CSCE 5290: Natural Language Processing

Group-15 Project Proposal

Title: NLP-Driven Query Answering and Summarization for Amazon Reviews

GitHub Repository: https://github.com/meghanasadhu/amazonreview

1. Motivation

The vast amount of data generated from online reviews can be overwhelming for buyers, making it difficult for them to get current information quickly. Although the number of stars helps some people, a deeper understanding of a product's pros and cons is hidden within the texts. This project seeks to solve this problem by creating a natural language processing (NLP) model that retrieves relevant answers to user's questions from product reviews and generates useful summaries.

2. Significance

In this rapidly growing world with technology, a solution that extracts information from huge amounts of reviews holds great value. This helps the user to make more useful decisions without having to read through reviews endlessly.

3. Objectives

- Develop a model to efficiently extract relevant segments that answer user's query.
- Implement a summarization mechanism that provides accurate summaries.
- Evaluate the performance of the model in terms of accuracy and relevance.

4. Features

- Question-Answering System: The ability to extract relevant information for user queries using techniques like semantic similarity.
- Automatic Summarization: The system generates a summary related to the query.

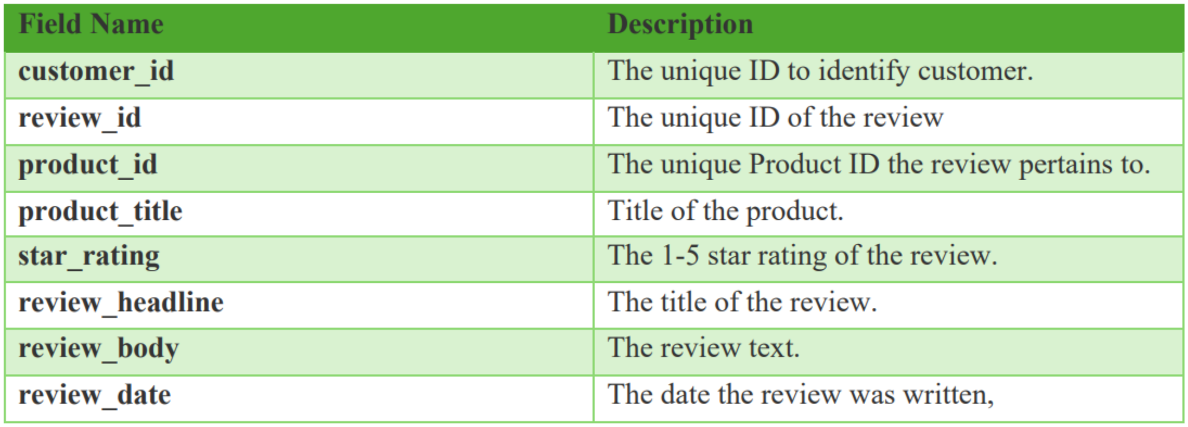
5. Dataset

- Source: The project will use the Amazon US Customer Reviews on Digital video games Dataset from Kaggle which is publicly available.
  Link: https://www.kaggle.com/datasets/cynthiarempel/amazon-us-customer-reviews-dataset?select=amazon_reviews_us_Digital_Video_Games_v1_00.tsv
- Type: Structured data fields like customer_id, review_id, product_id, product_title, star_rating, review_headline, review_body, review_date.
- Size: It has approx. 145,431 rows and 8 columns are used.

- Preprocessing: Text cleaning to remove unnecessary symbols, tokenization to split the text into tokens and stop words removal to reduce noise.

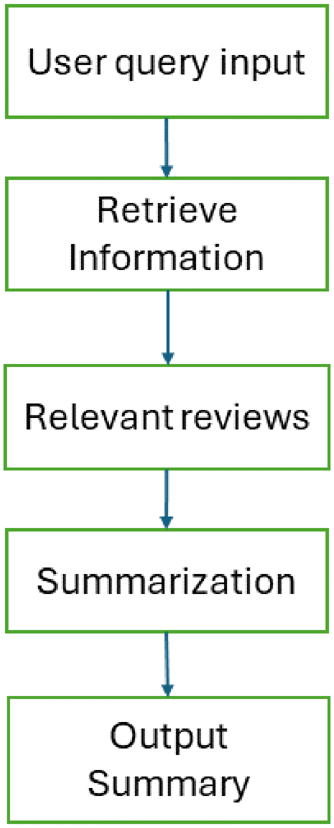
6. Visualization

Table: Dataset Overview



Field Name	Description
customer_id	The unique ID to identify customer.
review_id	The unique ID of the review
product_id	The unique Product ID the review pertains to.
product_title	Title of the product.
star_rating	The 1-5 star rating of the review.
review_headline	The title of the review.
review_body	The review text.
review_date	The date the review was written,

Workflow diagram:



This shows the workflow of the system step by step.

1. The user inputs a specific question regarding a product.
2. The system retrieves relevant reviews that match with the query.
3. The relevant reviews are extracted and passed to the next step.
4. The retrieved reviews are processed to generate a summary.
5. The system provides the user with a short summary of the reviews.