DATA 5901: Capstone 1
Project Definition

MSDS 25.4 “Voice of Market (VoM)”

Faculty Advisor: Dr. Elio Zhang

Sponsor Organization: Costco Wholesale

Team Members

Mustafa Bhavanagarwala
Hrishikesh Deepak Dhole
Lakshit Gupta
Priyal Sunil Joshi
Tejaswi Neelapu

Table of contents

1. Executive Summary

In today’s digital landscape, customer opinions shared on social media platforms like Reddit, Instagram, and Twitter (X) significantly impact brand perception and business strategy. Costco Wholesale, one of the world’s largest retailers, is committed to enhancing customer satisfaction and improving its product and service offerings. However, extracting meaningful insights from the vast volume of unstructured social media data is a challenge. Traditional methods such as surveys and focus groups are often biased, slow, and lack real-time responsiveness, making it difficult to adapt to evolving consumer needs.

This project leverages Natural Language Processing (NLP) and AI-driven sentiment analysis to transform raw customer feedback into structured insights. An automated pipeline will be developed to collect, preprocess, and analyze data using advanced techniques such as VADER for short-form sentiment analysis, BERT for contextual understanding, and AI-powered summarization.

Key deliverables include a sentiment analysis report, an interactive knowledge graph, an AI-powered chatbot, and a web-based visualization tool. These solutions will provide Costco with data-driven insights to improve decision-making, anticipate customer expectations, optimize its product line, and maintain a competitive edge in the evolving retail industry.

2. Background

Costco Wholesale is the third-largest retailer in the world known for providing valuable memberships to its customers. As of 2024, it serves approximately 137 million customers worldwide. High-quality and in-demand products are provided by Costco at very reasonable prices.[1]

Traditionally, companies rely on surveys, reviews from their own websites, and focus groups to understand customer needs. While these strategies do provide some information, they are time-consuming, often biased, and not real-time responsive. Additionally, Costco continuously looks for innovative ideas that can improve customer satisfaction and happiness. This can be done by anticipating customer needs through their opinions.

In today's fast-evolving digital world, social media platforms and direct company websites have become key channels for customer feedback. Customers express their thoughts on Costco's products and services through reviews, product ratings, survey responses, forum discussions, and social media posts. They share their experiences whether it's praising a new product, requesting improvements, or expressing dissatisfaction in the form of comments, tweets and posts. These insights come from various sources, including Costco's website, Instagram, Reddit discussion threads, and Twitter from customers worldwide. However, extracting meaningful insights from this vast amount of unstructured text data presents a significant challenge.

To address this challenge, this project leverages sentiment analysis and AI-powered summarization to systematically analyze large volumes of social media data. Sentiment analysis plays a crucial role in identifying whether customer feedback expresses positive, negative, or neutral sentiments, enabling businesses to track sentiment trends, assess customer satisfaction, and respond proactively to emerging concerns.[2]

A critical part of the data preprocessing pipeline involves natural language processing (NLP) techniques, such as tokenization, stopword removal, and lemmatization, to clean and structure the text before sentiment analysis. Tokenization is the process of breaking down text into individual words or phrases to facilitate further analysis. Stopword removal eliminates common words that do not contribute significantly to meaning, reducing noise in the data. Lemmatization reduces words to their base form, ensuring consistency in word variations. These preprocessing techniques enhance the quality of text data, making sentiment classification and topic extraction more accurate and efficient.

For short-form text commonly found on platforms like Reddit and Twitter, we will use VADER (Valence Aware Dictionary and Sentiment Reasoner), a rule-based sentiment analysis tool specifically designed for informal and social media language. VADER is highly effective in processing text that includes slang, emojis, punctuation-based intensity, and capitalization, making it well-suited for analyzing customer discussions on Costco's products and services. By incorporating VADER into the sentiment analysis pipeline, we can extract quick and reliable sentiment scores from customer feedback, allowing Costco to gain actionable insights into consumer perceptions.

This project also includes an AI-powered chatbot that allows interactive querying of sentiment trends, topic analysis, and insights gathered from the data to aid in stakeholder decision making. The chatbot makes data exploration more effective and actionable by enabling stakeholders to rapidly retrieve key takeaways, sentiment summaries, and emerging discussion themes in place of manually analyzing reports. This will ultimately increase member satisfaction and drive business growth.

