2:00
Data Processing & Analysis	Embeddings Generation (BigQuery ML)	2:00
	K-Means Clustering & UMAP Reduction	3:45
System Overhead	Other GCP Operations	0:15
Total Automated Time		8:06

When compared to the manual research baseline (~180 minutes), this represents a time reduction of over 95%. This efficiency directly addresses RQ3 and demonstrates the tool's potential to significantly accelerate insight delivery in consultancy workflows.

10.2 User Experience and Thematic Feedback

Semi-structured interviews and walkthroughs with a senior consultant provided targeted feedback on usability, utility, and areas for enhancement.

Table 10.3 Summary of User Testing Feedback (Consultant Perspective)

This table classifies and summarizes key feedback points gathered from user testing sessions with the senior consultant, providing insights into the tool's perceived utility, usability, and areas for future enhancement.

Category	*Specific Observation / Feedback Point*	*Implication for Project Evaluation / Future Development*
Overall Utility & Novelty	The consultant found sentiment analysis to be very useful for high-level understanding of data.	Confirms the value of integrated sentiment analysis for quick, actionable insights.
	The team had not used any similar dedicated social listening tools previously.	Highlights the tool's novelty and addresses a previously unmet need within the consultancy.
	The consultant provided the 3-hour estimate for manual data collection and initial analysis.	Serves as the crucial baseline for quantifying and validating the significant time savings demonstrated by the automated tool.
Data Quality & Pre-processing	Suggestion to remove single-word comments/replies from the dataset.	Indicates potential noise in raw text for certain analytical contexts; a future refinement for data cleansing or configurable pre-processing.
User Interface (UI) & Workflow	Suggestion to add quality-of-life (QoL) features in Google Sheets (e.g., a button to delete all previous links for easier data replacement).	Identifies opportunities to enhance user experience, streamline operational workflows, and improve efficiency through UI refinements.
Advanced Analytical Control	Desire for the ability to alter the prompt used to generate topic modeling labels (to tailor results for specific report objectives).	Highlights a key area for future development to provide advanced customization, allowing users to optimize insight quality and relevance based on their specific analytical goals for a particular report.
Data Export & Interoperability	Suggestion to add functionality to export collected social media data (e.g., comments) to PDF format, enabling external use or feeding to other LLM tools.	Identifies a valuable feature for external data utility and flexible interoperability with other analytical workflows or tools.
Data Collection Completeness	Observation that collecting all comments (including those under "load more" buttons) would be beneficial for more complete data.	Pinpoints a current data completeness limitation in the scraping methodology, suggesting a valuable enhancement for future data acquisition.
Usability & Onboarding	The tool would require a detailed instruction manual; it was not fully self-explanatory.	Reinforces the importance of comprehensive documentation and intuitive UI design for seamless user onboarding and proper tool utilization.
	Identified potential for jargon mismatch or misunderstanding of terminology (e.g., "3rd page" vs. "3 pages worth").	Highlights the need for clear communication, intuitive UI labels, and robust user guidance to align technical terms with consultant understanding and expectations.

10.3 Overall Discussion

The development and deployment of the social listening tool mark clear progress toward its intended strategic goals: reducing costs, accelerating insight generation, enhancing accessibility for non-technical consultants, and enabling exploratory research at Sense Worldwide. Yet alongside its utility, trade-offs in interpretability, scope, and customization also emerge, shaping how and where the tool should be applied.

10.3.1 Cost-Efficiency vs. Analytical Depth

At less than $0.40 per analytical run, the tool offers a substantial cost reduction compared to manual analysis (£150–£450) or commercial platforms like Brandwatch ($749/month). However, this economic efficiency may come at the expense of analytical depth. Unlike higher-end platforms that pull from multiple sources such as Twitter, TikTok, and Instagram, and offer advanced features such as dynamic sentiment calibration, this system currently supports only Reddit and Quora. Its K-Means clustering and sentiment analysis, while efficient, lack domain-specific tuning or cross-platform integration.

10.3.2 Speed vs. Interpretive Nuance

Reducing a three-hour task to under nine minutes provides over 95% time savings—especially valuable in fast-moving consulting work. But this speed can come with a downside. The tool gives quick, high-level insights using clustering and sentiment analysis, but it may miss important context, contradictions, or deeper meaning that a human analyst would notice through closer reading. Automation can make the analysis faster and cheaper, but also more shallow if used on its own. Without human review, the results might feel too generic or miss what really matters. The tool works well for spotting early patterns or generating ideas, but for more sensitive or complex questions, expert interpretation is still essential.

10.3.3 Accessibility and Cognitive Load

Non-technical consultants now navigate the full pipeline through familiar interfaces like Google Sheets and Looker Studio. Feedback confirms that this ease of use enhances engagement and autonomy. Nonetheless, configuring clustering parameters or

interpreting UMAP visualizations still requires a baseline understanding, meaning the platform lowers barriers but does not eliminate them entirely.

10.3.4 Strategic Use Cases

Feedback confirms that the tool fills a previously unmet need supporting hypothesis generation, trend spotting, and early-stage framing. However, it is not designed for longitudinal or ethnographic analysis. Its value lies in surfacing "what's emerging," not explaining "why it matters." In this sense, it should complement not replace deep qualitative methods.

10.4 LIMITATIONS

10.4.1 Narrow Consultant Sample

Validation of the platform's utility and novelty is based on feedback from a small group of internal consultants at Sense Worldwide. While their responses were detailed and positive, they remain context-specific and anecdotal. Broader validation particularly across diverse consultancy environments might produce different results in terms of usability, perceived value, or strategic relevance.

This limitation primarily stems from practical constraints, including limited time and restricted access to external participants. As such, current findings should be interpreted as promising but preliminary.

10.4.2 Partial Usability Evaluation

Usability was assessed through functional walkthroughs and informal interviews rather than systematic UX studies. More comprehensive testing could uncover interface pain points, strategic misalignments, or edge-case failures not observed in this exploratory phase.

10.4.3 Limited Platform Scope

The choice to focus exclusively on Reddit and Quora was intentional and primarily driven by time constraints, scope boundaries, and the specific analytical focus on text-rich data suitable for NLP methods. Implementing data collection from every major social platform (e.g., TikTok, X/Twitter, Instagram) would have significantly increased complexity, time demands, and costs, making it unfeasible within the dissertation's

timeframe. Although justified, this choice restricts the comprehensiveness and representativeness of generated insights.

10.4.4 Limited Clustering Methodology

The platform relies solely on K-Means for topic modeling, which restricts the analytical flexibility of the system. While K-Means was chosen for its computational efficiency, interpretability, and native support within BigQuery ML, it represents only one approach to clustering. Other methods such as HDBSCAN or BERTopic offer benefits like automatic cluster detection and better handling of non-spherical or overlapping data distributions.

These alternatives were not implemented due to their higher computational complexity and lack of seamless integration with the cloud-native architecture adopted for this project. However, the reliance on a single method limits adaptability to different data characteristics and may reduce robustness across varied consulting scenarios.

10.4.5 Incomplete Data from Selective Scraping

The platform captures only top-level Reddit comments, excluding nested replies that might contain deeper context, counterpoints, or elaboration. This decision was made deliberately to reduce scraping costs, processing time, and data storage demands keeping the system lightweight and aligned with its goal of rapid, cost-effective insight generation.

However, this approach limits data completeness. By omitting these, the analysis may miss important nuances or misrepresent the full shape of discourse, particularly in topics where reply chains contain rich debate or divergent views.

10.4.6 Low-Code Design Choices

The tool's low-code design favors accessibility over deep customizability, which may limit appeal to advanced analysts. More technical users may find the interface restrictive for custom model tuning or parameter experimentation. The system was explicitly designed for non-technical consultants, prioritizing ease of use over configurability. The assumption is that strategic insight arises from domain expertise, and even basic NLP tools can become powerful when paired with human interpretation.

10.5 Future Work

Further development of the platform can strengthen its flexibility, analytical depth, and user experience. Several areas for improvement have emerged across data coverage, interface functionality, and analytical capabilities.

10.5.1 Data Expansion and Completeness

Expanding the range of data sources beyond Reddit and Quora would improve representativeness and insight diversity. Including platforms that rely on short-form or multimedia content such as TikTok, Instagram, or X could uncover different behavioral signals and audience segments. This would require adapting the system to handle image and video data, possibly through integration with computer vision or audio transcription tools.

In addition, future iterations should aim to collect full Reddit thread structures, not just top-level comments. Many meaningful conversations occur deeper in the thread hierarchy, and capturing this nested content would provide a more complete view of public discourse.

10.5.2 Interface and Usability Enhancements

Improving the user interface could significantly boost adoption and satisfaction. Adding utility features such as bulk link deletion, confirmation prompts, or preset filters would streamline workflows. Greater customizability such as adjusting display options or refining clustering parameters through the interface would make the platform more adaptable to varied consultant needs.

Comprehensive usability testing with a broader user base could identify friction points, especially among new or infrequent users. Additional export formats, such as downloadable PDFs and CSVs, would also enable easier reporting and integration with other analysis tools.

10.5.3 Advanced Analysis Features

Enhancing the analytical pipeline would improve both the depth and reliability of insights. This includes adding automated pre-processing steps to detect spam, filter out low-value or excessively brief comments, and flag off-topic content.

Introducing alternative clustering methods such as HDBSCAN or BERTopic could increase robustness and accommodate a wider variety of data structures. Allowing users to modify the prompt templates used in LLM-based cluster labeling would offer more control over how insights are shaped and interpreted.

