Web application for automated collection, analysis, and visualization of YouTube data

Brandon Calderón Prieto

School of Systems and Computer Engineering

Faculty of Engineering

Universidad del Valle

Professional project

Advisor: Dr. Robinson Andrey Duque Agudelo

Ph.D., Universidad del Valle

Co-Advisor: Dr. Víctor Bucheli Guerrero

Ph.D., Universidad del Valle

December 15, 2025

Contents

List of Tables

Chapter 1

Abstract

YouTube, with 2,530 million monthly users, serves as a primary news source for 35% of U.S. adults. However, current YouTube analysis tools rely on proprietary, black-box methodologies that limit transparency and reproducibility, creating barriers for academic research requiring verifiable systems.

This thesis proposes to develop a web application to **process**, **analyze**, and **visualize** YouTube data using transparent and reproducible methods. The system comprises three integrated components: a data collection module using the YouTube Data API with natural language processing for query translation, a data processing pipeline for cleaning and normalizing responses, and an interactive visualization layer.

The methodology follows an adapted Scrum framework with iterative two-week sprints. Development includes technology stack selection, system architecture design, API integration with LLM-powered interfaces, database implementation, and comprehensive testing. All services are containerized using Docker for reproducible deployment.

The resulting application will provide a no-cost, transparent alternative to commercial tools, democratizing YouTube data analysis for academic researchers, non-profits, and independent investigators. By providing fully documented methodology and open architecture, this work promotes **reproducible science** and serves as a **foundational tool** for research

in social media analysis, polarization, and information dissemination.

Chapter 2

Problem statement

According to Statista Dixon, 2025, YouTube is the second most used social network globally, only behind Facebook, with 2,530 million active monthly users. Notably, 35% of U.S. adults regularly get news from this video-focused platform (St. Aubin & Liedke, 2025). These facts highlight how important it is to do any kind of analysis.

Although enterprise social-listening suites (like "Talkwalker", n.d.; "Meltwater", n.d.) and social science research tools (like "Communalytic", n.d.) monitor YouTube, their pipelines and models are proprietary, limiting methodological transparency, reproducibility, and academic scrutiny for topic-level studies. This proprietary approach contradicts with the principles of **transparent and reproducible science**.

Academic researchers often lack a verifiable system that automates collection, cleaning, descriptive analytics, and insight generation while exposing schemas for independent replication.

This need is exemplified by ongoing research projects that require reliable social media data for analysis. The PROMUEVA project (Computational Models of Social Networks Applied to Polarization in Valle del Cauca) at Universidad del Valle represents such a case. This multidisciplinary project aims to develop computational models to analyze, measure, and predict polarization in social networks, a phenomenon that has significantly impacted

Valle del Cauca and Cali specifically (Universidad del Valle, 2023). To achieve its objectives of developing mathematical models of social network characteristics and user cognitive biases, PROMUEVA requires access to transparent, reliable, and reproducible social media data collection systems. However, the project faces the same methodological barriers: proprietary tools provide data without transparency, while building custom systems demands substantial technical resources. This exemplifies the broader need for accessible, documented tools that enable rigorous academic research on social media phenomena.

2.1 Problem formulation

How to develop a web application to process, analyze and visualize YouTube data?

Chapter 3

Problem justification

The use of social media has resulted in profound changes in the pattern of human interaction in today's contemporary society. The presence of social media has emerged as a phenomenon that places itself as a crucial element in the drastic changes in the way people interact in an era where information technology has undeniable dominance in almost every field of life (Azzaakiyyah, 2023). As these platform become central to public discourse and information dissemination, the methods used to analyze them face increasing scrutiny. However, much of the existing analysis is conducted using proprietary, "black-box" tools that limit transparency and prevent academic certification. This creates a critical need for **an application built on a reproducible, auditable, and documented methodology** for a major platform like YouTube, thereby empowering researches and the public with verifiable insights.

3.1 Academic

This project directly addresses a critical methodological gap in academic research: the lack of a unified, verifiable system for YouTube data analysis. Its core academic contribution is the promotion of **reproducibility** and **methodological transparency**. By providing a **fully documented methodology within this thesis**, the application's architecture and data pipeline allow for the independent verification of research findings. This application

provides a **foundational tool**, enabling researchers to conduct reproducible analyses and serving as a **validated base for future work** in more advanced data analysis. This creates a valuable and reusable asset for the academic community, ensuring that future topic-level studies of YouTube can be conducted with higher methodological rigor and transparency.