3. Problem Statement

Costco needs an efficient way to analyze and interpret the vast amount of customer feedback available across social media platforms. Without a structured approach, valuable insights on product preferences, service experiences, and emerging trends remain hidden within unstructured text. This makes it difficult for Costco to make real-time, data-driven decisions that align with customer expectations.

The purpose of this project is to transform unstructured social media discussions into meaningful, structured insights that support strategic decision-making. By systematically processing and categorizing customer feedback, the solution will enable Costco to track emerging trends, measure sentiment, and refine its offerings based on real-world consumer experiences. The end goal is to develop a robust framework that helps Costco cautiously respond to customer needs, optimize products and services, and maintain its leadership in the retail industry.

4. Requirements

4.1 Data Collection and Preprocessing

The project requires collecting and analyzing social media data to understand customer sentiment toward Costco’s products and services. The data will be gathered from platforms like X (Twitter), Reddit and Instagram using web scraping tools such as BeautifulSoup, Scrapy, and Selenium.[3] Public APIs will also be used to streamline the data collection process.[4]

We have over 175,000 rows of data gathered from different social media platforms. Each row in the dataset which is a unique comment on a post contains the following fields:

Field Name	Description
Id	Unique identifier for each post
post_url	The URL of the post
comment_body	The text of the comment
post_title	The title of the post
source	The platform or source from which the post originates
post_year	The year the post was posted
post_month	The month the post was posted
comment_year	The year the comment was posted
comment_month	The month the comment was posted
category	Indicates whether the post relates to a product or service
token	Numerical token used for text processing

Table 1. Dataset Columns

The combined data volume is considered adequate to kickstart the analysis. This dataset will be analyzed to understand trends and discussions related to products. It will be updated every quarter as required to capture any shifts in sentiment or emerging topics.

4.1.1 Preprocessing

To handle the incoming data efficiently, a pre-optimized pipeline has been developed that can directly fetch updated data. However, for the current stage of the project, analysis will be conducted on the initially scraped data. During preprocessing, the data will be cleaned by removing duplicates, irrelevant content, and noise such as spam posts, ensuring that only authentic customer discussions contribute to the analysis. Spam detection will be implemented using a multi-layered approach combining rule-based filtering, machine learning classification, and NLP-based detection. Rule-based filtering such as excessive special characters, and suspicious links will also be used. Additionally, a TF-IDF and Logistic Regression model will help detecting statistical spam patterns, while an NLP-based BERT classifier will be used to identify contextually disguised spam messages.[11] This hybrid approach ensures accurate spam removal, preventing misleading data from influencing sentiment analysis and topic modeling. After filtering natural language processing (NLP) techniques including tokenization, stopword removal, and lemmatization will be applied to ensure that the text is ready for analysis.[5]

The data will be segmented based on product categories and services by analyzing the context and keywords present in customer discussions using topic modeling techniques, specifically Latent

Dirichlet Allocation (LDA). LDA is a probabilistic topic modeling method that identifies the main topics and how they are distributed throughout a collection of documents. LDA uses unsupervised learning to find latent themes in large text datasets without the need for predefined labels, in contrast to supervised learning techniques. [10]

LDA will be utilized in the project to categorize customer feedback into meaningful themes. To identify product-related feedback, dominant topics containing keywords associated with particular Costco products such as electronics, groceries, appliances, apparel etc were extracted. Feedback related to customer service will be categorized according to themes that represent experiences with checkout procedures, membership benefits, store operations etc. For example, the product category feedback section included comments like "The Kirkland Signature coffee is amazing and affordable," while the service feedback section included comments like "The self checkout lines at Costco are always too long."

As part of the data collection process, strict adherence to data privacy regulations will be maintained. Personal or sensitive data will not be collected in accordance with privacy laws such as the General Data Protection Regulation (GDPR) and the European Data Protection Board (EDPB).[7][8] Additionally, ethical scraping practices will be followed to avoid violating the platform's terms of service.

4.2 Sentiment Analysis and Insights

Sentiment analysis will classify customer feedback into positive, negative, or neutral categories. NLP models such as VADER and transformer-based models like BERT will be used as the foundation for performing sentiment classification. These models were selected based on their well-established effectiveness in analyzing sentiment from social media data.

