NATURAL LANGUAGE PROCESSING

Automated Content Creation and Personalization for E-commerce Product Descriptions

Presented By:

Sancia Fernandes (A012) | Yash Dudeja (A013) | Sherin Ouseph (A017)

TABLE OF CONTENT

Contents

1. Abstract

This project presents an automated system for generating personalized e-commerce product descriptions and recommendations tailored to individual user preferences and product features. Leveraging large datasets from Amazon and Flipkart, the system employs natural language processing (NLP) and generative AI techniques to analyze customer sentiment, extract key product attributes, and create customized descriptions. The end-to-end workflow begins with data collection, followed by text preprocessing to clean and standardize the input data. NLP tasks include sentiment analysis on customer reviews and feature extraction to identify significant product characteristics such as size, color, and material.

Using insights from these analyses, a generative AI model produces dynamic and personalized product descriptions, while a recommendation system suggests similar or complementary items based on the user's preferences and historical behavior. This solution is deployed on a web platform, allowing for real-time generation of product descriptions that adapt based on user input, thus enhancing user engagement and purchase likelihood. The project demonstrates the potential of combining NLP and generative AI for content automation, which can significantly improve user experience in e-commerce settings.

2. Introduction

2.1 Overview of the Project

In the highly competitive e-commerce landscape, delivering personalized and relevant content is essential to attracting and retaining customers. This project focuses on developing an automated system to generate personalized product descriptions and recommendations based on user behavior and product features. Utilizing datasets from Amazon and Flipkart, the system applies natural language processing (NLP) and generative AI to analyze customer reviews, extract key product attributes, and create tailored product descriptions. This approach enhances user engagement by presenting dynamically generated content that aligns with individual preferences, ultimately aiming to improve customer experience and increase sales.

2.2 Problem Statement

Manually creating personalized product descriptions and recommendations for a vast inventory is labor-intensive, time-consuming, and impractical for most e-commerce platforms. Generic descriptions often fail to engage customers, resulting in missed opportunities to connect with users based on their unique preferences. Current approaches to e-commerce content are largely static and lack the adaptability needed to cater to individual user profiles. There is a need for an automated solution that can analyze user behavior and product features to deliver personalized content, enhancing the relevance of product descriptions and recommendations to improve conversion rates.

2.3 Objective of the system

The primary objective of this system is to automate content generation for e-commerce platforms by creating personalized product descriptions and recommendations based on product attributes and user behavior. Specifically, the system aims to:

- Generate customized product descriptions that align with user preferences and sentiment analysis of product reviews.
- Provide personalized product recommendations by analyzing user input, review sentiment, and product features.
- Deploy the solution on a web platform where product descriptions and recommendations are generated in real-time, enhancing the overall user experience and likelihood of purchase.

2.4 Scope and Limitations

Scope: The system leverages large e-commerce datasets and applies NLP and generative AI to personalize content, covering processes such as data collection, text preprocessing, sentiment analysis, feature extraction, and recommendation generation. The solution is designed to integrate with e-commerce platforms, enabling real-time generation of personalized product descriptions and recommendations.

Limitations: The system's effectiveness depends heavily on the quality and quantity of the datasets used, as biases or insufficient data may impact the accuracy of sentiment analysis and feature extraction. Additionally, the AI models may require significant computational resources, potentially limiting real-time functionality on platforms with limited processing capabilities. The personalization relies primarily on product attributes and customer review sentiment, which may not capture the full range of individual user preferences. Lastly, any real-time deployment would need to address privacy concerns when analyzing user data for recommendation purposes.

3. Datasets

3.1 Amazon Product Reviews Dataset

The Amazon Product Reviews dataset provides a rich source of customer reviews and ratings across various product categories. This dataset includes valuable information such as review text, rating, product ID, and user information. By analyzing the reviews, we can extract customer sentiment and key insights into how users perceive different product features. These insights serve as input for generating personalized product descriptions, as they highlight common customer opinions, product strengths, and areas of concern. The sentiment analysis applied to this data also enables the system to categorize reviews into positive, neutral, or negative sentiments, which can be used to tailor descriptions in a way that resonates with potential buyers.