Concretely, this application will support active research initiatives such as the PROMUEVA project at Universidad del Valle, which requires social media data to develop computational models for analyzing polarization in Valle del Cauca. By providing a transparent data collection infrastructure, this work enables such projects to access YouTube data with full methodological documentation, supporting their research on polarization patterns and user behaviors in social networks.

3.2 Economic

The primary economic benefit is providing a **no-cost, high-value alternative** to expensive social listening tools. Commercial tools like Talkwalker or Meltwater represent a significant financial barrier for academic institutions, non-profits and independent researchers, limiting their ability to conduct large-scale analysis. By developing a **free tool**, this project **democratizes access** to these analytical capabilities, providing functionality that is otherwise locked behind costly subscriptions. Furthermore, the application **automates the entire data pipeline**, from collection and cleaning to analysis and visualization. This automation creates significant efficiency gains, saving countless hours of manual labor that researchers would otherwise spend on data preparation.

3.3 Social

Given that a substantial portion of the population now uses YouTube as a primary source for news and information, understanding the discourse on this platform is a matter of public interest. This project contributes to social good offering a **transparent** tool for analyzing

public opinion, which can be used by academic researchers. Unlike proprietary "black-box" systems, this tool's **publicly documented methodology** ensures that its processes are completely **transparent** and **auditable**. This fosters greater trust and empowers organizations to monitor topics like misinformation, public health discussions or political sentiment with a **verifiable** and **accessible** tool.

Chapter 4

Objectives

4.1 General objective

To develop a web application that automates the collection, processing, analysis and visualization of data from YouTube videos.

4.2 Specific objectives

1. **To design** the system architecture for a web application comprising three integrated components: data collection from YouTube Data API, data processing and analysis, and data visualization through charts.

2. **To implement** a data collection module featuring a Natural Language Interface (NLI) that utilizes a pre-trained Large Language Model (LLM) to validate and translate user requests into structured YouTube Data API query parameters.

3. **To implement** data processing and visualization components responsible for cleaning, normalizing, visualizing and structuring raw API responses into a standardized schema.

4. **To test** the implemented system through functional testing of each component and integration testing of the complete workflow from user query to visual output.

Chapter 5

Project Scope

5.1 Inclusions

This section details the specific features, deliverables, and technologies that are part of the project.

5.1.1 Functionality

The project will deliver a fully functional web application that automates the collection, processing, analysis, and visualization of data from YouTube. The core functionalities include:

- **Data collection module**: a configurable module will be implemented to collect video and associated comment data directly from the YouTube Data API using specified query filters.
- **AI query module**: a Natural Language Interface that integrates with a pre-trained AI model to translate informal user requests into valid YouTube Data API queries.
- **Data processing pipeline**: the application will feature a processing pipeline designed to clean, normalize, validate, and structure the raw data retrieved from the API.

- **Presentation layer**: a user-facing presentation layer will be built, featuring charts to visualize the analyzed data.

5.1.2 Technology

The specific technology stack for the application will be selected and justified as part of the implementation phase. However, the architecture will adhere to the following principles:

- **Backend service**: the core application logic and API will be developed using a modern web framework, with a relational database for data storage.
- **Frontend service**: a lightweight web server (e.g., Nginx) will be used to serve the static frontend assets (HTML, CSS, and JavaScript) built from a modern JavaScript framework (e.g., React, Vue).
- **Containerization**: all services (web server, backend application, database) will be containerized using Docker to ensure a consistent and reproducible deployment environment.

5.2 Exclusions

To maintain a clear focus and ensure the project is achievable within the defined timeframe, the following features and functionalities are explicitly out of scope:

- **Support for other platforms**: data collection and analysis will be limited exclusively to YouTube. The application will not support any other social media or video platforms (e.g., Facebook, Twitter, TikTok).
- **Training of NLP models**: the project will **integrate** existing, pre-trained models (for sentiment analysis and query translation). The scope does not include the **training**

or fine-tuning of new or custom natural language processing or machine learning models from scratch.

- **Real-time data streaming**: the data collection process is designed to be on-demand based on user queries. The project will not implement real-time or live-streaming data analysis.

- **Native mobile application**: the final deliverable is a web application accessible through standard desktop browsers. The development of native mobile applications for iOS or Android is not included.