4.2.1 VADER (Valence Aware Dictionary and sentiment Reasoner)

VADER is a rule-based sentiment analysis tool specifically designed for analyzing social media texts. Unlike traditional sentiment analysis models that rely on complex machine learning approaches, VADER is pre-trained and uses a lexicon of words and heuristics to assign sentiment scores to a given text. Each word in VADER’s dictionary is assigned a score ranging from -4 to +4, with -4 being the most negative and +4 being the most positive. Additionally, VADER considers textual cues such as capitalization, punctuation, and intensifiers (e.g., “great!!!” is more positive than just “great”) to determine sentiment intensity more accurately.

VADER is particularly effective for short, informal, and noisy texts, making it well-suited for platforms like Twitter, Reddit, and Instagram, where users frequently use slang, emojis, and abbreviations to express opinions. In this project, VADER will be applied to efficiently classify short-form customer feedback, enabling scalable and accurate sentiment analysis across Costco’s social media data.[9]

4.2.2 BERT (Bidirectional Encoder Representations from Transformers)

BERT is an advanced natural language processing model that analyzes text in both directions, taking into account both preceding and subsequent words to ascertain context. BERT is very effective for sentiment analysis because of its transformer-based architecture, which enables it to understand sentences in their entirety, unlike traditional models that read text sequentially.. Because it was pre-trained using Next Sentence Prediction (NSP) and Masked Language Modeling (MLM), it can handle the kind of informal, ambiguous, and complex text that is frequently found in social media conversations.[13]

BERT will be used in this project to categorize social media customer reviews into three groups: positive, negative, and neutral. It can also capture the complex viewpoints and conflicting emotions that are frequently present in conversations about Costco. While BERT's contextual understanding increases accuracy, traditional sentiment analysis models have trouble with sarcasm, informal language, and abbreviations. Furthermore, by identifying sentiment patterns and classifying reviews according to goods and services, it can offer a data-driven approach for understanding customer experiences.

BERT is the ideal tool for analyzing Costco's customer feedback as it can handle long reviews and responses with multiple sentences. Costco can obtain more precise and scalable insights from social media conversations by incorporating BERT into our sentiment analysis pipeline. This will allow for improved product and service optimization based on real customer sentiment trends.

4.3 AI-Powered Summarization and Chatbot

An AI-powered summarization model will be implemented to extract key insights from customer feedback. Both extractive and abstractive summarization techniques will be considered to ensure that the feedback is represented meaningfully.

Extractive Summarization: This method will select and highlight the most relevant sentences or segments from customer feedback without changing the wording. It helps to directly reflect the core ideas from the feedback.
Abstractive Summarization: This method involves generating a summary using a deeper understanding of the content. The model will create new, concise sentences that summarize the key points, providing a more comprehensive overview of customer sentiment.

The combination of these techniques ensures that detailed insights and high-level summaries can be extracted effectively, enabling Costco to quickly identify trends, common concerns, and emerging opportunities in customer feedback.[12]

Chatbot for Stakeholder Interaction

A chatbot will be developed to assist stakeholders in retrieving and interpreting sentiment insights from social media discussions. Instead of manually analyzing lengthy reports, stakeholders can interact with the chatbot to query customer sentiment trends, product-related discussions, and emerging concerns in real time. This will allow for faster decision-making and improved responsiveness to customer expectations.

To implement this, two chatbot frameworks are being considered:

4.3.1 Rasa

Rasa is an open-source framework designed for complex, multi-turn conversations, making it suitable for this project. It offers high customization, privacy control, and adaptability, allowing Costco to train the chatbot specifically for customer sentiment analysis. A notable benefit of Rasa is its on-premise deployment capability, allowing enterprises to maintain control over sensitive customer data and comply with internal security protocols. Additionally, Rasa’s customizable natural language understanding (NLU) pipeline facilitates the integration of specific components, such as sentiment analysis, enhancing the assistant’s ability to interpret nuanced customer feedback. This adaptability is crucial for tailoring the chatbot to address the unique requirements of analyzing Costco’s customer feedback.[14]

4.3.2 Dialogflow

Developed by Google, Dialogflow is known for its powerful natural language understanding (NLU) capabilities. The platform's ease of use and ability to integrate with multiple platforms (e.g., web apps, messaging apps, and voice assistants) are key strengths. For instance, KLM Airlines utilizes Dialogflow to handle customer interactions, integrating it across multiple communication platforms. Dialogflow’s ability to process sentiment data in real-time and integrate with various platforms will enhance the chatbot, allowing stakeholders to engage with actionable customer insights effectively.[15]

Criteria for Selecting the Best Chatbot Framework

To determine the most suitable framework for this project, the following factors will be evaluated:

1. **Integration with Sentiment Analysis Models** – Ability to incorporate real-time sentiment scores and topic analysis from Costco’s customer feedback data.
2. **Customization & Scalability** – Flexibility to adapt chatbot responses based on specific product and service categories.
3. **User Experience & Accessibility** – Ease of interaction for non-technical stakeholders, ensuring intuitive and meaningful insights.
4. **Data Privacy & Security** – Ensuring compliance with Costco’s data handling policies while protecting customer data.
5. **Cost & Integration Feasibility** – Evaluating based on integration cost with the model.