3.2 Flipkart Product Dataset

The Flipkart Product dataset contains detailed information on various products listed on Flipkart, including product descriptions, features, price, and other essential details. This dataset is used to identify key product attributes such as size, color, material, and other relevant features. By extracting these attributes, the system can create descriptions that emphasize specific qualities of each product, adding a layer of personalization based on typical customer preferences. This dataset complements the Amazon review data, providing structured product details that enable feature extraction, which, combined with sentiment analysis, enhances the overall personalization and relevance of the generated product descriptions.

4. Methodology

The methodology for automating content creation and personalization of e-commerce product descriptions follows a systematic process, involving data collection, preprocessing, natural language processing (NLP) tasks, generative AI tasks, and the development of a recommendation system. The workflow is broken down into several key stages:

4.1. Loading Datasets

The system begins by loading the product review datasets from Amazon and Flipkart. These datasets contain valuable product-related information in the form of reviews and descriptions. The Amazon dataset is loaded using the AmazonReviews.csv file, and the Flipkart dataset is loaded using the flipkart_com-ecommerce_sample.csv file. These datasets serve as the foundation for all subsequent data processing and analysis.

4.2. Data Cleaning

Data cleaning is performed to remove any noise and irrelevant information from the raw text. The clean_text function is implemented to:

- Remove HTML tags, ensuring that only the relevant content remains.
- Remove non-alphabetic characters, retaining only textual content.
- Convert the text to lowercase to standardize the format and prevent case-based discrepancies.
- Eliminate stop words, which are common words that don’t add significant meaning (e.g., "the", "and", "is"). This cleaned data is stored in new columns (cleaned_reviews for Amazon and cleaned_description for Flipkart) and saved for further analysis.

4.3. Sentiment Analysis

4.3.1. Sentiment Analysis Model

The sentiment of each review or product description is analyzed using the pre-trained distilbert-base-uncased-finetuned-sst-2-english model. This model is optimized for sentiment classification and is accessed through the Hugging Face pipeline. It analyzes whether the sentiment of a given review or description is positive or negative. The function get_sentiment processes each piece of text and classifies it accordingly. The results are stored in the sentiment column of the Amazon dataset.

4.3.2. Result Storage

The sentiment analysis results are saved into a new CSV file, sentiment_amazon_reviews.csv, which contains the original reviews along with their corresponding sentiment labels.

4.4. Feature Extraction

4.4.1. Product Feature Extraction

To enhance product recommendations, the system extracts key features from the Flipkart product descriptions. These features include:

- **Product Type:** The category or type of the product (e.g., shirt, laptop).
- **Material:** The material used to make the product (e.g., cotton, leather).
- **Size:** The size of the product (e.g., M, L).
- **Color:** The color of the product (e.g., red, blue).

The extract_features function employs regular expressions to identify these features and store them in a new column (features) in the dataset. This extracted information helps tailor recommendations based on specific product attributes.

4.5. Personalized Product Description Generation

4.5.1. Combining Sentiment and Features

To create personalized product descriptions, the system combines the sentiment and features extracted from each product's description. For instance, a description might include a positive sentiment and detailed features such as "material: cotton" and "color: red." This combination allows the system to generate product descriptions that align with the sentiment and key features of interest.

4.5.2. Saving Personalized Descriptions

The personalized product descriptions are stored in a new CSV file, flipkart_product_personalized_descriptions.csv. This file contains the product names alongside their newly generated descriptions, which integrate sentiment and product features for a more tailored user experience.

4.6. Feature Vectorization for Product Recommendation

4.6.1. Vectorizing Product Descriptions

The system uses the TfidfVectorizer from sklearn to convert the product descriptions into numerical vectors. The vectorization captures the importance of terms in the text based on their

frequency. The model uses an ngram_range=(1, 2) to extract both unigrams and bigrams from the product descriptions, allowing it to capture both individual words and short phrases. The max_features=10 parameter restricts the analysis to the top 10 most important features.

4.6.2. Recommendation System

Cosine similarity is calculated between the vectorized product descriptions and the user input. The function recommend_products uses the cosine similarity to identify the top 5 products most similar to the user's query. The system ranks products based on similarity scores and returns the closest matches to the user.