Chapter 6

Reference framework

6.1 Glossary

- **Methodological transparency (new term)**: a practice in research where the methods, procedures, and analysis techniques are documented and shared in sufficient detail to allow other researchers to understand exactly how the study was conducted.
- **API (Application Programming Interface)**: "a set of functions and procedures that allow the creation of applications which access the features or data of an operating system, application, or other service, enabling third parties to use the functionality of that software application". ("Application Programming Interface", n.d.)
- **LLM (Large Language Model)**: "a narrow artificial intelligence (AI) system that has been trained on a massive amount of text data to interpret natural language and generate human-like responses to text-based prompts or questions". (Almarie et al., 2023)
- **Computational reproducibility**: "obtaining consistent results using the same input data, computational methods, and conditions of analysis". (National Academies of Sciences, Engineering, and Medicine, 2019)

- **Replicability**: "an attempt by a second researcher to replicate a previous study is an effort to determine whether applying the same methods to the same scientific question produces similar results". (National Academies of Sciences, Engineering, and Medicine, 2019)

- **Social network site**: "a networked communication platform in which participants (a) have uniquely identifiable profiles that consist of user-supplied content, content provided by other users, and/or system-level data; (b) can publicly articulate connections that can be viewed and traversed by others; and (c) can consume, produce, and/or interact with streams of user-generated content provided by their connections on the site". (Aichner et al., 2021)

- **Social listening**: "monitoring and analyzing conversations that take place on social and digital channels to gain insights into customer opinions, preferences, and trends. It involves tracking mentions of a brand, product, or relevant keywords, and analyzing the sentiment and context of these conversations". (Emplifi, n.d.)

- **Sentiment analysis**: "can be stated as the procedure to identify, recognize, and/or categorize the users' emotions or opinions for any service like movies, product issues, events, or any attribute as positive, negative, or neutral". (Bordoloi & Biswas, 2023)

6.2 State of art

6.2.1 LLM-based query generation

Recent advancements in natural language processing have demonstrated the efficacy of Large Language Models (LLMs) in translating informal user intent into structured technical commands. A prominent application of this capability is Text-to-SQL generation, which lowers the barrier to entry for interacting with complex database systems by allowing users to query data using natural language. The reliability of this translation process has improved

significantly with modern models; as noted by Zhu et al. (2024), the enhanced inference and generalization capabilities of current LLMs allow them to generate correct SQL queries with a high degree of accuracy. For instance, models such as GPT-4 have achieved state-of-the-art performance on benchmarks like Spider, outperforming previous rule-based approaches in natural language understanding.

Beyond database querying, this capability extends to API integration, which can be achieved effectively when clear documentation is available. Recent work demonstrates that LLMs can generate accurate API calls directly from natural language instructions (Tesfagiorgis & Monteiro Silva, 2023). This application validates the proposed strategy of using pre-trained LLMs as semantic translation layers between user intent and technical system interfaces, proving that effective integration is achievable without the need for resource-intensive model training or domain-specific fine-tuning.

This established precedent supports the proposed architecture of using a pre-trained LLM as a natural language interface for the YouTube Data API, where user queries are translated into valid API parameters through semantic understanding of intent and systematic mapping to API specifications.

6.2.2 Commercial social listening tools

Commercial social listening tools represent the closest existing solutions to the proposed system in terms of core functionality, offering capabilities such as multi-platform monitoring, sentiment analysis, and trend detection. Prominent platforms like Hootsuite (with Talkwalker) and Meltwater are optimized for business intelligence, brand monitoring, marketing analytics, and customer experience management across multiple social media platforms simultaneously, including Facebook, X (formerly Twitter), Instagram, TikTok, LinkedIn, YouTube, and Reddit, processing millions to hundreds of millions of data sources daily.

However, despite functional similarities, these commercial solutions operate under a fundamentally different scope and design philosophy. These tools prioritize real-time monitoring,

competitive intelligence, and actionable business insights over methodological transparency and scientific reproducibility. Their proprietary algorithms, closed-source architectures, and lack of access to raw data or processing methods make it impossible for researchers to verify results, replicate analyses, or understand the underlying computational processes, requirements that are essential for rigorous academic research.

Furthermore, enterprise pricing models (often ranging from thousands to tens of thousands of dollars annually) create significant accessibility barriers for individual researchers, small research teams, and academic institutions with limited budgets. While these platforms offer sophisticated AI-powered analytics and visualization capabilities suitable for corporate decision-making, they are inherently unsuitable for transparent, reproducible scientific research where methodological clarity, open access to data processing methods, and cost-effective accessibility are fundamental requirements. This gap between commercial enterprise tools and academic research needs establishes the motivation for developing a research-focused "social listening platform" that maintains methodological transparency while providing the analytical capabilities required for scientific study.