By selecting the most suitable chatbot framework, this project will enable efficient analysis of customer feedback, provide stakeholders with real-time sentiment insights, and support data-driven decision-making to improve products and services.

4.4 Visualization and Reporting

A visualization dashboard will be created using Power BI. The dashboard should present sentiment trends and customer engagement metrics.

A knowledge graph will also be created which is a structured representation of relationships between entities, helping to organize and connect data in a meaningful way. In this project, the knowledge graph will visualize links between products, customer sentiment, and key discussion themes. By mapping these relationships, stakeholders can quickly identify trends, emerging concerns, and product preferences. This approach enables a more intuitive and insightful analysis of customer feedback, facilitating better decision-making for product improvements and service enhancements.

4.4.1 Power BI

Power BI will be used in the project because it offers an easy-to-use interface, allowing stakeholders to view and understand data without technical expertise. It integrates well with various data sources, ensuring smooth data flow. The tool provides customizable dashboards that help present sentiment trends, customer engagement metrics, and product relationships effectively.[16] Costco also utilizes Power BI for internal analytics, leveraging its capabilities to improve decision-making and enhance operational efficiency.

4.5 Deliverables

The final deliverables for the project will include the following key components:

1. **Cleaned Dataset:** A cleaned and organized dataset from public social media platforms relevant to Costco.

2. **Sentiment Analysis Report:** A detailed sentiment analysis report categorizing customer feedback on Costco products and services.
3. **Recommendations and Knowledge Graph:** Suggestions for new products/services based on customer sentiment, with a knowledge graph that visualizes product trends, key attribution tags, and scoring.
4. **AI-powered Chatbot Prototype:** A working prototype of an AI-powered chatbot designed to provide interactive insights from customer feedback.
5. **Code Notebook:** A well-documented code notebook containing the complete solution, ensuring clarity and ease of use.
6. **Reusable DS Model/Framework:** A reusable data science model or framework that can be applied to future projects, providing flexibility and efficiency.

4.6 Stretch Goals (Advanced Features)

If time permits, additional features such as web real-time sentiment monitoring, advanced topic modeling, and chatbot integration with live social media streams may be explored.

5. Proposed Solution

5.1 Approach to Solve the Problem

This project suggests a multi-step AI-driven solution that automates data collection, processing, sentiment classification, summarization, and visualization in order to efficiently analyze customer sentiment from open social media platforms like Reddit, Instagram etc. In order to improve decision-making about Costco's goods and services, the objective is to extract valuable insights from unstructured customer feedback by utilizing Natural Language Processing (NLP) techniques.

In order to ensure a steady and scalable flow of pertinent conversations, the solution starts with data collection from open social media platforms using web scraping and API integration. To improve text quality and eliminate extra information or noise, preprocessing techniques such as tokenization, stopword removal, lemmatization, and spell correction are applied to the raw text data. To make sure that only real customer conversations are examined, spam detection techniques like rule-based filtering and machine-learning classifiers will also be used.

Topic modeling and sentiment analysis form the basis of the solution. For sentiment classification, transformer-based models such as BERT will be employed, offering context-aware sentiment labeling (positive, negative, or neutral) with greater accuracy than conventional techniques. This helps in identifying the primary factors influencing consumer opinions. In order to identify hidden themes in customer conversations and classify feedback into relevant areas like product quality, pricing issues, or service experiences, this project uses Latent Dirichlet Allocation (LDA) for topic modeling.

Both extractive and abstractive summaries of customer sentiment will be produced by an AI-powered summarization module to make the results easier to understand. While abstractive summarization rephrases information in a more understandable format, extractive summarization highlights important phrases straight from the text. This makes it possible to rapidly understand general sentiment patterns without having to go through a lot of feedback by hand.