Finally, expanding the dashboard with additional visualizations such as heatmaps, comparative timelines, or interactive filters would provide consultants with richer tools for exploration and storytelling.

11 GenAI Usage

11.1 Writing Support: Multi-LLM Approach

Three distinct LLMs were employed for dissertation writing support, each demonstrating unique strengths that complemented different aspects of the writing process:

ChatGPT (Custom GPTs): The "Dissertation Buddy" custom GPT was configured to maintain consistency with preferred academic writing conventions and stylistic preferences. This approach proved most effective for maintaining voice consistency across chapters.

Gemini for Structural Planning: Gemini's extended context window (up to 1M tokens) (Comanici et al., 2025) enabled processing of entire chapter drafts for high-level structural feedback and outline development. However, outputs exhibited characteristics typical of AI-generated text verbose, jargon-heavy prose that required substantial human editing.

Claude for Logical Argumentation: Despite context window limitations requiring more targeted, specific inputs, Claude consistently produced the most logically structured outputs. This made it particularly valuable for developing coherent arguments and identifying potential counterarguments during the analytical phases of writing.

11.2 Code Generation Analysis

Gen AI specifically Gemini was instrumental in the code development process due to its extended context window. This allowed it to retain awareness of previously generated API structures, variable naming conventions, and system logic across multiple prompts.

As a result, creating interconnected systems (e.g., Google Cloud Functions for APIs, Google Apps Script for client-side integration, and BigQuery for data storage) became a smoother, more cohesive process.

Unlike shorter-context models, Gemini could "remember" how different components were supposed to interact. For example, when an API endpoint was defined, it could immediately generate the corresponding function call in Apps Script without requiring a full re-prompt. This continuity significantly reduced friction during iteration.

11.3 Topic Labels Generation and Evaluation

After performing K-Means clustering, Gemini used Google Vertex AI to generate human-readable labels and summaries for each cluster, replacing the need for manual annotation or keyword frequency-based methods.

To evaluate label quality, a report was generated using social media data concerning diet soda. The UMAP projection in Figure 9.10 illustrates the clustered embeddings and their Gemini-generated labels.

Table 11.1 Evaluation of Gemini-Generated Topic Labels Across Clusters.

Cluster Label	*Document Count*	*Label Quality*	*Observations*
Soda Taste Preferences & Opinions	538	Accurate but broad	Largest cluster; captured dominant themes like sweetener comparisons but label reflects inherent thematic diversity
Drink Taste and Quality Concerns	172	Accurate summary, masked variation	Summarized sentiment well but concealed meaningful differences between product feedback, brand perception, and health complaints
Expressing Gratitude and Thanks	49	Optimal performance	Internally consistent content yielding accurate, specific label
Diet Soda and Weight Loss	6	Highly accurate	Small cluster with precise labeling but limited strategic value
Order Fulfillment Communication	3	Highly accurate	Micro-cluster representing outliers with limited strategic value

These findings demonstrate that Gemini performs well when clusters are internally consistent, regardless of size. However, label clarity degrades when clusters are too broad or fragmented indicating that the effectiveness of automated topic labeling is

dependent on the quality of upstream clustering, not on hallucination or misrepresentation by the model.

11.4 Ethical Considerations and Research Integrity

The research maintained clear boundaries between AI assistance and original contribution. Writing support focused on clarity and structure enhancement without contributing substantive content or arguments. Code generation provided scaffolding for predetermined approaches rather than architectural decisions. Academic integrity was preserved through transparent acknowledgment of AI usage, human validation of all outputs, and retention of final responsibility for all research deliverables.

12 References

Blei, D., Ng, A. and Jordan, M. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, [online] 3(993-1022), pp.993–1022. Available at: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf.

Brandwatch (2025) Pricing plans. Available at: https://www.brandwatch.com/pricing [Accessed: 21 July 2025].

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C. and Hesse, C. (2020). Language Models Are Few-Shot Learners. arxiv.org, [online] 4(33). Available at: https://arxiv.org/abs/2005.14165

Chaffey, D. and Ellis-Chadwick, F. (2019). Digital Marketing: Strategy, Implementation and Practice. 8th ed. Harlow, England: Pearson.

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., Marris, L., Petulla, S., Gaffney, C., Aharoni, A., Lintz, N., Pais, T.C., Jacobsson, H., Szpektor, I., Jiang, N.-J. and Haridasan, K. (2025). Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. [online] arXiv.org. Available at: https://arxiv.org/abs/2507.06261 [Accessed 28 Jul. 2025].

Consultancy.uk (2024) Consulting industry fees & rates. Available at: https://www.consultancy.uk/consulting-industry/fees-rates [Accessed: 21 July 2025].

Culnan, M.J., Mchugh, P. and Zubillaga, J.I. (2010). (PDF) How Large U.S. Companies Can Use Twitter and Other Social Media to Gain Business Value. [online] ResearchGate. Available at: https://www.researchgate.net/publication/279893388_How_Large_US_Companies_Can_Use_Twitter_and_Other_Social_Media_to_Gain_Business_Value.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. [online] ArXiv. Available at: https://arxiv.org/abs/1810.04805.

Donkin, R. (2010) The Future of Work. [Online]. Palgrave Macmillan UK. Available from: doi:10.1057/9780230274198.

Fan, W. and Gordon, M.D. (2014). The power of social media analytics. Communications of the ACM, [online] 57(6), pp.74–81. doi:https://doi.org/10.1145/2602574.

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv:2203.05794 [cs]. [online] doi:https://doi.org/10.48550/arXiv.2203.05794.

Hasan, R. (2024). Sentiment analysis and social media analytics in brand management: Techniques, trends, and implications. World Journal of Advanced Research and Reviews, 23(2), pp.287–296. doi:https://doi.org/10.30574/wjarr.2024.23.2.2369.

Hevner, A., March, S., Park, J. and Ram, S. (2004). Design Science in Information Systems Research. MIS Quarterly, [online] 28(1), pp.75–105. doi:https://doi.org/10.2307/25148625.

Kamburugamuve, S., Ekanayake, S., Pathirage, M. and Fox, G. (2016). Towards High Performance Processing of Streaming Data in Large Data Centers. 2016 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW). doi:https://doi.org/10.1109/ipdpsw.2016.103.

Kaplan, A.M. and Haenlein, M. (2010). Users of the world, unite! the Challenges and Opportunities of Social Media. Business Horizons, [online] 53(1), pp.59–68. doi:https://doi.org/10.1016/j.bushor.2009.09.003.

Lakshmanan, V. (2022). DATA SCIENCE ON THE GOOGLE CLOUD PLATFORM : implementing end-to-end real-time data... pipelines. S.L.: O'reilly Media.

Laney, D., 2001. 3D data management: Controlling data volume, velocity and variety. META group research note, 6(70), p.1.

Liu, B. (2022). Sentiment Analysis and Opinion Mining. Springer Nature.

Malhotra, N.K. (2019). Marketing Research : An Applied Orientation. 7th ed. New York, Ny: Pearson.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. [online] arXiv.org. Available at: https://arxiv.org/abs/1301.3781.

Nesta (2011) Sense Worldwide « OpenBusiness. [Online]. 2011. Available from: https://web.archive.org/web/20110830101558/http://www.openbusiness.cc/2010/01/06/sense-worldwide/ [Accessed: 31 July 2021].

Peffers, K. et al. (2007) 'A Design Science Research Methodology for Information Systems Research', Journal of Management Information Systems, 24(3), pp. 45–77. doi: 10.2753/MIS0742-1222240302.

radimrehurek.com. (n.d.). Gensim: topic modelling for humans. [online] Available at: https://radimrehurek.com/gensim/index.html.

Sense Worldwide. (2024). The Sense Network - Sense Worldwide. [online] Available at: https://senseworldwide.com/innovate-with-the-sense-network/ [Accessed 8 May 2025]

Sprout Social (2025) Pricing plans. Available at: https://sproutsocial.com/pricing [Accessed: 21 July 2025].