6.2.3 Existing open-source and academic tools

The landscape of open-source YouTube analysis tools presents a fundamentally fragmented ecosystem where individual components exist in isolation, but no comprehensive, integrated solution addresses the complete workflow required for reproducible academic research focused on YouTube data analysis. This fragmentation represents a critical gap in the available infrastructure for transparent YouTube-based social media research.

While numerous open-source tools provide specific functionalities for YouTube data analysis, such as comment extraction libraries (e.g., YouTube Data API wrappers in Python like `youtube-search-python` Mercerind, n.d.), sentiment analysis packages (e.g., VADER with 4.5K GitHub stars, TextBlob with 9K stars, spaCy with 30K stars), and data visualization frameworks (e.g., Matplotlib, Plotly, D3.js), none combine data collection, processing, analysis,

and visualization into a unified, documented, and methodologically transparent platform designed specifically for academic research purposes. Examples of fragmented tools include standalone projects like `youtube-comments-sentiment-analyzer` Deniz, n.d. and `YouTube Data Tools` "YouTube Data Tools", n.d., each addressing isolated components but requiring manual integration, custom scripting, and technical expertise to create a complete research pipeline.

This modular approach requires researchers to manually integrate disparate tools, each with different installation requirements, documentation standards, data formats, and API authentication methods, creating substantial technical barriers and introducing potential points of failure in reproducibility. Furthermore, existing tools typically lack the methodological documentation frameworks essential for academic research: documented workflows and processing pipelines that enable independent verification of research findings.

The value proposition of an integrated YouTube social listening tool for academic research thus extends beyond merely providing free alternatives to commercial tools. It centers on establishing documented, validated, and reproducible methodologies within a unified platform that:

1. Abstracts API complexity and quota management, including the translation of informal, natural language requests into valid API queries.

2. Provides standardized data collection and processing workflows.

3. Offers built-in sentiment analysis and visualization capabilities.

Most importantly, such a tool must explicitly acknowledge and document its dependency on the YouTube API as a fundamental limitation and design consideration, implementing strategies to mitigate API variability through consistent query parameters, timestamp documentation, and transparent reporting of data collection conditions.

This gap between fragmented, component-level tools and comprehensive, methodologically transparent research platforms, combined with the inherent fragility of API-dependent data

collection, establishes the primary motivation for developing an integrated free "YouTube social listening system" designed specifically to meet academic research standards for transparency, reproducibility, and methodological rigor, while explicitly addressing the constraints and limitations imposed by platform API dependencies.

6.2.4 Key academic studies

The academic community has increasingly utilized YouTube as a significant data source for social science research. This research is diverse: some studies focus on video content (such as education, health information, or politics); others investigate the recommendation algorithm (examining its relation to news, misinformation, or radicalization); and a significant portion specifically analyzes user comments to understand topics like hate speech, political ideology, gender differences, or sentiment (Deubel et al., 2024).

6.3 Theoretical framework

6.3.1 Interactions in the YouTube social network

Understanding the structure of interactions within YouTube as a social network is fundamental to designing a social listening tool that captures meaningful data for research purposes.

Roles within the YouTube ecosystem Sui et al. (2022) identify three primary roles that individuals occupy within the YouTube social network, each representing progressively higher levels of platform engagement.

- **Viewer**: represents represents the most basic level of engagement, consisting of individuals who interact with YouTube solely through passive video consumption. This role requires no account and represents the broadest, least specific form of platform engagement.

- **User**: extends beyond passive viewing to include active engagement mechanisms available through a Google account, such as leaving likes or dislikes on videos, posting comments, replying to existing comments, liking or disliking comments, and subscribing to channels.
- **Creator(or YouTuber)** : occupies the highest degree of platform engagement, actively producing content by posting videos, writing descriptions, curating channels, and building subscriber communities that can elevate them to micro-celebrity status.

These roles are not mutually exclusive (creators are simultaneously users and viewers) but they represent distinct analytical categories for understanding platform interactions.

Interaction mechanisms Building upon the framework proposed by Giglietto et al. (2012), Sui et al. (2022) categorize YouTube interactions into distinct types based on the nature of engagement.

- **Audience interactions**: the views.
- **Social interactions**: the likes, dislikes and comments.
- **Platform interactions**: the metadata like title, date, ID, etc.