For stakeholders, an interactive chatbot built with Dialogflow or Rasa will act as a query-based interface. This chatbot will enhance the system's usability for business decision-making by enabling stakeholders to retrieve sentiment trends, product-related insights, and new customer concerns in real time. Additionally, all processed data will be visualized using Power BI, presenting sentiment trends, engagement metrics, and a knowledge graph that maps relationships between products and customer concerns.

Traditional techniques like surveys and manual reviews lack the scalability, real-time monitoring, and automation that the suggested solution offers. The framework guarantees accuracy, efficiency, and actionable insights by combining machine learning based classification, AI-driven summarization, and sophisticated NLP models. This strategy also establishes the groundwork for upcoming enhancements, like extending to more social media channels, incorporating real-time monitoring, and honing the chatbot for more in-depth sentiment analysis exchanges.

The performance of the solution will be evaluated based on multiple metrics. Sentiment classification accuracy will be measured using precision, recall, and F1-score, ensuring that models correctly distinguish between positive, negative, and neutral sentiments. The effectiveness of AI-generated summaries will be assessed using ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores, which compare AI-generated summaries to manually written reference summaries.[17] Human-written references will be obtained through manually labeled sentiment data, where a subset of comments will be reviewed by human annotators to establish a benchmark for evaluating AI-generated outputs. Additionally, the usability of the chatbot will be measured using engagement metrics, such as the frequency of interactions, response accuracy, and user satisfaction ratings. The goal is to achieve a decent F1-score ranging between 70-80% for sentiment classification, generate summaries that align closely with human interpretations, and ensure that stakeholders can make data-driven product and service decisions based on the provided insights.

5.2 Management Perspective

From a management perspective, the project's feasibility is supported by a structured timeline. The data collection and preprocessing phase is expected to take 4-6 weeks, followed by sentiment analysis and modeling over the next 6-8 weeks. Chatbot development and integration will be completed in 6-8 weeks, with a final testing and deployment phase spanning approximately 4 weeks.

To facilitate efficient and large-scale data collection, automated web scraping and API integration tools will be utilized. Two primary tools considered for this task are Apify and PhantomBuster. Apify is a cloud-based web scraping and automation platform that enables structured data extraction from websites and social media platforms. It will be used to scrape relevant customer feedback, product discussions, and sentiment-related content from sources such as Reddit, Twitter (X), and Instagram. PhantomBuster is a no-code automation tool designed to extract public data from social media platforms efficiently. It allows for automated profile scraping, comment extraction, and keyword-based filtering, reducing manual effort while ensuring continuous data collection. The project will require ongoing API subscriptions for these tools, with an estimated monthly cost of $49 for Apify and $69 for PhantomBuster.[18][19] These costs must be factored into the project budget to ensure seamless data retrieval and integration. Additionally, computational costs will primarily involve cloud-based infrastructure for AI model training and chatbot deployment.

To ensure successful implementation, access to social media APIs (while needed) and insights from Costco's marketing and merchandising teams will be essential. Stakeholder involvement will be critical for refining the chatbot's usability and ensuring that insights align with Costco's business needs. This solution will have significant impacts on various stakeholders, including decision-makers who will gain real-time insights into customer sentiment, merchandising teams that can make informed stocking and marketing decisions, and customers who will benefit from Costco's responsiveness to market trends. Ethical considerations will also be taken into account to ensure transparency in AI-generated recommendations while complying with data privacy regulations.

5.3 Team Capabilities

The success of this project relies on our diverse team expertise in data science, NLP, data engineering, and software development. Agile project management will ensure iterative improvements, delivering actionable market insights for Costco.

Each team member brings unique strengths, expertise in machine learning, data visualization, and software development; strong foundations in data science, clear communication, and collaboration; proficiency in data analysis, model building, and workflow optimization; skills in machine learning and handling complex datasets; and problem-solving, effective communication, and teamwork. Areas for growth include chatbot frameworks, concise idea summarization, confidence in presentations, navigating complex datasets, and time management. Together, we will execute this project efficiently and drive data-driven decision-making.