13 APPENDIX

13.1 UNIFIED_SOCIAL_CONTENT_ITEMS VIEW

This appendix provides the complete SQL definition for the *unified_social_content_items* BigQuery View. This view is a crucial component of the data model, responsible for flattening hierarchical social media content and standardizing schemas from various sources into a single, unified dataset for downstream analytical processing.

```
CREATE OR REPLACE VIEW `social-listening-sense.social_listening_data.unified_social_content_items` AS

WITH
  -- CTE for Reddit Posts: Rely on t.post_id for stable_post_id, filter out rows where post_id is NULL.
  reddit_posts_base AS (
    SELECT
        t.post_id,
        t.url,
        t.user_posted,
        t.title,
        t.description,
        t.num_upvotes,
        t.date_posted,
        t.community_name,
        t.photos,
        t.videos,
        t.timestamp,
        t.comments,
        t.snapshot_id,
        t.error_code,
        t.error,
        t.warning_code,
        t.warning,
        t.post_id AS stable_post_id
    FROM
      `social-listening-sense.social_listening_data.reddit_data` AS t
    WHERE
      t.post_id IS NOT NULL
      AND (t.title IS NOT NULL OR t.description IS NOT NULL)
  ),

  -- CTE for Reddit Comments: Generate stable comment_item_id without UUID.
  reddit_comments_base AS (
    SELECT
        p.stable_post_id,
        c.url AS comment_url,
        c.user_commenting,
        c.comment,
        c.num_upvotes,
        c.date_of_comment,
        c.replies,
        -- Generate a stable ID for the comment using deterministic fields
        CONCAT('reddit_comment_gen_', FARM_FINGERPRINT(
            CONCAT(
                COALESCE(p.stable_post_id, ''),
                COALESCE(c.user_commenting, ''),
                FORMAT_TIMESTAMP('%Y%m%d%H%M%S%F', COALESCE(c.date_of_comment, CAST('1970-01-01 00:00:00 UTC' AS TIMESTAMP))),
                COALESCE(c.comment, '')
            )
        )) AS stable_comment_id
    FROM
      reddit_posts_base AS p,
      UNNEST(p.comments) AS c
    WHERE
      c.comment IS NOT NULL
  ),

  -- CTE for Reddit Replies: Generate stable reply_item_id without UUID.
  reddit_replies_base AS (
    SELECT
        c.stable_post_id,
        c.stable_comment_id,
        r.user_url,
        r.user_replying,
        r.num_upvotes,
        r.date_of_reply,
        -- Using r.user_replying as content
        r.user_replying AS reply_content_placeholder,
        -- Generate a stable ID for the reply using deterministic fields
        CONCAT('reddit_reply_gen_', FARM_FINGERPRINT(
            CONCAT(
                COALESCE(c.stable_comment_id, ''),
                COALESCE(r.user_replying, ''),
                FORMAT_TIMESTAMP('%Y%m%d%H%M%S%F', COALESCE(r.date_of_reply, CAST('1970-01-01 00:00:00 UTC' AS TIMESTAMP))),
                COALESCE(r.user_replying, '') --  Used for hash
            )
```

```
        )) AS stable_reply_id
    FROM
      reddit_comments_base AS c,
      UNNEST(c.replies) AS r
    WHERE
      r.user_replying IS NOT NULL -- Filtering on user_replying,
  ),

  -- CTE for Quora Questions: Rely on t.post_id for stable_post_id, filter out rows where post_id is NULL.
  quora_questions_base AS (
    SELECT
        t.post_id,
        t.url,
        t.author_name,
        t.title,
        t.post_text,
        t.upvotes,
        t.post_date,
        t.pictures_urls,
        t.videos_urls,
        t.top_comments,
        t.timestamp,
        t.snapshot_id,
        t.post_id AS stable_post_id
    FROM
      `social-listening-sense.social_listening_data.quora_data` AS t
    WHERE
      t.post_id IS NOT NULL
      AND (t.title IS NOT NULL OR t.post_text IS NOT NULL)
  ),

  -- CTE for Quora Comments: Generate stable comment_item_id.
  quora_comments_base AS (
    SELECT
        q.stable_post_id,
        q.url AS question_url,
        c.commenter_name,
        c.comment,
        c.comment_date,
        c.replys,
        -- Generate a stable ID for the comment
        CONCAT('quora_comment_gen_', FARM_FINGERPRINT(
            CONCAT(
                COALESCE(q.stable_post_id, ''),
                COALESCE(c.commenter_name, ''),
                FORMAT_TIMESTAMP('%Y%m%d%H%M%S%F', COALESCE(c.comment_date, CAST('1970-01-01 00:00:00 UTC' AS TIMESTAMP))),
                COALESCE(c.comment, '')
            )
        )) AS stable_comment_id
    FROM
      quora_questions_base AS q,
      UNNEST(q.top_comments) AS c
    WHERE
      c.comment IS NOT NULL
  ),

  -- CTE for Quora Replies: Generate stable reply_item_id.
  quora_replies_base AS (
    SELECT
        c.stable_post_id,
        c.stable_comment_id,
        r.commenter_name,
        r.comment,
        r.comment_date,
        -- Generate a stable ID for the reply
        CONCAT('quora_reply_gen_', FARM_FINGERPRINT(
            CONCAT(
                COALESCE(c.stable_comment_id, ''),
                COALESCE(r.commenter_name, ''),
                FORMAT_TIMESTAMP('%Y%m%d%H%M%S%F', COALESCE(r.comment_date, CAST('1970-01-01 00:00:00 UTC' AS TIMESTAMP))),
                COALESCE(r.comment, '')
            )
        )) AS stable_reply_id
    FROM
      quora_comments_base AS c,
      UNNEST(c.replys) AS r
    WHERE
      r.comment IS NOT NULL
  )

--- Final UNION ALL for the Unified View ---
SELECT * FROM (
  SELECT
      'Reddit' AS source,
      'post' AS content_type,
      t.stable_post_id AS content_item_id,
      CAST(NULL AS STRING) AS parent_content_item_id,
      t.stable_post_id AS top_level_post_id,
      t.url AS content_item_url,
      t.user_posted AS author_username,
      t.title AS primary_text,
      t.description AS full_text_context,
      CAST(t.num_upvotes AS INT64) AS engagement_score,
      t.date_posted AS content_timestamp,
      t.community_name AS community_or_channel_name,
      ARRAY<STRING>[] AS hashtags,
      ARRAY<STRING>[] AS mentions,
      t.photos AS media_urls,
      t.videos AS video_urls,
```

```
      t.timestamp AS record_load_timestamp,
      t.snapshot_id,
      ROW_NUMBER() OVER(PARTITION BY t.stable_post_id ORDER BY t.date_posted DESC) as rn
  FROM
      reddit_posts_base AS t
)
WHERE rn = 1

UNION ALL

SELECT * FROM (
  SELECT
      'Reddit' AS source,
      'comment' AS content_type,
      c.stable_comment_id AS content_item_id,
      c.stable_post_id AS parent_content_item_id,
      c.stable_post_id AS top_level_post_id,
      c.comment_url AS content_item_url,
      c.user_commenting AS author_username,
      c.comment AS primary_text,
      CONCAT(COALESCE(rp.title, ''), ' ', COALESCE(rp.description, ''), ' ', COALESCE(c.comment, '')) AS full_text_context,
      CAST(c.num_upvotes AS INT64) AS engagement_score,
      c.date_of_comment AS content_timestamp,
      rp.community_name AS community_or_channel_name,
      ARRAY<STRING>[] AS hashtags,
      ARRAY<STRING>[] AS mentions,
      ARRAY<STRING>[] AS media_urls,
      ARRAY<STRING>[] AS video_urls,
      rp.timestamp AS record_load_timestamp,
      rp.snapshot_id,
      ROW_NUMBER() OVER(PARTITION BY c.stable_comment_id ORDER BY c.date_of_comment DESC, rp.date_posted DESC) as rn
  FROM
      reddit_comments_base AS c
  JOIN
      reddit_posts_base AS rp
  ON
      c.stable_post_id = rp.stable_post_id
)
WHERE rn = 1

UNION ALL

SELECT * FROM (
  SELECT
      'Reddit' AS source,
      'reply' AS content_type,
      r.stable_reply_id AS content_item_id,
      r.stable_comment_id AS parent_content_item_id,
      r.stable_post_id AS top_level_post_id,
      r.user_url AS content_item_url,
      r.user_replying AS author_username,
      r.reply_content_placeholder AS primary_text,
           CONCAT(COALESCE(rp.title,   ''),   '   ',   COALESCE(rp.description,   ''),   '   ',   COALESCE(rc.comment,   ''),   '   ',
COALESCE(r.reply_content_placeholder, '')) AS full_text_context,
      CAST(r.num_upvotes AS INT64) AS engagement_score,
      r.date_of_reply AS content_timestamp,
      rp.community_name AS community_or_channel_name,
      ARRAY<STRING>[] AS hashtags,
      ARRAY<STRING>[] AS mentions,
      ARRAY<STRING>[] AS media_urls,
      ARRAY<STRING>[] AS video_urls,
      rp.timestamp AS record_load_timestamp,
      rp.snapshot_id,
      ROW_NUMBER() OVER(PARTITION BY r.stable_reply_id ORDER BY r.date_of_reply DESC, rc.date_of_comment DESC, rp.date_posted DESC) as rn
  FROM
      reddit_replies_base AS r
  JOIN
      reddit_comments_base AS rc
  ON
      r.stable_comment_id = rc.stable_comment_id
  JOIN
      reddit_posts_base AS rp
  ON
      r.stable_post_id = rp.stable_post_id
)
WHERE rn = 1

UNION ALL

SELECT * FROM (
  SELECT
      'Quora' AS source,
      'post' AS content_type,
      t.stable_post_id AS content_item_id,
      CAST(NULL AS STRING) AS parent_content_item_id,
      t.stable_post_id AS top_level_post_id,
      t.url AS content_item_url,
      t.author_name AS author_username,
      t.title AS primary_text,
      t.post_text AS full_text_context,
      CAST(t.upvotes AS INT64) AS engagement_score,
      t.post_date AS content_timestamp,
      CAST(NULL AS STRING) AS community_or_channel_name,
      ARRAY<STRING>[] AS hashtags,
      ARRAY<STRING>[] AS mentions,
      t.pictures_urls AS media_urls,
      CASE WHEN t.videos_urls IS NOT NULL THEN [t.videos_urls] ELSE ARRAY<STRING>[] END AS video_urls,
      t.timestamp AS record_load_timestamp,
      t.snapshot_id,
```

```
      ROW_NUMBER() OVER(PARTITION BY t.stable_post_id ORDER BY t.post_date DESC) as rn
  FROM
      quora_questions_base AS t
)
WHERE rn = 1

UNION ALL

SELECT * FROM (
  SELECT
      'Quora' AS source,
      'comment' AS content_type,
      c.stable_comment_id AS content_item_id,
      c.stable_post_id AS parent_content_item_id,
      c.stable_post_id AS top_level_post_id,
      q.url AS content_item_url,
      c.commenter_name AS author_username,
      c.comment AS primary_text,
      CONCAT(COALESCE(q.title, ''), ' ', COALESCE(q.post_text, ''), ' ', COALESCE(c.comment, '')) AS full_text_context,
      CAST(NULL AS INT64) AS engagement_score,
      c.comment_date AS content_timestamp,
      CAST(NULL AS STRING) AS community_or_channel_name,
      ARRAY<STRING>[] AS hashtags,
      ARRAY<STRING>[] AS mentions,
      ARRAY<STRING>[] AS media_urls,
      ARRAY<STRING>[] AS video_urls,
      q.timestamp AS record_load_timestamp,
      q.snapshot_id,
      ROW_NUMBER() OVER(PARTITION BY c.stable_comment_id ORDER BY c.comment_date DESC, q.post_date DESC) as rn
  FROM
      quora_comments_base AS c
  JOIN
      quora_questions_base AS q
  ON
      c.stable_post_id = q.stable_post_id
)
WHERE rn = 1

UNION ALL

SELECT * FROM (
  SELECT
      'Quora' AS source,
      'reply' AS content_type,
      r.stable_reply_id AS content_item_id,
      r.stable_comment_id AS parent_content_item_id,
      r.stable_post_id AS top_level_post_id,
      q.url AS content_item_url,
      r.commenter_name AS author_username,
      r.comment AS primary_text,
       CONCAT(COALESCE(q.title, ''), ' ', COALESCE(q.post_text, ''), ' ', COALESCE(qc.comment, ''), ' ', COALESCE(r.comment, '')) AS
full_text_context,
      CAST(NULL AS INT64) AS engagement_score,
      r.comment_date AS content_timestamp,
      CAST(NULL AS STRING) AS community_or_channel_name,
      ARRAY<STRING>[] AS hashtags,
      ARRAY<STRING>[] AS mentions,
      ARRAY<STRING>[] AS media_urls,
      ARRAY<STRING>[] AS video_urls,
      q.timestamp AS record_load_timestamp,
      q.snapshot_id,
      ROW_NUMBER() OVER(PARTITION BY r.stable_reply_id ORDER BY r.comment_date DESC, qc.comment_date DESC, q.post_date DESC) as rn
  FROM
      quora_replies_base AS r
  JOIN
      quora_comments_base AS qc
  ON
      r.stable_comment_id = qc.stable_comment_id
  JOIN
      quora_questions_base AS q
  ON
      r.stable_post_id = q.stable_post_id
)
WHERE rn = 1
```