Extended data categories

- **Engagement metrics**: quantifiable representations of viewer and user interactions with creators and the platform, including views, likes/dislikes, comments, comment replies, comment likes/dislikes, and subscriber counts. These metrics provide measurable indicators of content popularity, audience sentiment, and community activity.
- **Video/channel characteristics**: structural data independent of user interactions, including total video uploads, video duration, channel start date, video posting frequency, and upload schedules. These characteristics enable analysis of creator behavior patterns and content production strategies.

- **Textual data**: language and discourse elements available for qualitative and quantitative analysis, including video transcripts (extractable through YouTube's built-in transcript function), video titles, video descriptions, channel "About" pages, video tags, comments, and comment replies. This textual data provides rich material for content analysis, discourse analysis, sentiment analysis, and thematic exploration of community perspectives.

- **Visual data**: image and video content elements including the videos themselves (analyzable as complete recordings or extracted screenshots), video thumbnails, visual banners and annotations, environmental settings, camera angles, and visual representations of creators themselves (enabling analysis of demographic characteristics such as ethnicity, age, and gender presentation). Visual data facilitates ethnographic research, visual rhetoric analysis, and examination of multimodal communication strategies.

6.3.2 YouTube Data API limitations and computational reproducibility challenges

Beyond the diversity of methodological approaches employed by researchers, the YouTube Data API itself presents inherent technical limitations that must be carefully considered to ensure research reproducibility. Crucially, all YouTube data collection tools, whether open-source or commercial, depend entirely on the YouTube Data API, which is controlled by Google and subject to unilateral modifications without academic input. The discontinuation of the `relatedToVideoId` parameter in August 2023, which was previously essential for expanding datasets through recommendation networks, exemplifies how platform decisions can retroactively invalidate established research methodologies and prevent replication of earlier studies.

Recent empirical investigations reveal that the API's Search endpoint exhibits significant temporal variability in returned results. Efstratiou (2025) conducted a systematic audit by

running identical queries at 5-day intervals over 12 weeks across six diverse topics, finding that Jaccard similarity (measure of the similarity of two datasets) between video sets declined substantially over time, demonstrating that datasets collected using the exact same historical query may differ vastly based solely on when queries are executed.

These limitations underscore the critical importance of rigorous methodological practices in YouTube-based research. While the API's temporal inconsistency and platform dependency cannot be fully eliminated, researchers can mitigate its impact through transparent documentation of collection timestamps and API parameters and explicit acknowledgment of these platform-imposed constraints in their methodology sections. Tools and frameworks that systematically document API dependencies, timestamp data collection, and maintain transparent records of all methodological decisions are essential for conducting replicable YouTube research within these known limitations.

6.3.3 Scrum

Scrum is an agile project management framework that facilitates team collaboration through iterative and incremental development (Schwaber & Sutherland, 2020). The framework prescribes for teams to break work into goals completed within time-boxed iterations called sprints, typically lasting two to four weeks (Schwaber & Sutherland, 2020). Scrum is based on three pillars (transparency, inspection, and adaptation) and emphasizes self-organization, continuous feedback, and flexibility in response to changing requirements (Schwaber & Sutherland, 2020). Originally developed for software development, Scrum has since been successfully applied to various complex projects requiring iterative progress and stakeholder collaboration (Schwaber & Sutherland, 2020).

Chapter 7

Methodology

This project adopts an adapted Scrum framework to guide the software development process. While Scrum is traditionally designed for team-based development, this thesis applies specific Scrum artifacts and events that provide particular value for individual academic software projects, especially in maintaining transparency and systematic progress tracking.

The choice of Scrum is motivated by two primary factors: its **flexibility** through iterative development and its emphasis on **structured documentation** through artifacts. This aligns well with the exploratory nature of thesis development, where requirements may evolve as the research progresses.

7.1 Scrum artifacts adapted for this project

7.1.1 Product backlog

The product backlog serves as an emergent, ordered list of what is needed to improve the product and functions as the single source of work for the project (Schwaber & Sutherland, 2020). For this thesis, the **product backlog** will document all planned features, technical requirements, and improvements for the web application. This artifact provides transparency in planning and allows for systematic prioritization of development tasks based on their

contribution to thesis objectives.

7.1.2 Sprint backlog

Composed of the **sprint goal**, selected **product backlog** items, and an actionable plan for delivering an increment (Schwaber & Sutherland, 2020), the **sprint backlog** will serve as the operational plan for each development iteration. For this individual project, sprints will be two weeks periods during which specific functionality will be developed and integrated into the application.