6. Deliverables and Timeline

6.1 Phase 1: Planning and Setup (Weeks 1-2)

- Tasks:
    - Define scope, deliverables, and timeline with Costco stakeholders (1 week)
    - Confirm team roles and responsibilities (0.5 weeks)
    - Identify data sources (Reddit, Instagram, X) and set up Agile tools (Jira) (0.5 weeks)
    - Finalize technical environments, tools and libraries (1 week)
    - Team Allocation: All members involved
- Dependency: Must be completed before data collection begins

6.2 Phase 2: Data Collection and Preprocessing (Weeks 3-7)

- Tasks:
    - Extract data using APIs/web scraping tools (2 weeks)
    - Store raw data in a centralized database (trying find resource)
    - Clean data (remove duplicates, noise) (1 week)
    - Segment data by product and service mentions(0.5 weeks)
    - Conduct quality assurance checks (0.5 weeks)
- Team Allocation: 2-3 members focused on data extraction; 2 members on data cleaning
- Dependency: Planning and setup must be completed before extraction begins

6.3 Phase 3: Sentiment Analysis and Knowledge Graph Development (Weeks 7-12)

- Tasks:
    - Perform exploratory data analysis (EDA) (1 week)
    - Implement sentiment analysis using NLP models (VADER, TextBlob, BERT) (2 weeks)
    - Optimize sentiment models for accuracy (1 week)
    - Develop a knowledge graph linking products, sentiment, and trends (2 weeks)
    - Use propensity scoring to rank products or services based on customer likelihood to engage (1 week)

- Team Allocation: 2-3 members on sentiment analysis, 2 members on knowledge graph
- Dependency: Requires cleaned and preprocessed data from Phase 2

6.4 Phase 4: Dashboard and Visualization Development (Weeks 13-15)

- Tasks:
  - Develop dashboard interface (Tableau, Dash, Plotly) (1.5 weeks)
  - Integrate sentiment analysis and knowledge graph into dashboard (1 week)
  - Take feedback from stakeholders through testing and make improvements based on their input (0.5 weeks)
- Team Allocation: 3 members on dashboard development, 2 on integration/testing
- Dependency: Requires sentiment analysis and knowledge graph from Phase 3

6.5 Phase 5: Summarization Chatbot Development (Weeks 16-17)

- Tasks:
  - Develop chatbot using Rasa, Dialogflow, or OpenAI APIs (1 week)
  - Train chatbot to summarize insights dynamically (0.5 weeks)
  - Integrate chatbot with dashboard (0.5 weeks)
  - Internal testing to improve the responses (1 week)
- Team Allocation: 3 members on chatbot development, 2 on integration/testing
- Dependency: Requires completed dashboard and sentiment analysis

6.6 Phase 6: Final Deliverables and Presentation (Week 18)

- Tasks:
  - Compile cleaned dataset, analysis report, and knowledge graph (0.5 weeks)
  - Finalize dashboard and chatbot (0.5 weeks)
  - Prepare project day presentation and documentation (1 week)
  - Present findings to Costco and take feedback (presentation day)
- Team Allocation: All members involved
- Dependency: Requires completion of all previous phases

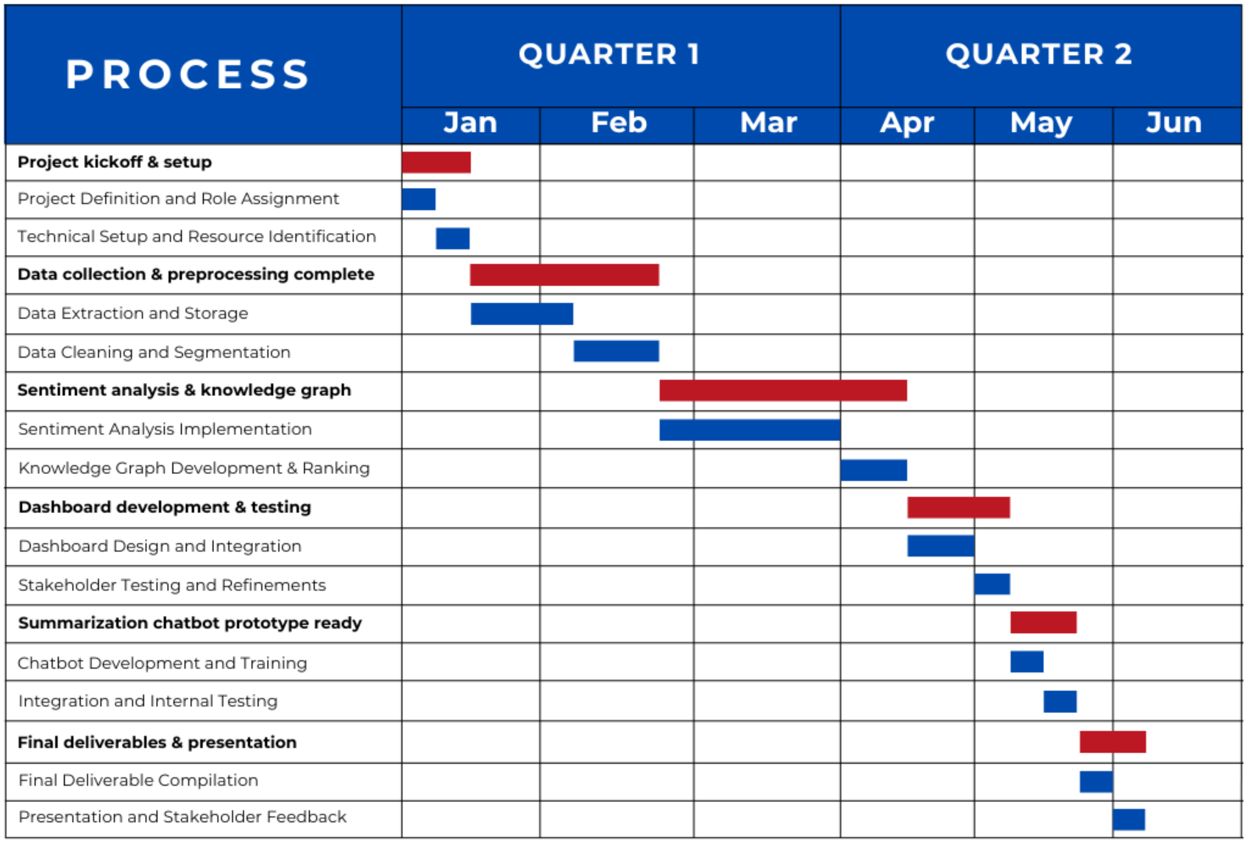
6.7 Stretch Goals (Optional, Time Permitting)