13.2 Data Infrastructure and Workflow Snapshots

This section includes screenshots that show how the system was built and how it works behind the scenes. The images capture how tools like Google BigQuery and Brightdata were used to collect, store, and process social media data. They show key steps such as setting up search queries, gathering content, running analysis jobs, and organizing topics for review. These visuals help demonstrate the structure and logic of the system and show the practical work that went into building it.

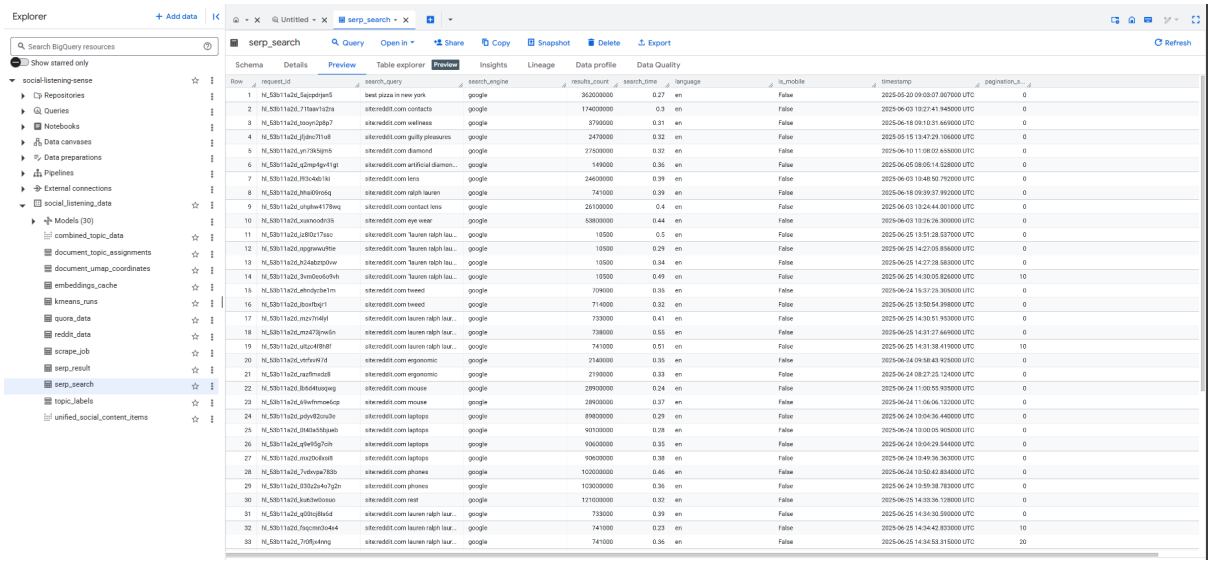


Figure 13.1 BigQuery serp_search Table – This table logs all search engine queries made through the Brightdata SERP API, including keywords, timestamps, estimated result counts, and search metadata.

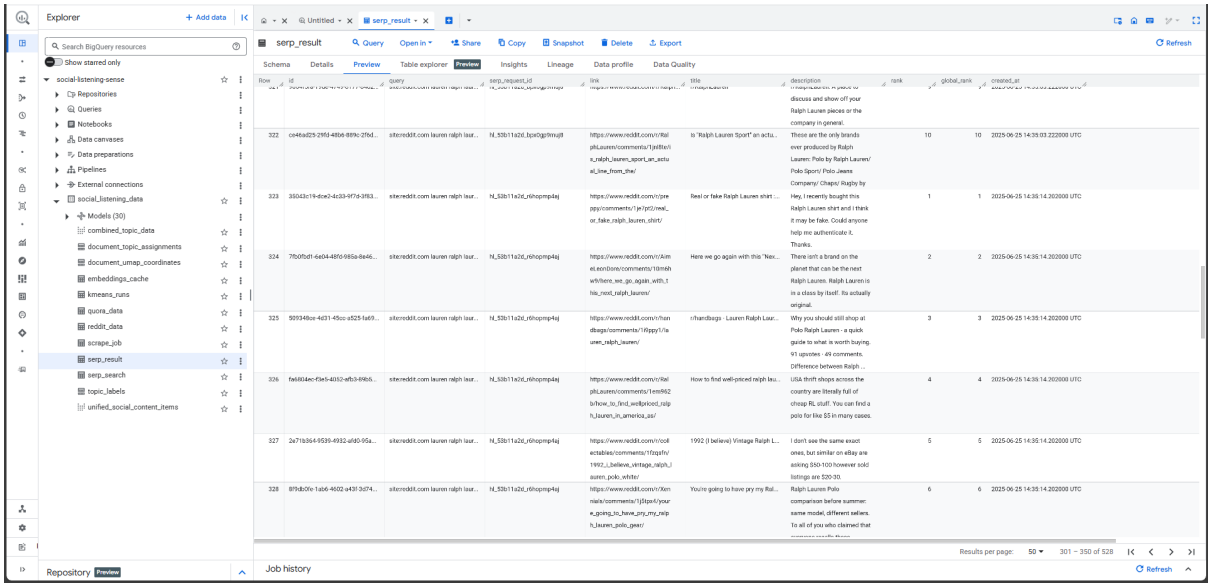


Figure 13.2 BigQuery serp_result Table – Stores individual search results (URLs, titles, metadata) retrieved per SERP query, enabling link tracking and selection for scraping.

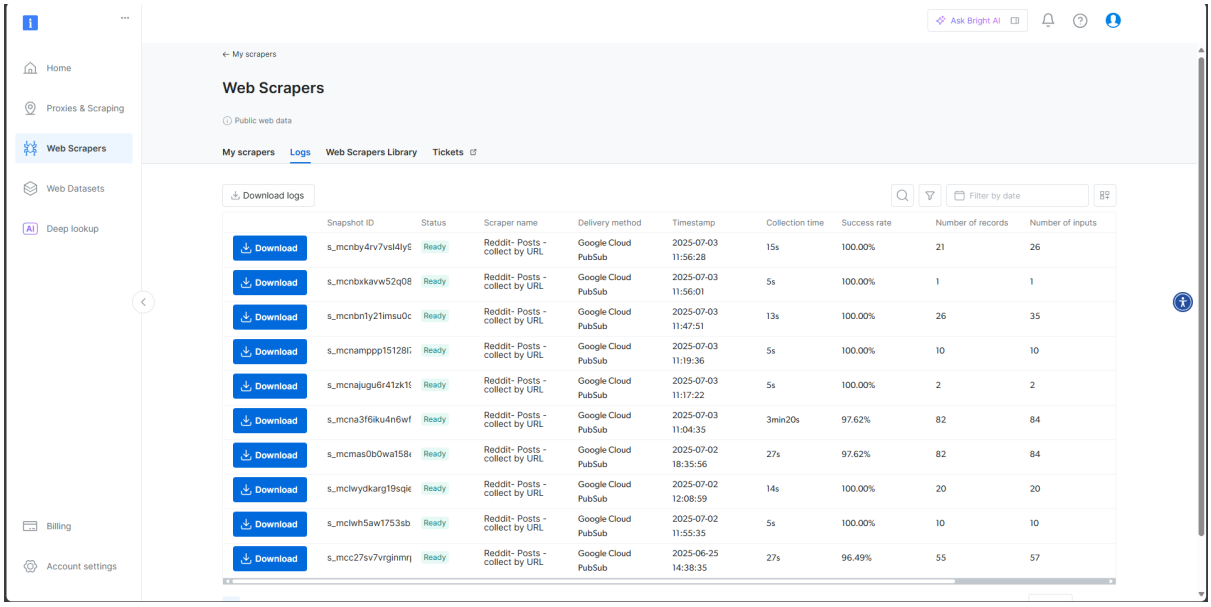


Figure 13.3 Brightdata Web Scraper Dashboard – Overview of Reddit and Quora scraping tasks, showing the number of URLs processed, success rate, and data delivery method.