7.2 Scrum events adapted for this project

While Scrum events are designed for team collaboration, certain events will be adapted to provide structure and facilitate self-reflection.

7.2.1 Sprint planning

At the beginning of each sprint, sprint planning will determine what can be delivered in the upcoming sprint and how that work will be achieved (Schwaber & Sutherland, 2020). This will involve selecting **product backlog** items and defining a clear **sprint goal**.

7.2.2 Sprint retrospective

The sprint Retrospective provides an opportunity to plan ways to increase quality and effectiveness (Schwaber & Sutherland, 2020). For this individual project, retrospectives will be conducted at the end of each sprint to reflect on what went well, what challenges were encountered, and what adjustments should be made for subsequent sprints. This practice supports continuous improvement and systematic documentation of the development process.

Although Scrum is traditionally conceived as a framework for teams, with defined roles including **Scrum master**, **product owner**, and **developers** (Schwaber & Sutherland, 2020),

this project adapts its core principles to the context of individual thesis development. The emphasis remains on the empirical pillars of **transparency**, **inspection**, and **adaptation** (Schwaber & Sutherland, 2020), which are particularly valuable for maintaining rigor and documentation in academic software development projects.

7.3 Activities to be carried out

7.3.1 Specific objective 1 1

Table 7.1: Activities to be carried out for the specific objective 1

ID	Activity	Expected result
A1.1	Evaluate and select the technology stack for the software development project, including frontend framework, backend framework, database management system, and LLM API integration approach.	Technical report documenting the selected technology stack with justification based on project requirements.
A1.2	Conduct a literature review on YouTube Data API functionality, parameters, available endpoints, quota management, and documented limitations.	Technical document describing YouTube Data API capabilities, constraints, and best practices for research-oriented data collection.
A1.3	Design the database schema for storing collected YouTube data (videos, comments, replies), user projects, API queries, and processing metadata.	Database entity-relationship diagram with table specifications, relationships, and data types.

7.3.2 Specific objective 2 2

Table 7.2: Activities to be carried out for the specific objective 2

ID	Activity	Expected result
A2.1	Implement the backend data collection component with YouTube Data API integration, including authentication, query configuration, quota management, and error handling.	Source code for the data collection component with API integration, request handling, and response processing functionality.
A2.2	Implement the LLM integration module that processes natural language user requests and generates valid YouTube API query parameters using an existing LLM model via API.	Source code for the NLP interface component that translates user queries into structured API parameters with validation mechanisms.

7.3.3 Specific objective 3 3

Table 7.3: Activities to be carried out for the specific objective 3

ID	Activity	Expected result
A3.1	Implement the data processing pipeline for cleaning, normalizing, and structuring raw API responses.	Source code for the processing pipeline with data transformation, validation, and schema mapping functionality.
A3.2	Implement the database layer with the designed schema for storing videos, comments, replies, and associated metadata.	Database implementation with tables, relationships, and indexes following the designed schema.
A3.3	Implement data visualization components with charts.	Source code for visualization features.
A3.4	Implement user project management functionality allowing users to create, save, and manage multiple data collection projects.	Source code for project management features including CRUD operations for user projects and saved queries.

7.3.4 Specific objective 4 4

Table 7.4: Activities to be carried out for the specific objective 4

ID	Activity	Expected result
A4.1	Conduct functional testing of each component (data collection, LLM integration, processing pipeline, presentation layer) using synthetic data.	Test report documenting test cases, procedures, and results for each individual component with evidence of functional correctness.
A4.2	Conduct integration testing of the complete workflow from natural language input through data collection, processing, and visualization output.	Test report documenting end-to-end integration tests with evidence that all components work together correctly.
A4.3	Validate system functionality using synthetic data from two YouTube topics, ensuring all requirements meet the general objective.	Validation report with evidence of system operation using real test cases and documentation of results against requirements.

7.4 Schedule of activities

The following Gantt chart (Figure 7.1) illustrates the temporal distribution of all activities across the project timeline from February to October. Activities are organized according to their corresponding specific objectives, with dependencies and overlaps clearly indicated to ensure efficient resource allocation and logical progression of the development process.