- Real-time sentiment monitoring and trend analysis
- Enhanced chatbot with live social media integration
- Scalable web application for broader usability

Milestone	**Week**	**Dependencies**
Project kickoff & setup	2	None
Data collection & preprocessing complete	6	Planning & setup
Sentiment analysis & knowledge graph ready	12	Data collection & preprocessing
Dashboard development & testing complete	15	Sentiment analysis & knowledge graph
Summarization chatbot prototype ready	17	Dashboard development
Final deliverables & stakeholder presentation	18	All previous phases

Table 2. Milestone table

Gantt Chart



PROCESS	QUARTER 1			QUARTER 2		
	Jan	Feb	Mar	Apr	May	Jun
Project kickoff & setup	🟥					
Project Definition and Role Assignment	🟦					
Technical Setup and Resource Identification	🟦					
Data collection & preprocessing complete	🟥	🟥				
Data Extraction and Storage	🟦	🟦				
Data Cleaning and Segmentation		🟦				
Sentiment analysis & knowledge graph		🟥	🟥	🟥		
Sentiment Analysis Implementation		🟦	🟦			
Knowledge Graph Development & Ranking				🟦		
Dashboard development & testing				🟥	🟥	
Dashboard Design and Integration				🟦		
Stakeholder Testing and Refinements					🟦	
Summarization chatbot prototype ready					🟥	
Chatbot Development and Training					🟦	
Integration and Internal Testing					🟦	
Final deliverables & presentation					🟥	🟥
Final Deliverable Compilation					🟦	
Presentation and Stakeholder Feedback						🟦

Figure 1. Timeline Gantt Chart

7. Conclusion

Conventional methods of consumer sentiment measurement, like surveys and focus groups, are usually biased, ineffective, and unable to provide timely insights. AI-driven sentiment analysis, however, offers a scalable, data-driven means of understanding customer sentiment. Large amounts of social media customer feedback are processed and labeled using Natural Language Processing (NLP) methods, such as VADER for contextual sentiment evaluation of short-text, and BERT for considering context from both directions. The system facilitates the effective extraction of actionable information by stakeholders through the integration of AI-based summarization and conversational chatbot functionality. Also, a knowledge graph and visualization dashboard offer a systematic means of viewing sentiment trends, enabling businesses to base their decisions on actual client experience.

Even though this method greatly improves decision-making, issues like noise filtering, computational costs, and data privacy still need to be resolved. AI-driven techniques offer more flexibility and accuracy than conventional sentiment analysis models, but their efficacy is contingent upon appropriate data preprocessing and contextual analysis. Future developments include real-time sentiment analysis, sophisticated topic modeling to uncover new trends, and the creation of an interactive web application for easy access to insights. With these innovations in place, this project hopes to equip businesses with the tools they need to increase customer engagement, maximize product strategies, and keep a competitive edge in the rapidly changing retail environment.