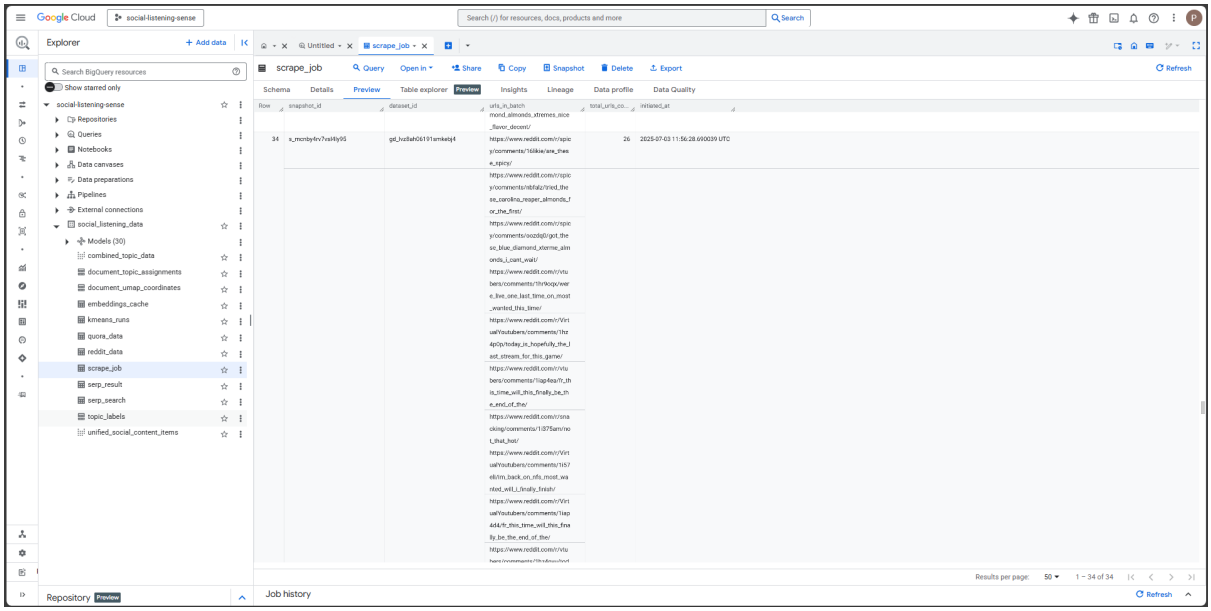


Figure 13.4 BigQuery scrape_job Table – Maintains records of scraping jobs triggered from selected links, including dataset and snapshot identifiers for traceability.

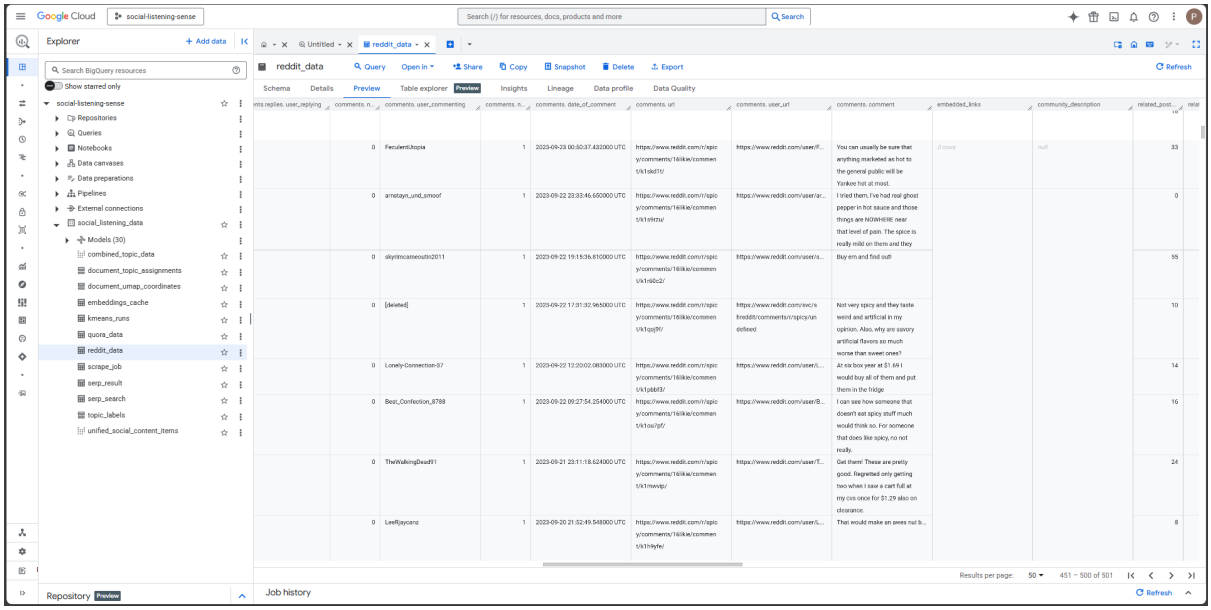


Figure 13.5 BigQuery reddit_data Table – Captures raw scraped Reddit content including post bodies, comments, engagement metrics, and author metadata.

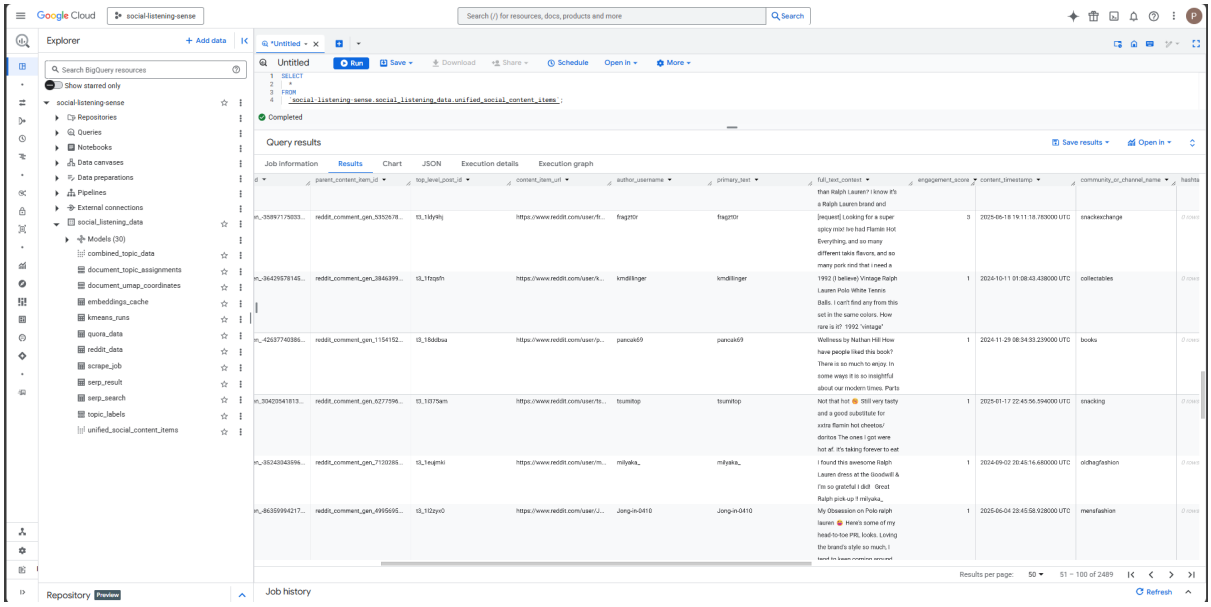


Figure 13.6 Query Result View – Unified content extracted from Reddit and Quora, including text, timestamps, and community-level metadata to support downstream NLP analysis.

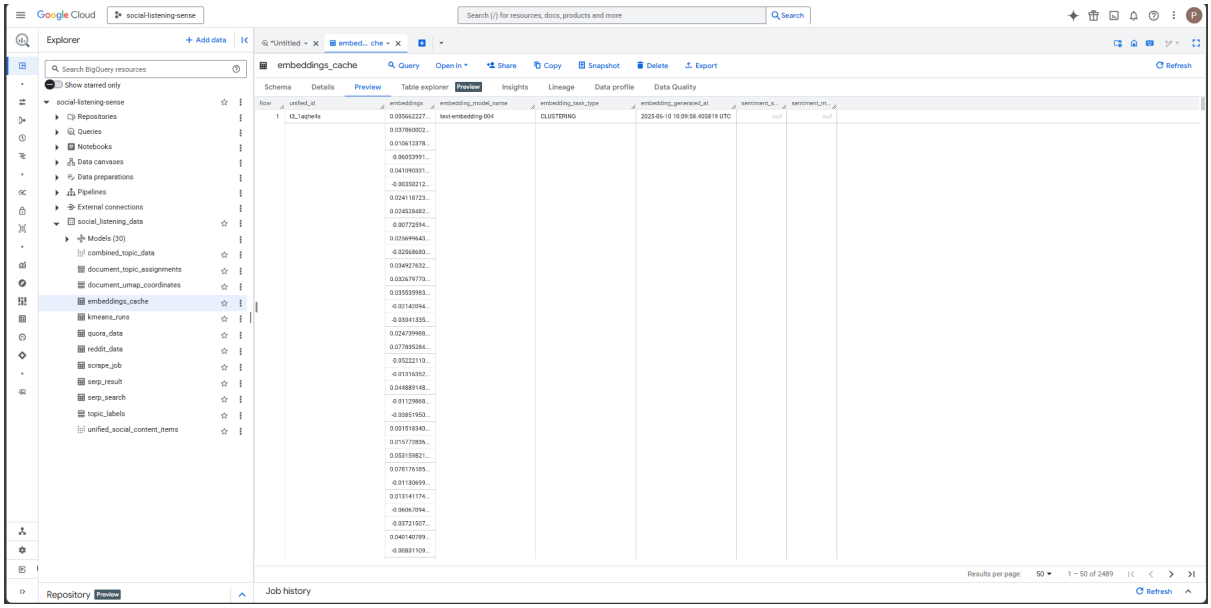


Figure 13.7 BigQuery embeddings_cache Table – Stores generated text embeddings and sentiment scores for each content item, avoiding redundant computation.