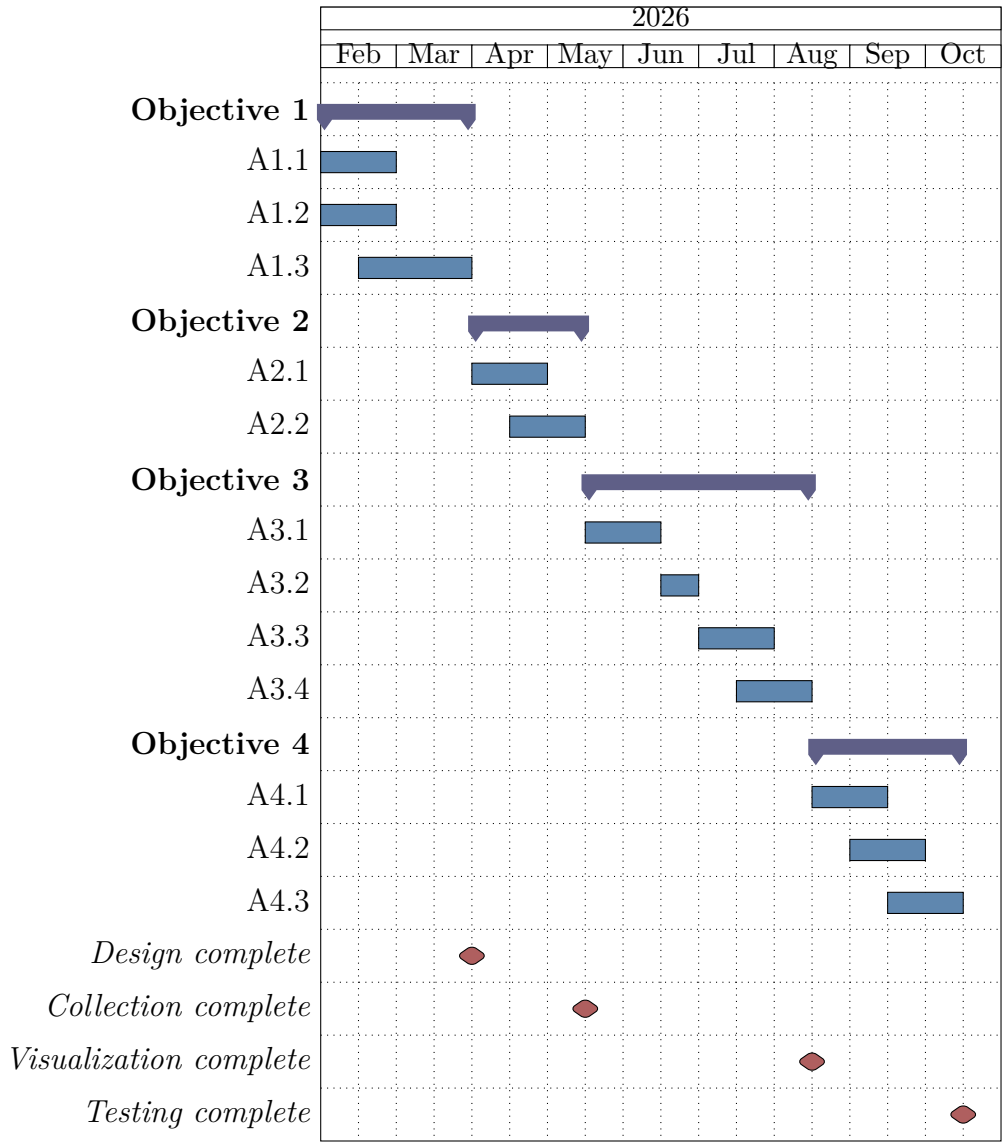

Figure 7.1: Gantt chart showing the schedule of activities from February to October, organized by specific objectives with dependencies indicated by connecting arrows.

Chapter 8

Budget

8.1 Human resources

Table 8.1: Human resources table

Person	**Weekly dedication**	**Hourly rate**	**Total**
Advisor	1	$\$110,000$	$\$3,960,000$
Co-Advisor	1	$\$110,000$	$\$3,960,000$
Student	20	$\$15,000$	$\$10,800,000$
		Total	$\$18,720,000$

8.2 Materials and services

Table 8.2: Materials and services table

Category	**Item**	**Cost**
Infrastructure and equipment	Internet service	$\$450,000$
Training and development	Technology courses	$\$300,000$
	Technical books and resources	$\$200,000$
Operational costs	Electricity	$\$360,000$
	Transportation	$\$100,000$
	Total	$\$1,410,000$