8. Appendix

1. **Statista.** (2024). *Number of Costco card holders worldwide from 2014 to 2024*. Retrieved February 9, 2025, from, https://www.statista.com/statistics/718406/costco-number-memberships/
2. **IBM**. (n.d.). *What is Sentiment analysis?* Retrieved February 5, 2025, from, https://www.ibm.com/think/topics/sentiment-analysis
3. **Dorian, L.** (2021, March 5th). *Scraping Medium with Python & BeautifulSoup*. Retrieved February 9, 2025, from, https://dorianlazar.medium.com/scraping-medium-with-python-beautiful-soup-3314f898bbf5
4. **GitHub.** (n.d.). *Public APIs [Repository]*. Retrieved February 12, 2025, from, https://github.com/public-apis/public-apis
5. **GeeksforGeeks.** (n.d.**)**. *Natural language processing (NLP) – Overview*. GeeksforGeeks. Retrieved February 12, 2025, from, https://www.geeksforgeeks.org/natural-language-processing-overview/
6. **Zonka Feedback.** (n.d.). *Sentiment analysis of customer feedback: A complete guide*. Retrieved February 12, 2025, from, https://www.zonkafeedback.com/blog/sentiment-analysis-customer-feedback
7. **Wikipedia contributors.** (n.d.). *General Data Protection Regulation*. Wikipedia, The Free Encyclopedia. Retrieved February 18, 2025, from, https://en.wikipedia.org/wiki/General_Data_Protection_Regulation
8. **European Data Protection Board**. (n.d.). *European Data Protection Board Guidelines*. Retrieved February 18, 2025, from, https://www.edpb.europa.eu/our-work-tools/our-documents/publication-type/guidelines_en
9. **Slavanya, R.** (n.d.). *VADER: A comprehensive guide to sentiment analysis in Python*. Medium. Retrieved March 5, 2025, from https://medium.com/@rslavanyageetha/vader-a-comprehensive-guide-to-sentiment-analysis-in-python-c4f1868b0d2e
10. **IBM**. (n.d.). *Latent Dirichlet Allocation (LDA)*. IBM. Retrieved March 5, 2025, from https://www.ibm.com/think/topics/latent-dirichlet-allocation
11. **V. Rakesh Reddy , S. Rakesh.** (2024). *Next-generation email security: Enhancing spam filtering with NLP algorithm*. International Journal of Research Publication and Reviews, 5(11), 7005–7009.https://ijrpr.com/uploads/V5ISSUE11/IJRPR35629.pdf
12. **DigitalOcean.** (n.d.). Extractive and Abstractive Summarization Techniques. Retrieved February 9, 2025, from https://www.digitalocean.com/community/tutorials/extractive-and-abstractive-summarization-techniques
13. **TechTarget.** (n.d.). *BERT language model*. TechTarget. Retrieved March 5, 2025, from https://www.techtarget.com/searchenterpriseai/definition/BERT-language-model
14. **Rasa.** (n.d.). *How to choose an enterprise chatbot platform*. Rasa. Retrieved March 5, 2025, from https://rasa.com/blog/how-to-choose-enterprise-chatbot-platform/
15. **Randelli, G.** (2024, April 15). *Create multimodal conversational experiences with Google Cloud Dialogflow CX and Gemini Vision*. Medium. Retrieved February 23, 2025,

from https://medium.com/@gabrielerandelli/create-multimodal-conversational-experiences-with-google-cloud-dialogflow-cx-and-gemini-vision-8d39035b985d

16. **Microsoft.** (n.d.). *Power BI overview.* Microsoft Learn. Retrieved February 22, 2025, from https://learn.microsoft.com/en-us/power-bi/fundamentals/power-bi-overview
17. **Bronsdon, C.** (2024, December 4). *Understanding ROUGE in AI: What it is and how it works.* Galileo AI. Retrieved February 23, 2025, from https://www.galileo.ai/blog/rouge-ai
18. **PhantomBuster.** (n.d.). *Billing & pricing.* PhantomBuster. Retrieved February 23, 2025, from https://phantombuster.com/3164602463919830/billing
19. **Apify.** (n.d.). *Pricing.* Apify. Retrieved February 23, 2025, from https://apify.com/pricing