4.6.3. User Input and Recommendations

A user can input a product description or preference, and the system will generate a list of recommended products based on the similarity of descriptions. The recommendations include details such as the product name and extracted features (e.g., size, color, material) to help users make informed decisions.

4.7. Building the Web Application Using Streamlit

4.7.1. Streamlit Interface

To make the product recommendation system accessible, it is integrated into a web interface using Streamlit. The interface includes:

- **Product Information Display:** A section where users can select a product and view its personalized description and sentiment.
- **Personalized Recommendations:** A user input field where users can enter a query, and the system generates personalized recommendations based on the query's similarity to the product descriptions.

4.7.2. Streamlit Deployment

The app is deployed using ngrok, making it accessible through a public URL. This deployment allows users to interact with the recommendation system in real-time, enhancing the overall user experience.

5. Implementation Details

5.1. Technologies Used

5.1.1. Hugging Face Models

Hugging Face is a leading library in the field of Natural Language Processing (NLP), offering a variety of pre-trained models through its transformers library. These models are essential for a range of NLP tasks, including text generation, sentiment analysis, and text classification, all of which are central to our project. Specifically, the project leverages a combination of models such as BERT, DistilBERT, and GPT for various tasks.

- **BERT (Bidirectional Encoder Representations from Transformers):** BERT is a transformer-based model that excels at understanding the contextual relationships between words in a sentence by reading text bidirectionally (considering both the preceding and following words). This bidirectional approach enables BERT to capture deeper contextual meaning, which is essential for handling tasks such as text classification, question answering, and sentiment analysis with high accuracy.

  In this project, BERT is primarily used for sentiment analysis. By analyzing the sentiment of existing product descriptions, BERT classifies them into positive, negative, or neutral categories. This sentiment analysis is crucial for personalizing the generated product descriptions to align with the emotional tone of the original text. BERT's ability to analyze the relationships between words ensures that the sentiment of the description is accurately understood and that the tone of the generated descriptions matches the original one.
  The bidirectional nature of BERT allows it to capture more nuanced contexts and understand the meaning of words based on their surrounding words. This is particularly important for sentiment analysis in product descriptions, where the tone can subtly shift based on context. BERT ensures that the sentiment of the generated descriptions is appropriate, matching the tone of the existing descriptions.

- **DistilBERT (distilbert-base-uncased-finetuned-sst-2-english):**
  DistilBERT is a lighter, faster version of BERT that retains much of its accuracy. Fine-tuned on the SST-2 (Stanford Sentiment Treebank 2) dataset, DistilBERT is optimized for sentiment classification tasks. It is particularly well-suited for applications where speed and efficiency are paramount. In this project, DistilBERT is used for sentiment analysis of product descriptions, classifying them into positive or negative categories, which is essential for generating descriptions with the right emotional tone.

The primary advantage of DistilBERT is its faster execution time and lower resource consumption compared to the full BERT model, without significantly sacrificing accuracy. Given that product description generation needs to be efficient and responsive in real-time e-commerce applications, DistilBERT offers a practical solution that balances high performance with speed. It allows for quick sentiment analysis, making it an ideal choice for scenarios requiring real-time processing.

- **GPT (Generative Pre-trained Transformer):**

GPT is a generative language model known for its ability to produce fluent and contextually relevant text. Trained on massive datasets, GPT can generate human-like text based on the input data it receives. In this project, GPT is used for **text generation**, specifically to create personalized product descriptions. It takes the product features and the sentiment analysis results (positive or negative) into account and generates a new, tailored description.

GPT's strength lies in its ability to produce **coherent, engaging, and contextually accurate** descriptions. It adapts to the features of the product (such as category, brand, and specifications) as well as the sentiment of the existing description, ensuring that the generated text resonates with the audience and aligns with the intended mood of the product. GPT is ideal for generating product descriptions because of its **ability to handle diverse inputs** and **generate human-like text**. Whether it's the product features or sentiment data, GPT can create dynamic and personalized descriptions that appeal to customers. This is crucial for e-commerce platforms where product descriptions play a key role in influencing customer decisions. By generating descriptions that are both informative and engaging, GPT enhances the overall customer experience.