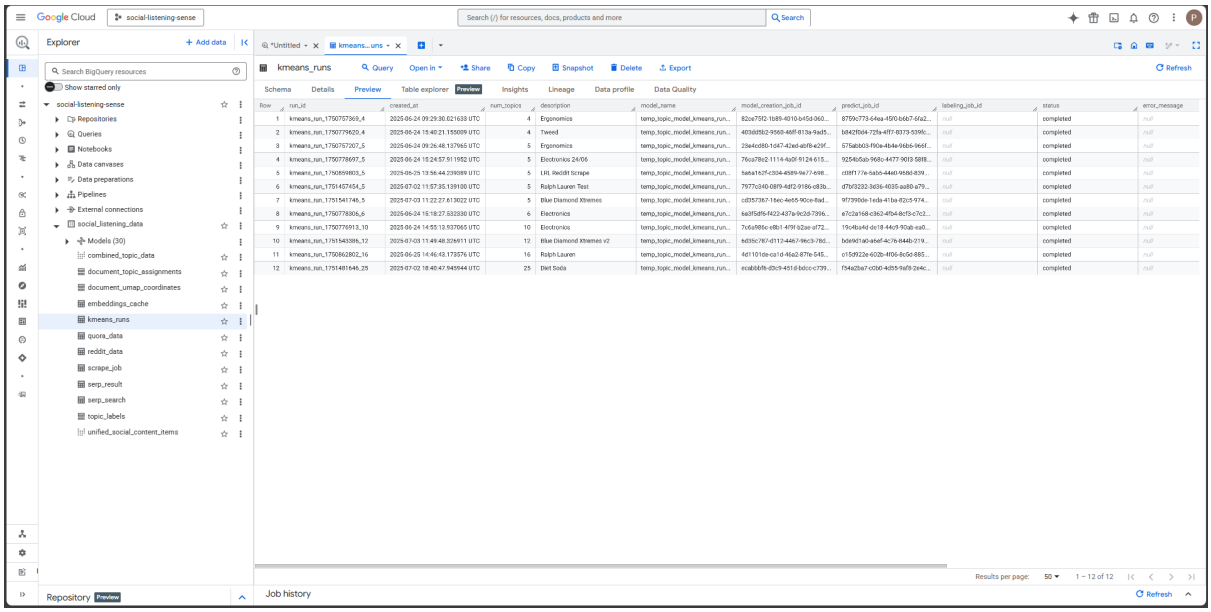


Figure 13.8 BigQuery kmeans_run Table – Metadata repository tracking topic modeling runs, cluster counts, and associated parameters.

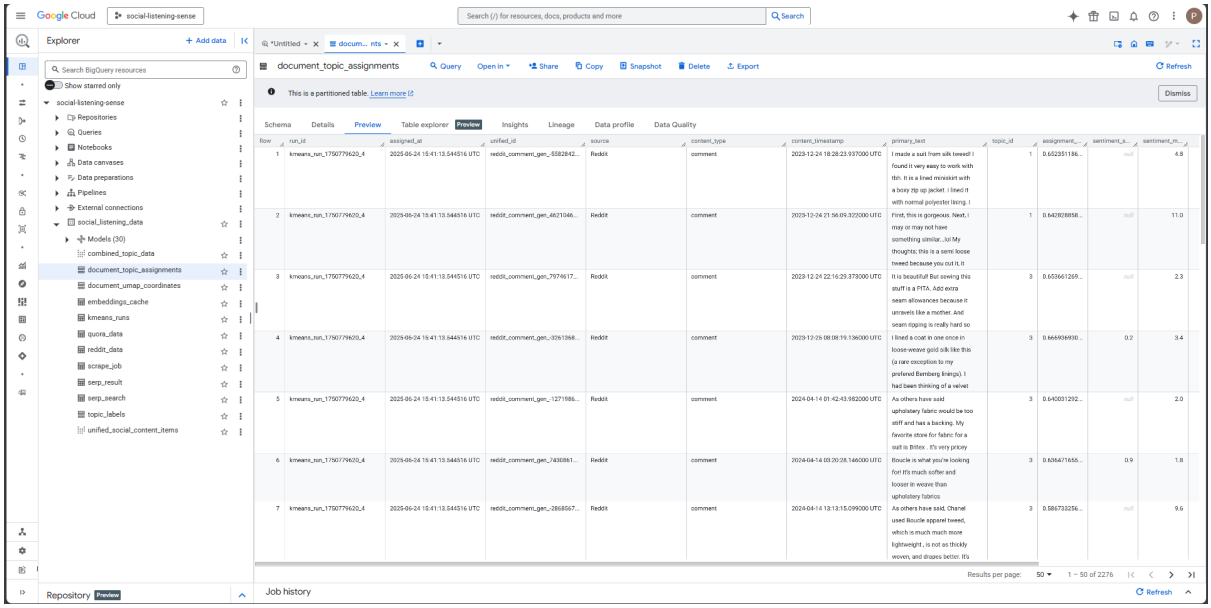


Figure 13.9 BigQuery document_topic_assignments Table – Links each content item to its assigned topic cluster, supporting both analytics and visualization.

13.3 Project Management

13.3.1 Critical Reflections

I initially planned for the project phases to follow a strict sequence starting with data collection, then system setup, and so on. However, I soon realized that these phases were more interconnected than expected. For example, I had to consider elements of the system architecture while still developing the topic modeling solution. As a result, the phases overlapped and influenced one another, which differed significantly from my original plan.

The planned two-month development and one-month writing split generally worked well. However, in hindsight, allocating some writing time earlier in the process might have allowed for more in-depth reflection and analysis of the development phase. It's a trade-off, but for future projects, I would consider a more integrated approach that blends writing and development more fluidly.

13.3.2 Project timeline

I used Miro and Google Calendar/Tasks to manage my project timeline. The timeline views illustrate task duration, overlap, and milestones.

Tasks were color-coded by category to enhance clarity. The categories included:

Figure 13.10 Color-coded task categories

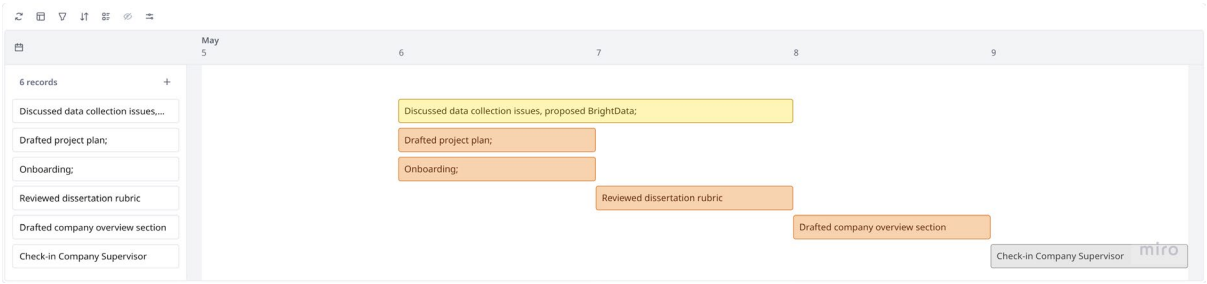


Figure 13.11 Miro Timeline May 5 - May 9

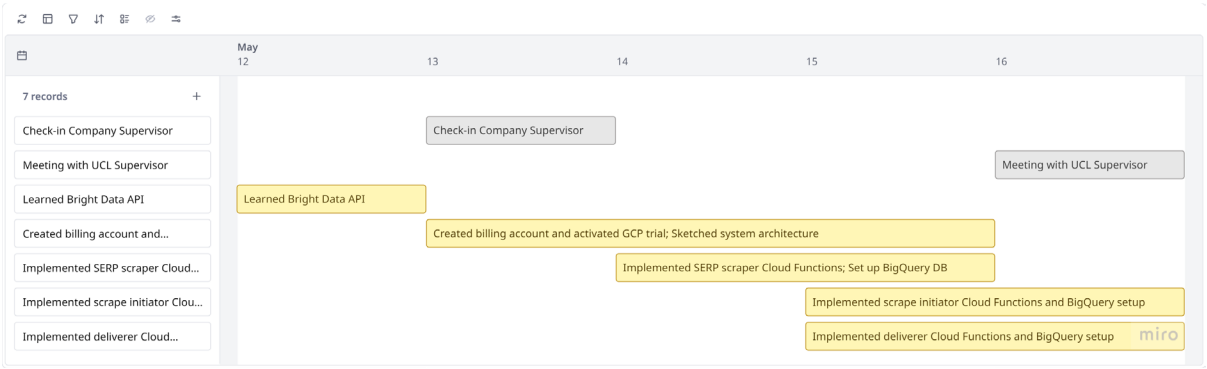


Figure 13.12 Miro Timeline May 12 - May 16

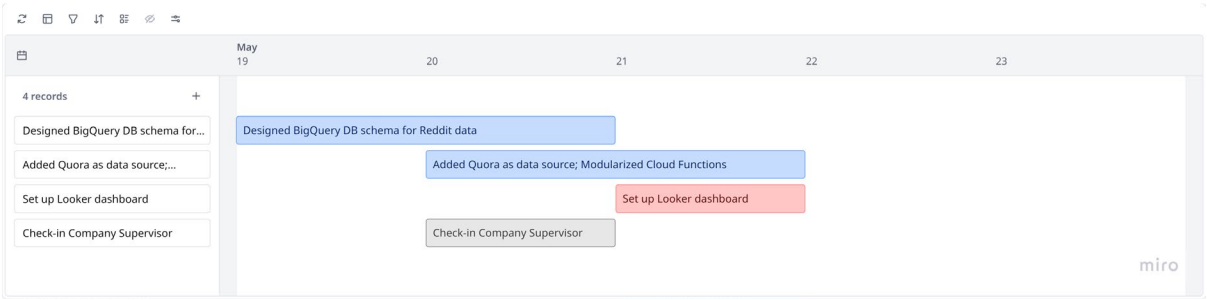


Figure 13.13 Miro Timeline May 19 - May 23

Figure 13.14 Miro Timeline May 26 - May 30

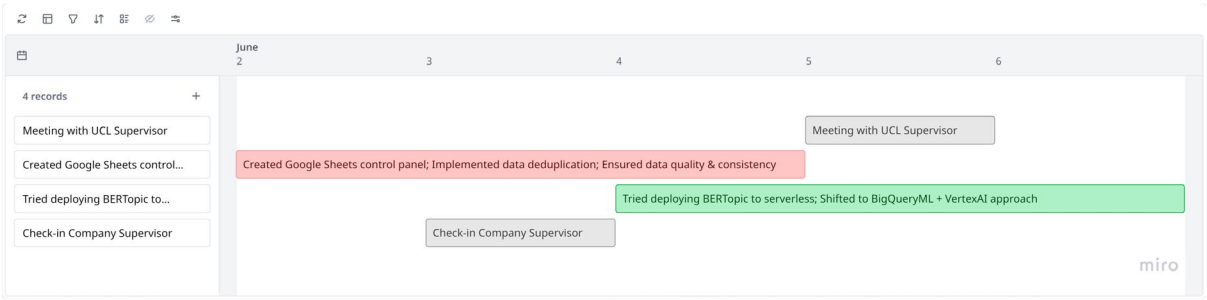


Figure 13.15 Miro Timeline June 2 - June 6

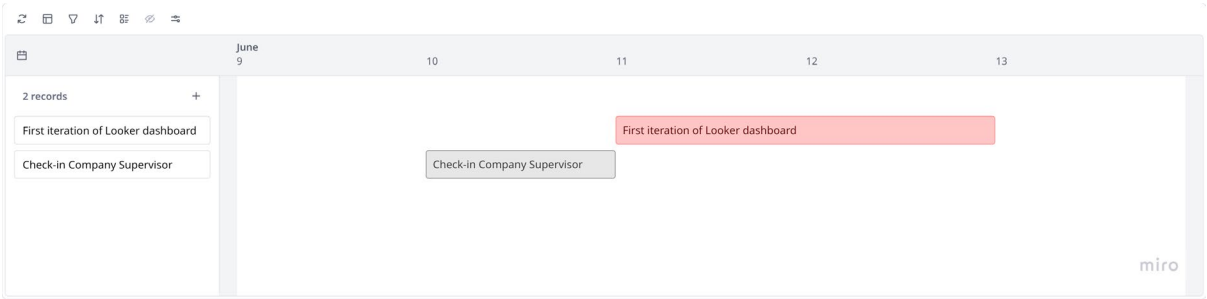


Figure 13.16 Miro Timeline June 9 - June 13

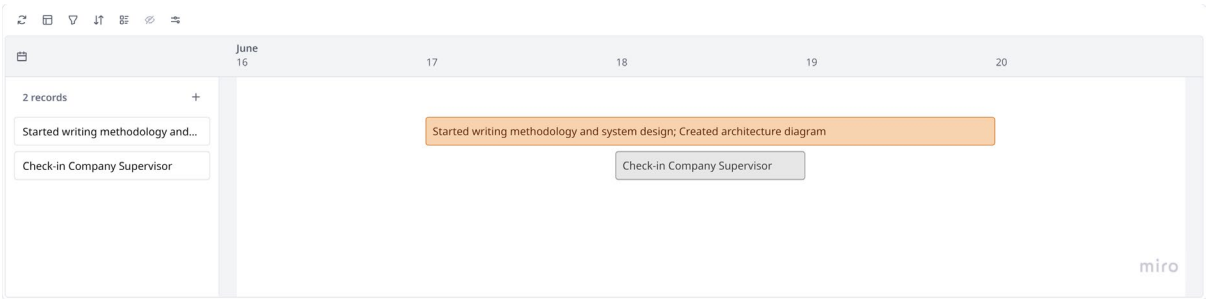


Figure 13.17 Miro Timeline June 16 - June 20

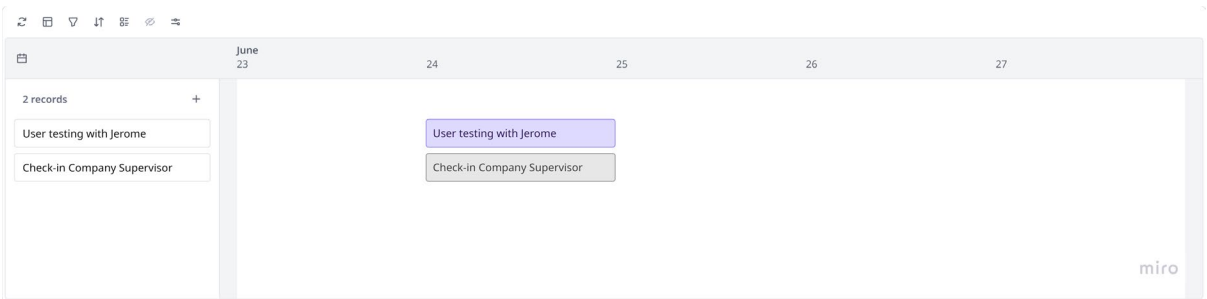


Figure 13.18 Miro Timeline June 23 - June 27

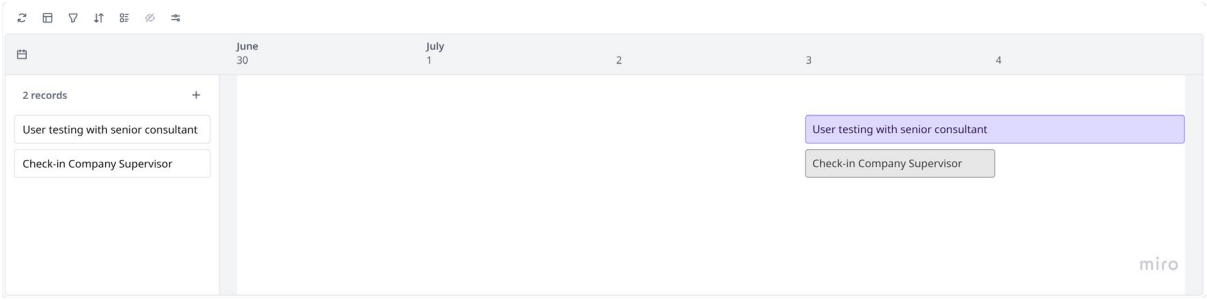


Figure 13.19 Miro Timeline June 30 - July 4

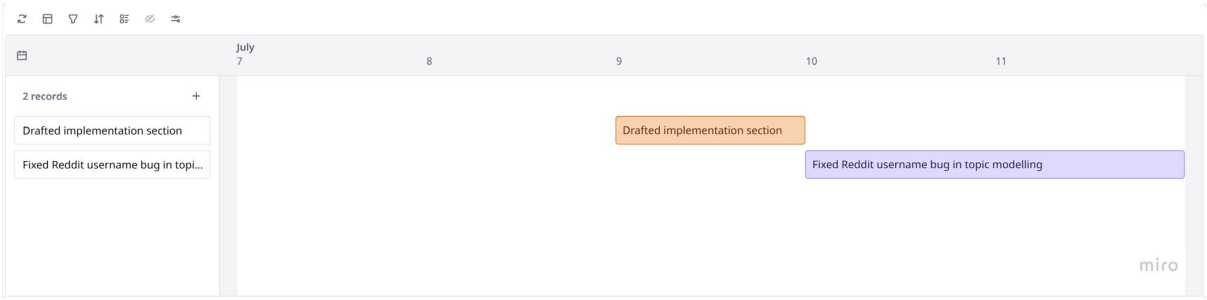


Figure 13.20 Miro Timeline July 7 - July 11

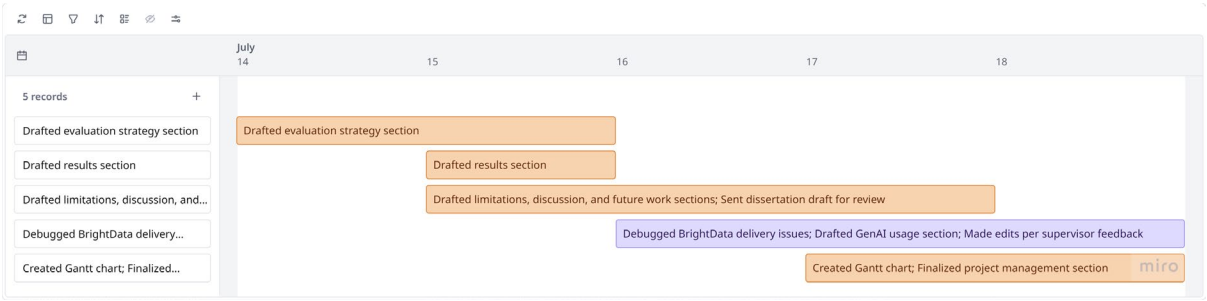


Figure 13.21 Miro Timeline July 14 - July 18

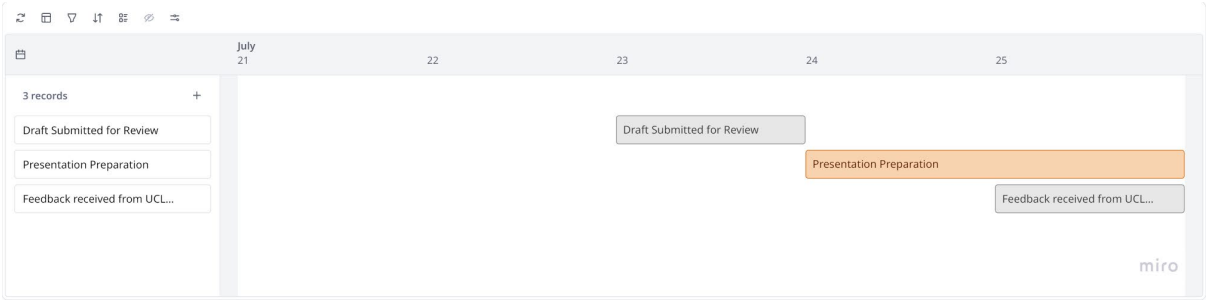


Figure 13.22 Miro Timeline July 21 - July 25

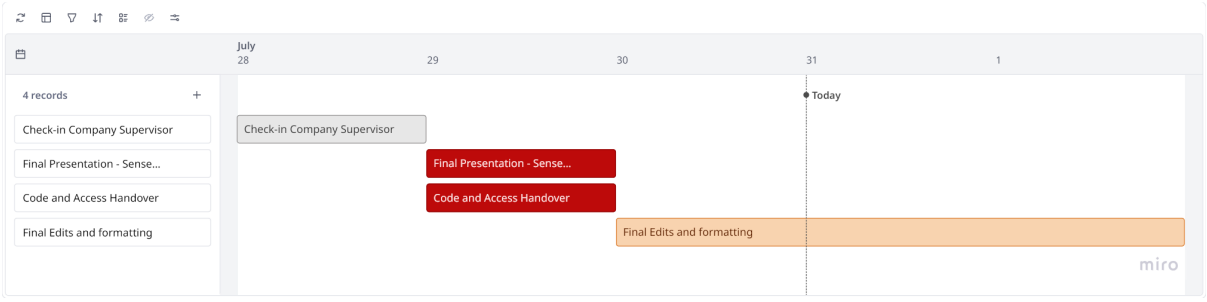


Figure 13.23 Miro Timeline July 28 - August 1

13.3.3 Meeting Notes

Onboarding Meeting: Me x Leila

Date: Tuesday, 6 May

Time: 10:15 – 11:15am

- setup access to necessary accounts and share permissions
- introduction of R&D at sense worldwide - past projects with ucl
- my project approach discussed
- initial expectations for project
- objective for the day - rough project plan

End of Day check-in: Me x Leila

Tuesday, 6 May 4:30 – 5:00pm

- discuss developed project plan
- setup a meeting with senior consultant for better understanding of company

End of Week Check-in: Me x Leila

Friday, 9 May 10:00 – 10:30am

- potential problems - data and scraping
- solutions found - brightdata services

Project Overview // Me x Freddie

Friday, 9 May 11:30am – 12:00pm

- In-depth understanding of Sense Worldwide's current needs and strategic context.
- priority should be making it easy to collect recent social media data (1year)
- Nice to have would be able to see how the trend changed over the years.

- Innovation Consultancy - Key Notes:
- Core Function: Helps businesses navigate challenges & identify growth opportunities.
- Client Input: Ranges from rough ideas to detailed problems.
- Example: Contact lens company - declining youth adoption.
- Phase 1: "Mind Expansion"
- Thorough desk research is conducted.
- Analysis includes:
- Company profiles
- Industry reports
- Competitor landscapes
- Other relevant data
- Goal: Comprehensive understanding of the issue.
- Phase 2: Collaboration & Qualitative Insights
- Close collaboration with the client.
- Leverages internal network for interviews.
- Goal: Gather rich qualitative data.
- Outcomes:
- Develop informed hypotheses.
- Test and validate hypotheses.
- Deliver validated ideas.
- Provide actionable strategies.
- Aim for tangible client results.
- Social listening tool will be used in addition to desk research and interview data to give a more holistic view of the current market, industry, and customers.

Check-in // Me x Leila

- Tuesday, 13 May 10:00 – 10:45am
- discussed system architecture
- google cloud - team most familiar with it

- brightdata api working shown
- need for frontend discussed
- backend technology choices - cloud function and pub sub
- bigquery as database
- data collection how to handle pagination
- how to get relevant post links
- solution use search api and use keywords from consultants

Dissertation Check in // Me x Tjun Hoh

Friday, 16 May

- concerns about data scraping discussed
- system architecture presented
- feedback given to start prototyping asap

Check-in // Me x Leila

- Tuesday, 20 May 10:00 – 10:30am
- duplicate record handling strategy - bigquery schema design (row number and latest timestamp)
- tracking serp jobs and serp results in same table or different tables
