CSCE 5290: Natural Language Processing

Project Proposal

Title: Product Review Summarization

Group No: 6

Team Members:

SNo	Name	UNT ID	Email
1	Koushik Vemuri	11590466	KoushikVemuri@my.unt.edu
2	Bhargav Ram Pushadapu	11647795	BhargavRamPushadapu@my.unt.edu
3	Rikhita Koganti	11641222	Rikhitakoganti@my.unt.edu
4	Padmaja Kodati	11647780	PadmajaKodati@my.unt.edu

Github Repository: https://github.com/KoushikVemuri/Product-Review-Summarization

Motivation

The primary goal of product customer review summarizing is to effectively extract important insights from a huge number of evaluations. The effort attempts to solve the problem of corporations and consumers having too much information. Customers find it challenging to sort through the vast number of product reviews available due to the growth of online shopping platforms and to make well-informed judgments. Similar to this, companies may find it difficult to extract useful information from the overwhelming amount of input they get. This initiative aims to give businesses useful input to enhance their products and services, while also streamlining the decision-making process for consumers by consolidating reviews into clear and insightful summaries.

Significance

A critical project that aims to filter several reviews into important insights for the benefit of businesses and customers alike is the summary of customer reviews for products. This solution tackles the issue of excessive information when purchasing online by simplifying the process of making decisions for customers. It facilitates educated purchasing decisions and saves time and effort by providing brief descriptions of product experiences. Furthermore, this project offers businesses insightful feedback to better understand client mood and improve product offerings. In the end, analyzing condensed insights increases customer satisfaction and loyalty by pointing out trends, recurring problems, and opportunities for development. Additionally, within the wider framework of electronic commerce, it cultivates confidence, transparency, and effectiveness in virtual exchanges. Reaching the project's objective is crucial because it gives customers more power in the online market, encouraging wise decisions, better goods, and improved consumer experiences.

Objectives

- Our project is in the field of natural language processing(NLP). It involves the raw text from customer reviews and it is being analyzed , categorized and summarized.
- Our objective is to develop and train a summarization model which is capable of effectively summarizing product reviews which can capture key insights while maintaining coherence and relevance.
- We can determine the success of our objective by achieving high ROUGE scores indicating similarity between generated summaries and reference summaries.

- By the project’s finish we will explore multiple extractive and abstractive summarization techniques to identify the most suitable approach for the given dataset and task.
- We can evaluate these techniques by using our evaluation techniques and decide on the best performing.
- If time permits, we hope to Create a user interface that allows users to input product reviews and receive a summary of the review.
- We can easily evaluate the success of this objective based on the coherence and relevance of the generated summary by the model.

Features

Technical Characteristics and Deliverables

- Conducting Data Cleaning and Data Pre-processing to prepare the date for Feature extraction.
- We are considering both extractive and abstractive summarization techniques for developing our summarization model.
- Gensim library in python has extractive summarization techniques such asTextRank and Lex Rank or we can implement LTMS (Latent Topic Modeling Summarization) and RNN which are abstractive summarization techniques.
- Employing a model based on the chode summarization technique.
- We will evaluate our model’s performance using ROUGE scores and other evaluation metrics.
- We are hoping to add a User interface which enables users to input any review and summarize it.

Milestones

- Data Collection and Preprocessing
- Feature extraction
- Model Development and Training (Considering both extractive and abstractive approaches)
- Evaluation and Optimization
- Testing and Validation
- Possibly develop a UI to enable custom reviews to be summarized.

Distinctive features

- For building our model, we are going to choose the Summarization techniques which works the best with our chosen dataset
- We will evaluate our project using ROUGE scores and other evaluation techniques to show the effectiveness of our model
- We are hoping to add an UI to summarize custom reviews given by users in addition to summarizing the reviews in the dataset. We believe that these are the features which are distinct to our project.

Dataset

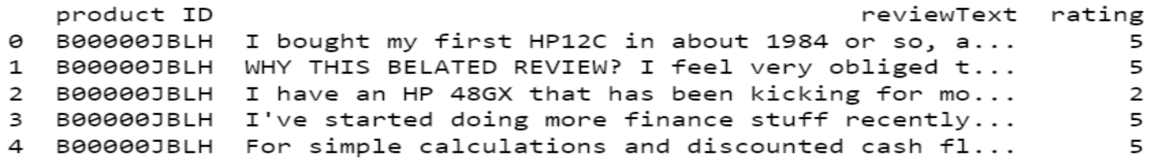
https://cseweb.ucsd.edu/~jmcauley/datasets/amazon/links.html

We are using the Amazon office product reviews dataset for our project.