References

Aichner, T., Grünfelder, M., Maurer, O., & Jegeni, D. (2021). Twenty-five years of social media: A review of social media applications and definitions from 1994 to 2019. *Cyberpsychology, Behavior and Social Networking*, *24*(4), 215–222. https://doi.org/10.1089/cyber.2020.0134

Almarie, B., Teixeira, P. E. P., Pacheco-Barrios, K., Rossetti, C. A., & Fregni, F. (2023). Editorial – the use of large language models in science: Opportunities and challenges. *Principles and Practice of Clinical Research*. https://doi.org/10.21801/ppcrj.2023.91.1

Application programming interface. (n.d.). ScienceDirect Topics. Retrieved October 21, 2025, from https://www.sciencedirect.com/topics/computer-science/application-programming-interface

Azzaakiyyah, H. K. (2023). The impact of social media use on social interaction in contemporary society. *Technology and Society Perspectives (TACIT)*. https://doi.org/10.61100/tacit.v1i1.33

Bordoloi, M., & Biswas, S. K. (2023). Sentiment analysis: A survey on design framework, applications and future scopes. *Artificial Intelligence Review*. https://doi.org/10.1007/s10462-023-10442-2

Communalytic. (n.d.). Communalytic. Retrieved November 6, 2025, from https://communalytic.org

Deniz, C. (n.d.). *Youtube comments sentiment analyzer*. Retrieved November 6, 2025, from https://github.com/coskundeniz/youtube-comments-sentiment-analyzer

Deubel, A., Breuer, J., Kohne, J., & Mohseni, M. R. (2024). *Overview of working with youtube data* (tech. rep.). GESIS - Leibniz Institute for the Social Sciences. https://doi.org/10.60762/ggdbd24012.1.0

Dixon, S. J. (2025, March). *Most popular social networks worldwide as of February 2025, by number of monthly active users*. Statista. Retrieved October 2, 2025, from https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users

Efstratiou, A. (2025). On youtube search api use in research. *Proceedings of the 2025 ACM Internet Measurement Conference*. https://doi.org/10.1145/3730567.3764492

Emplifi. (n.d.). *Social listening*. Emplifi. Retrieved October 21, 2025, from https://emplifi.io/definitions/social-listening/

Giglietto, F., Rossi, L., & Bennato, D. (2012). The open laboratory: Limits and possibilities of using facebook, twitter, and youtube as a research data source. *Journal of Technology in Human Services*. https://doi.org/10.1080/15228835.2012.743797

Meltwater. (n.d.). Meltwater. Retrieved November 6, 2025, from https://www.meltwater.com

Mercerind, A. (n.d.). *Youtube-search-python: Search for youtube videos, channels & playlists*. Retrieved November 6, 2025, from https://github.com/alexmercerind/youtube-search-python

National Academies of Sciences, Engineering, and Medicine. (2019). *Reproducibility and replicability in science*. The National Academies Press. https://doi.org/10.17226/25303

Schwaber, K., & Sutherland, J. (2020, November). *The scrum guide: The definitive guide to scrum: The rules of the game*. Retrieved October 27, 2025, from https://scrumguides.org/scrum-guide.html

St. Aubin, C., & Liedke, J. (2025, September). *Social media and news fact sheet* [Survey conducted August 18-24, 2025]. Pew Research Center. Retrieved October 2, 2025, from https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet

Sui, M., Sui, Y., Zhang, H., Hu, Y., & Zhang, Y. (2022). Analyzing users' sentiment toward popular video games on youtube. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2022.940724

Talkwalker. (n.d.). Talkwalker. Retrieved November 6, 2025, from https://www.talkwalker.com

Tesfagiorgis, Y. G., & Monteiro Silva, B. M. (2023). *Large language models as an interface to interact with api tools in natural language* [Bachelor's thesis]. Linnaeus University. Retrieved December 15, 2025, from https://lnu.diva-portal.org/smash/get/diva2:1801354/FULLTEXT01.pdf

Universidad del Valle. (2023). *Promueva: Computational models of social networks applied to polarization in valle del cauca* [Research project at Universidad del Valle]. Retrieved November 1, 2025, from https://sites.google.com/view/promueva

Youtube data tools. (n.d.). Digital Methods Initiative. Retrieved November 6, 2025, from https://ytdt.digitalmethods.net/index.php

Zhu, X., Li, Q., Cui, L., & Liu, Y. (2024). Large language model enhanced text-to-sql generation: A survey. https://arxiv.org/abs/2410.06011