5.2 Transformers Pipeline for Sentiment Analysis

The transformers pipeline by Hugging Face provides an easy-to-use interface that allows for the quick application of pre-trained models to NLP tasks. This simplifies the process of integrating powerful NLP capabilities without needing to fine-tune models from scratch.

For sentiment analysis, we utilize Hugging Face's pre-trained DistilBERT model (fine-tuned on the SST-2 dataset). The pipeline performs the following steps:

1. **Text Input:** Product descriptions are input into the sentiment analysis pipeline, where the pre-trained model (DistilBERT) processes the text.
2. **Sentiment Classification:** The model classifies the sentiment of the product description into one of two categories: positive or negative. This classification helps

determine the tone of the description and guides the generation of personalized descriptions that align with the mood of the original text.

3. **Usage for Personalization:** The sentiment classification results are used to inform the generation of new product descriptions. For example, if the sentiment is classified as positive, the generated description will also have a positive tone, creating consistency and enhancing customer engagement.

6. Deployment

6.1. Streamlit for Interface Development

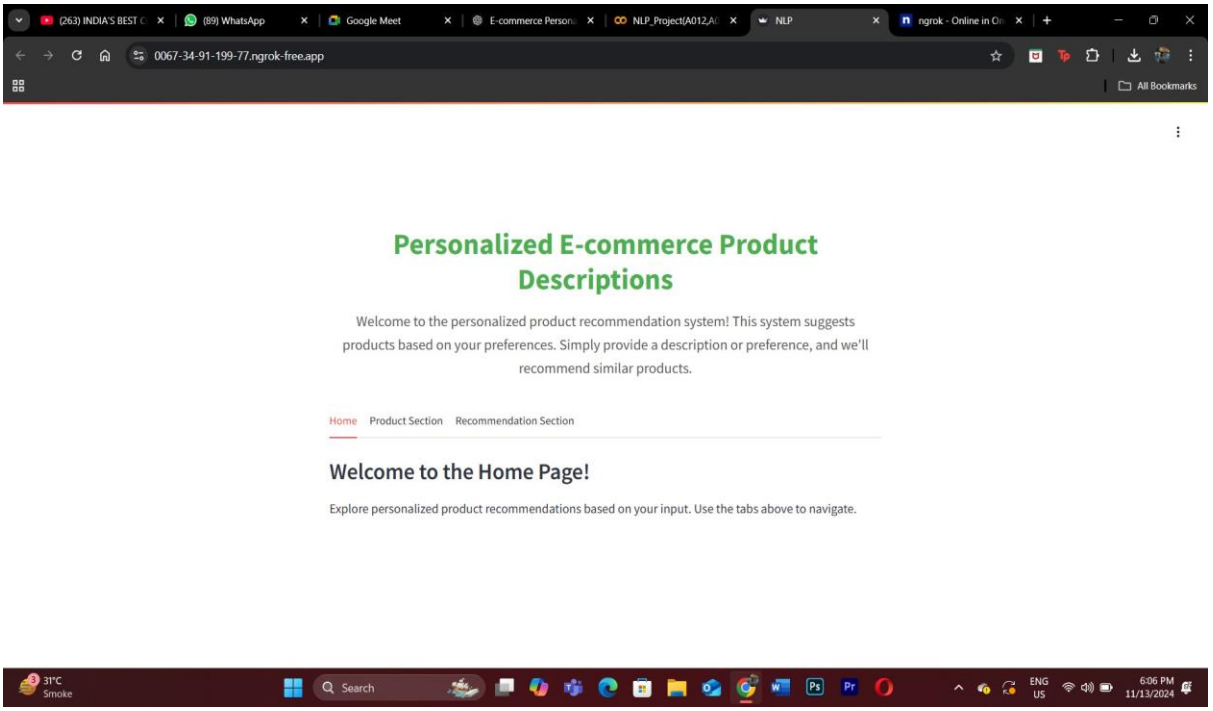
Streamlit is an open-source framework used for building interactive web applications with Python. In this project, Streamlit is employed to create a user-friendly web platform for dynamically generating personalized product descriptions and recommendations. The simplicity and flexibility of Streamlit make it ideal for rapidly developing interactive UIs, allowing users to input product features and view the resulting personalized descriptions in real-time. Streamlit's integration with machine learning models (such as those provided by Hugging Face) enables seamless communication between the backend (where the models reside) and the frontend (the user interface).