Sample:

```
{
  "reviewerID": "A2SUAM1J3GNN3B",
  "asin": "0000013714",
  "reviewerName": "J. McDonald",
  "helpful": [2, 3],
  "reviewText": "I bought this for my husband who plays the piano.
He is having a wonderful time playing these old hymns.  The music  is
at times hard to read because we think the book was published for
singing from more than playing from.  Great purchase though!",
  "overall": 5.0,
  "summary": "Heavenly Highway Hymns",
  "unixReviewTime": 1252800000,
  "reviewTime": "09 13, 2009"
}
```

In our project we are only going to use the columns: product ID(asin), overall (rating) and reviewText and generate our own summaries. And we will drop the remaining columns which are not necessary for our project.



```
   product ID                                         reviewText  rating
0  B00000JBLH  I bought my first HP12C in about 1984 or so, a...       5
1  B00000JBLH  WHY THIS BELATED REVIEW? I feel very obliged t...       5
2  B00000JBLH  I have an HP 48GX that has been kicking for mo...       2
3  B00000JBLH  I've started doing more finance stuff recently...       5
4  B00000JBLH  For simple calculations and discounted cash fl...       5
```

The dataset has 53,258 user reviews of office products purchased by the customers.

Visualizations To Gain More Insight Into The Dataset

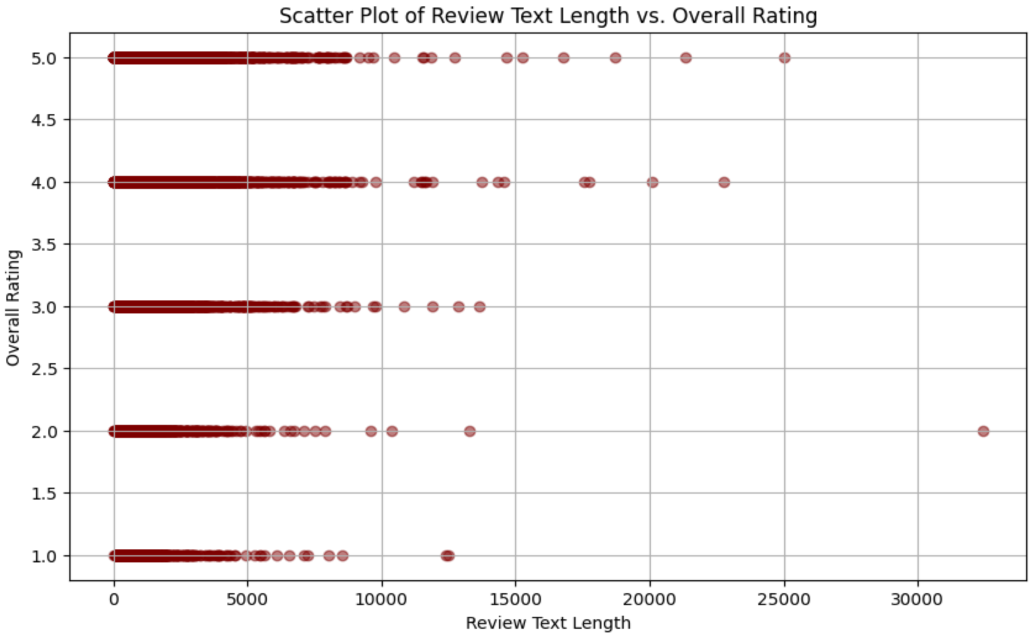


Fig 1: Length of Review Text vs Respective Rating

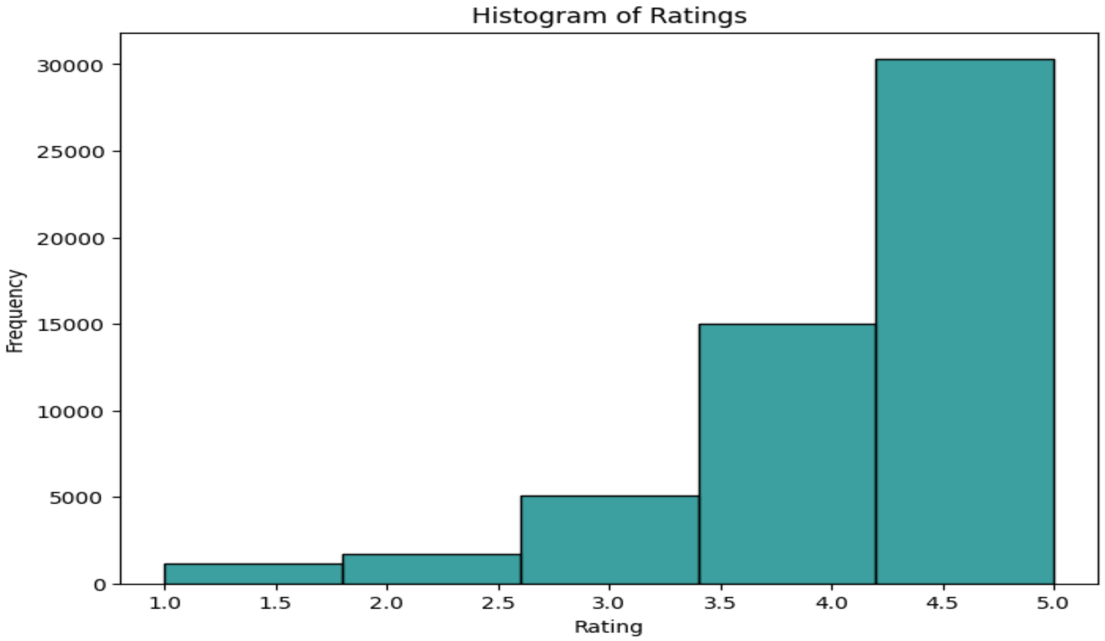


Fig 2: Distribution of Ratings In The Dataset

