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. Background

Costco Wholesale is the third-largest retailer in the world known for providing beneficial memberships to its customers. As of 2024, it has about 136.8 million happy customers. High-quality and in-demand products are provided by Costco at very reasonable prices.[1]

Traditionally, companies look over surveys, reviews on their own websites, and focus groups to understand their 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, sources like social media, and direct company websites have become major platforms for customer feedback. Customers share their opinions about Costco's products through forms like reviews, product ratings, survey responses, and forum discussions.[2] Praise for a new product, a request for improvements, or dissatisfaction with a service are shared in the form of comments. Sources for these comments include Costco's website, Instagram posts, discussion threads on Reddit, and Twitter posts managed by Costco. These comments capture customer sentiment. However, extracting meaningful information from this kind of data is a challenge.

Thus by using sentimental analysis and AI-powered summarization, large amounts of data from social media and other sources can be quickly analyzed to spot trends and understand customer preferences. Sentiment analysis is the process of analyzing large volumes of text to determine whether it expresses a positive, negative, or neutral sentiment. These techniques are effective as they can provide real-time insights, are scalable, and are accurate.

This project aims to combine sentiment analysis, AI-driven summarization, a web-based tool, and an interactive chatbot to help Costco with clearer, more meaningful insights. By converting data from sources like social media and direct company websites into actionable recommendations, Costco can make data-driven decisions on new and emerging products, and services, thereby improving member satisfaction and business growth.

2. Problem Statement

Costco customers share their experiences, preferences, and concerns on social media sites like Instagram, Twitter (X), and Reddit. These discussions offer insightful information that, with careful analysis, can support strategic choices. But it can be very difficult to glean useful information from this enormous volume of unstructured text. By putting in place a data-driven solution to assess consumer sentiment, spot new trends, and strengthen evidence-based business decisions, this project seeks to address this challenge.

Through sentiment analysis and natural language processing (NLP), the project will turn unstructured social media data into structured insights. Real-time analysis and ongoing data collection are guaranteed by an automated data pipeline. The final deliverables will include a sentiment analysis report, an interactive knowledge graph, an AI-enabled chatbot, and a web-based tool for delivering real-time actionable insights. By using these solutions, Costco will be able to better predict consumer demands, enhance its product line, and keep a competitive edge in the retail industry.

3. Requirements

3.1 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 around 100,000 rows of data gathered from different social media platforms. 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.

3.1.1 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. 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 also be segmented based on product categories and services. Product-related feedback will be separated from service-related feedback, allowing for a more detailed and focused sentiment analysis.[6]

As part of the data collection process, strict adherence to data privacy regulations will be maintained. Personal or sensitive data will not be collected, ensuring compliance with privacy laws. Ethical web scraping practices will be followed to avoid violating platform terms of service.

3.2 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 to improve accuracy. These

specific models were chosen based on their well-established effectiveness in analyzing sentiment from social media data.[7]

3.2.1 VADER (Valence Aware Dictionary and sentiment Reasoner)

This tool is especially suited for analyzing social media content because it is designed to handle the short, informal, and often slang heavy language used in platforms like Twitter, Reddit, and Instagram. It is a rule based model and is efficient in detecting sentiment from text with emoticons, punctuation, and capitalization, making it highly suitable for social media posts.[8]

3.2.2 BERT (Bidirectional Encoder Representations from Transformers)

BERT is a state of the art, transformer based model that excels in understanding the context of words in a sentence. Unlike traditional models, BERT considers both the left and right context of a word, which is crucial for detecting sentiment accurately in more complex customer feedback. Its higher accuracy for sentiment classification, especially in nuanced and longer texts, makes it the preferred choice for more detailed and complex sentiment analysis.[9]

The analysis will also focus on identifying sentiment trends over time and across different product categories, enabling the detection of patterns and shifts in customer sentiment. This will provide a deeper understanding of how Costco's products and services are being perceived.

A propensity scoring system will be developed to rank products and services based on their sentiment impact and engagement levels. This system will assess how customer sentiment influences the popularity and engagement of each product, helping to prioritize areas for improvement or further marketing efforts.[10]

3.3 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 choice of these techniques is based on their ability to handle the complex and varied nature of customer feedback, ensuring that both detailed insights and high-level summaries can be extracted effectively.[11]

A chatbot will also be developed using frameworks like Rasa and Dialogflow. The chatbot will make it easier for stakeholders to quickly access and understand key information about customer sentiment.