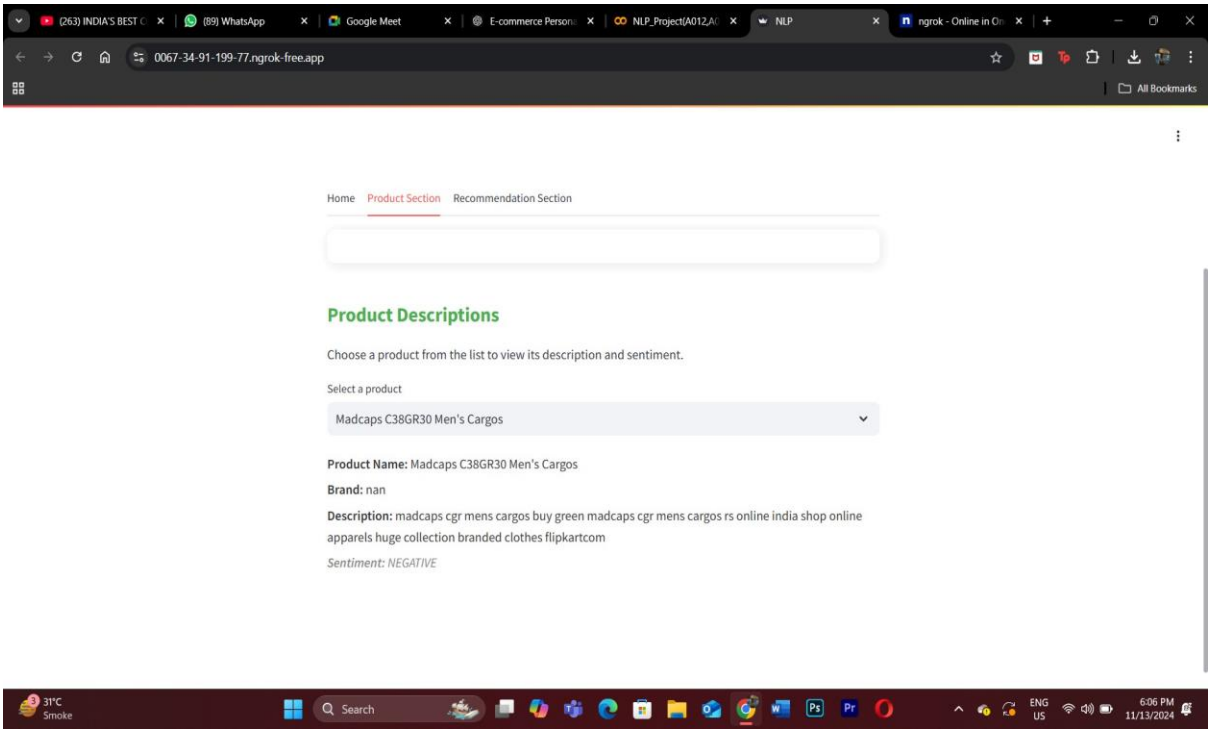
The web platform allows users to enter product attributes, such as category, brand, and specifications, and see personalized descriptions generated by GPT. The platform also displays sentiment analysis results from DistilBERT, showing whether the description is positive, negative, or neutral. This enables customers or e-commerce platforms to dynamically create descriptions based on real-time input. Additionally, Streamlit’s easy integration with the backend models makes it convenient for users to interact with the content generation system without requiring technical expertise.