- bigquery upsert issues
- how to handle multiple sources (platforms reddit / quora etc) - solution - unified data model

Check-in // Me x Leila

Friday, 30 May 11:00 – 11:30am

- topic modeling prototype shown
- dashboard platform - looker studio chosen
- visualizations to include discussed

Check-in // Me x Leila

Tuesday, 3 June 10:00 – 10:30am

- deployment strategies of analytics to allow custom runs - bigquery ml vs docker containers
- control panel frontend platform - google sheets chosen (low maintainance , familiarity of consultants)

Dissertation Check in // Me x Tjun Hoh

Thursday, 5th June

- Attempted creating a microservice for topic modeling (failed due to deployment limits in Google Cloud Functions).
- Explored alternative cloud-based scalable solutions like BigQuery ML and Vertex AI for embedding and modeling.
- Commended for building an end-to-end working prototype.
- Praised for being ahead of schedule and demonstrating initiative.
- Key Advice:
- Emphasize prototyping, not production software.
- Include reflections on what analysts need and how tool compares to previous manual methods.
- Include business and theoretical justifications in the report.
- Suggested conducting informal interviews with company analysts.

Check-in // Me x Leila

Tuesday, 10 June 5:00 – 5:30pm

- dashboard plan - sentiment distribution, topic map, word cloud, topic distribution

Check-in // Me x Leila

Wednesday, 18 June 10:00 – 10:30am

- tool working demonstrated
- user testing meeting setup

User Testing // Jerome x Me

Tuesday, 24 June

- quick informal user demo
- hands on use.
- searched brand lauren ralph lauren
- they would not be able to use if i didnt explain.
- interesting aspect discovered - fake luxury market within reddit

Check-in // Me x Leila

Thursday, 3 July 10:30 – 11:00am

- final presentation scheduled

Me x Freddie Dashboard review

Thursday, 3 July 12:00 – 12:30pm

- summarized in results section

Dissertation Feedback // Tjun Hoh

Thursday 17 July

- get rid of signposting
- add high level overview of methodology
- add sources/references
- elaborate on feedback and introduction in relevance to project context and the company goals

Dissertation Feedback // Tjun Hoh

Thursday 17 July

- get rid of signposting
- formatting issues
- elaborate on discussion limitation future work
- Considers alternative interpretations/arguments.
- Identifies and justifies limitations with reasonable arguments.
- Draws conclusions/recommendations that are fully consistent with the evidence presented.

- critical evaluation of GenAI
- progress made towards business goals
- elaborate project planning with meeting summaries and critical reflections

Me x Leila // Check-in

Monday, 28 July 10:00 – 10:30am

- presentation feedback - add slide for how genAI used in topic labelling

UCL 2025 Project - Final Presentation (Sense Worldwide)

Tuesday, 29 July 10:00 – 10:40am

- interesting discussion points
- bias if 1000s of reddit post and limited inputs from other sources
- meaning of the axes in topic map visualization
- data privacy concerns
- page number control in data collection
- positive acknowledgement of the control panel (google sheets)
- k means no of clusters control how it works - bigger k more specific cluster smaller more broader
- opinions talked on social media don't always equal to what people think

13.4 Source Code Repository

The full source code for this project has been made publicly available on GitHub. The repository includes all components required to replicate the system, including:

- Google Cloud Functions (Python) for orchestration and analysis
- SQL models for BigQuery (embedding, clustering, sentiment, topic labeling)
- Google Apps Script files for the low-code Google Sheets control panel

GitHub Repository: https://github.com/prathamskk/social-listening-tool-dissertation