3.3.1 Why These Tools?

The following tools will be considered based on specific criteria to select the best solution:

Rasa

Rasa is an open source framework designed for building highly customizable chatbots. It is well suited for complex, multi-turn conversations and can be tailored to provide a personalized user experience. The ability to integrate with various data sources and customize the chatbot's behavior makes it a strong contender for this project.[12]

Dialogflow

This is a popular framework developed by Google, known for its ease of use and powerful natural language understanding (NLU) capabilities. It allows easy integration with multiple platforms, including web apps, messaging apps, and voice assistants. Dialogflow's ability to handle a wide range of languages and its pre-built models for sentiment analysis and conversation flow are key reasons for considering it.

The final decision on the tool will depend on the following factors like how easily it integrates with the existing data pipeline and visualization tools, the level of customization needed for specific project needs like personalized sentiment reports, its scalability to handle growing data and more complex interactions, cost-effectiveness, especially when comparing open-source options like Rasa with commercial tools like Dialogflow or OpenAI APIs, and the user experience, ensuring that stakeholders can easily use the chatbot and get useful insights without needing technical knowledge.[13]

This approach will ensure that the best possible chatbot framework is chosen based on the specific requirements and limitations of the project.

3.4 Visualization and Reporting

A visualization dashboard will be created using Power BI. The dashboard should present sentiment trends, customer engagement metrics, and a knowledge graph showing relationships between products and customer expectations.

Power BI

Power BI will be used in our 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.

3.5 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.

3.6 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.

4. Schedule

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

- Milestone: Kickoff and Project Alignment
- 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

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

- Milestone: Completion of Data Collection and Cleaning

- 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

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

- Milestone: Delivery of Sentiment Analysis and Knowledge Graph
- 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

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

- Milestone: Completion of Interactive Dashboard and Initial User Testing
- 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

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

- Milestone: Summarization Chatbot Prototype Ready
- 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

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

- Milestone: Final Project Submission and Stakeholder Presentation
- 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

4.7 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 1. Project Timeline and Milestones.

5. Appendix

1. **Statista.** (2024). Number of Costco card holders worldwide from 2014 to 2024. Statista. Retrieved February 9, 2025, from https://www.statista.com/statistics/718406/costco-number-memberships/
2. **Costco.** (n.d.). Costco. Wikipedia. Retrieved February 8, 2025, from https://en.wikipedia.org/wiki/Costco
3. **Dorian, L.** (2021, March 5). Scraping Medium with Python & BeautifulSoup. Medium. Retrieved February 9, 2025, from https://dorianlazar.medium.com/scraping-medium-with-python-beautiful-soup-3314f898bbf5
4. **Hugging Face.** (n.d.). Formatting inference API call for LLaMA-2. Retrieved February 9, 2025, from https://discuss.huggingface.co/t/formatting-inference-api-call-for-llama-2/54901
5. **DeepLearning.AI.** (n.d.). Natural language processing. DeepLearning.AI. Retrieved February 9, 2025, from https://www.deeplearning.ai/resources/natural-language-processing/
6. **Unwrangle.** (n.d.). Costco Product Reviews API Documentation. Retrieved February 9, 2025, from https://docs.unwrangle.com/costco-product-reviews-api/#__tabbed_1_2
7. **IBM**. (n.d.). Sentiment analysis. IBM Think. Retrieved February 5, 2025, from https://www.ibm.com/think/topics/sentiment-analysis
8. **Swayanshu.** (n.d.). VADER (Valence Aware Dictionary and Sentiment Reasoner) – Sentiment Analysis. Medium. Retrieved February 9, 2025, from https://swayanshu.medium.com/vader-valence-aware-dictionary-and-sentiment-reasoner-sentiment-analysis-28251536698
9. **NVIDIA.** (n.d.). BERT (Bidirectional Encoder Representations from Transformers). Retrieved February 9, 2025, from https://www.nvidia.com/en-us/glossary/bert/
10. **DIME Wiki.** (n.d.). Propensity score matching. The World Bank. Retrieved February 9, 2025, from https://dimewiki.worldbank.org/Propensity_Score_Matching
11. **DigitalOcean.** (n.d.). Extractive and Abstractive Summarization Techniques. Retrieved February 9, 2025, from https://www.digitalocean.com/community/tutorials/extractive-and-abstractive-summarization-techniques
12. **Botpress.** (n.d.). Open-Source Chatbots. Retrieved February 9, 2025, from https://botpress.com/blog/open-source-chatbots
13. **Google Cloud.** (n.d.). Conversational Agents. Retrieved February 9, 2025, from https://cloud.google.com/products/conversational-agents