Workflow Diagram

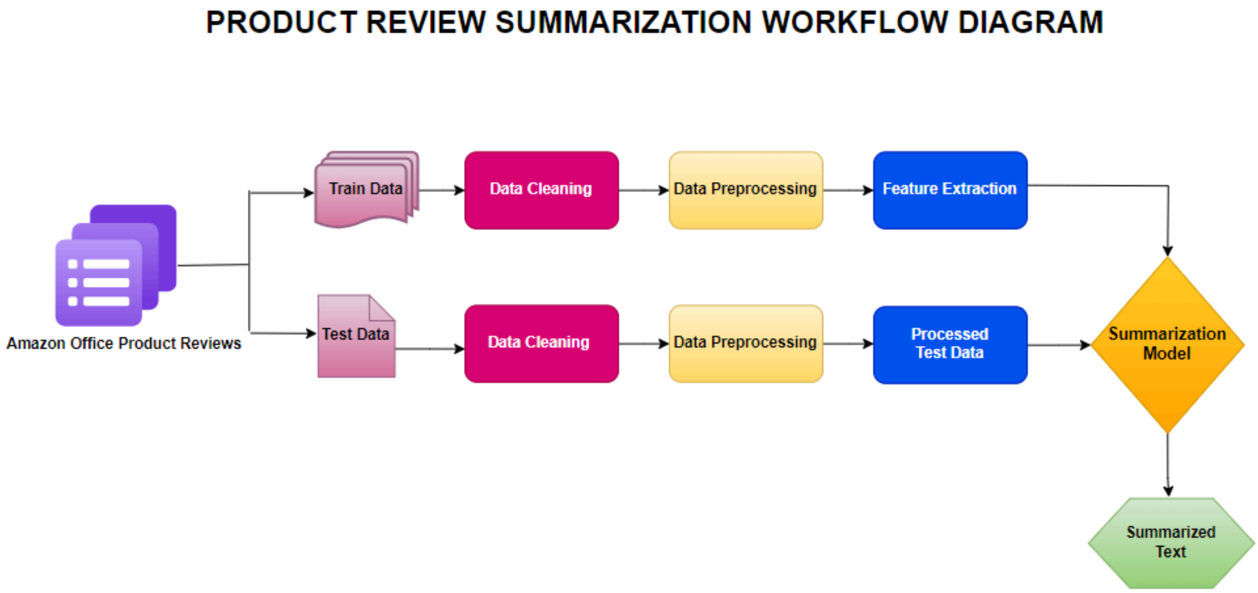


Fig 3: Product Review Summarization Workflow Diagram

Data Cleaning: In the data cleaning phase, we perform the following actions: converting all text to lowercase, removing unwanted characters, and eliminating stop words to prepare the data for Pre-processing.

Data Pre-processing: In the data pre-processing stage, we conduct tokenization and lemmatization, essential steps to transform raw text data into a structured format. Additionally, we may apply techniques such as stemming or removing HTML tags.

Feature Extraction: In feature extraction, we compute TF-IDF scores and use word embeddings to convert text into numerical data. We also explore methods like Word2Vec or GloVe embeddings to better capture word meanings.

Summarization model: In the summarization model, we use extractive or abstractive summarization techniques to generate summaries of the reviews. Gensim library in python has extractive summarization techniques such asTextRank and Lex Rank or we can implement LTMS (Latent Topic Modeling Summarization) and RNN which are abstractive summarization techniques

Summary

In summary, our project aims to summarize customer reviews using the Amazon Movie and TV dataset. We start by cleaning and organizing the data. Then, we process the text to make it easier to analyze. After that, we extract key features from the text. Finally, we use different techniques like TextRank, LexRank, LTMS, and RNN to generate summaries of the reviews.