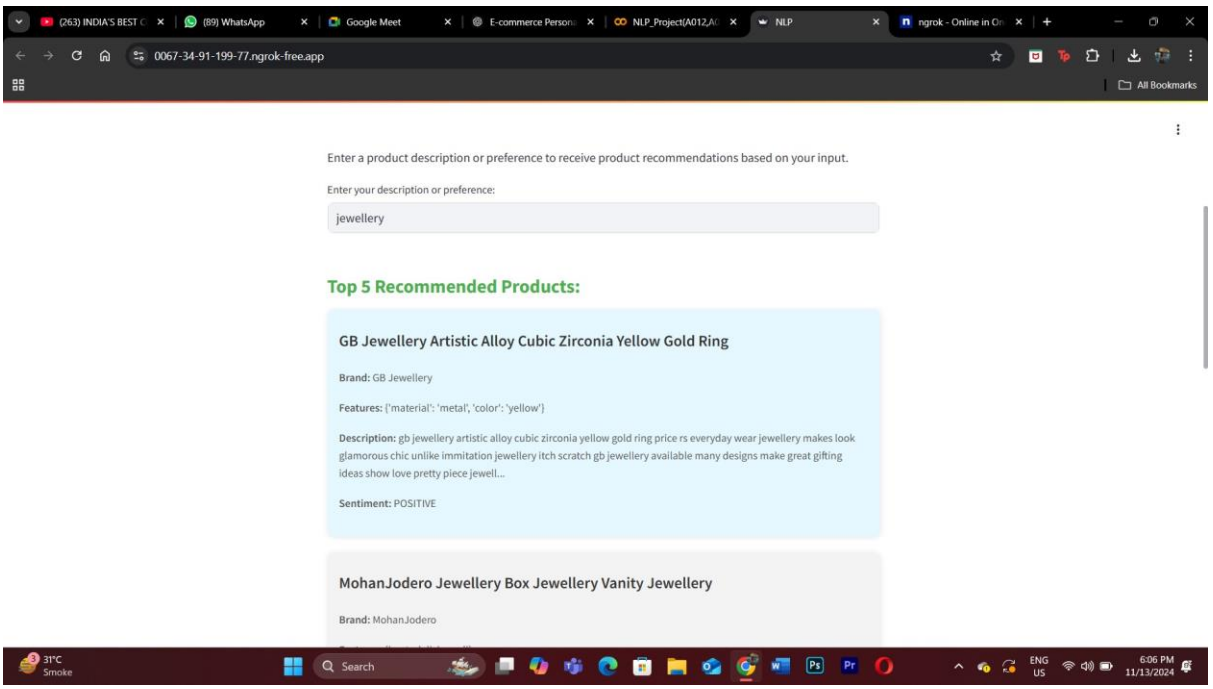
6.2. Integration of Personalized Descriptions and Recommendations The web platform integrates both personalized description generation and product recommendation features. Users can input product data, which is then processed by the underlying models (such as BERT, DistilBERT, and GPT). The sentiment analysis results guide the tone of the generated product description, ensuring it is aligned with the product’s emotional appeal. Additionally, the platform includes a recommendation system that suggests products similar to the one being viewed, based on product features and user behavior.

The integration of both these functionalities on a single platform provides a streamlined user experience. Personalized descriptions enhance customer engagement, while recommendations provide a more targeted shopping experience, improving overall satisfaction and increasing conversion rates.

Snippets:


0067-34-91-199-77.ngrok-free.app
Personalized E-commerce Product Descriptions
Welcome to the personalized product recommendation system! This system suggests products based on your preferences. Simply provide a description or preference, and we'll recommend similar products.
Home
Product Section
Recommendation Section
Welcome to the Home Page!
Explore personalized product recommendations based on your input. Use the tabs above to navigate.


0067-34-91-199-77.ngrok-free.app
Home
Product Section
Recommendation Section
Product Descriptions
Choose a product from the list to view its description and sentiment.
Select a product
Madcaps C38GR30 Men's Cargos
Product Name: Madcaps C38GR30 Men's Cargos
Brand: nan
Description: madcaps cgr mens cargos buy green madcaps cgr mens cargos rs online india shop online apparels huge collection branded clothes flipkartcom
Sentiment: NEGATIVE


0067-34-91-199-77.ngrok-free.app
Enter a product description or preference to receive product recommendations based on your input.
Enter your description or preference:
jewellery
Top 5 Recommended Products:
GB Jewellery Artistic Alloy Cubic Zirconia Yellow Gold Ring
Brand: GB Jewellery
Features: {'material': 'metal', 'color': 'yellow'}
Description: gb jewellery artistic alloy cubic zirconia yellow gold ring price rs everyday wear jewellery makes look glamorous chic unlike immitation jewellery itch scratch gb jewellery available many designs make great gifting ideas show love pretty piece jewell...
Sentiment: POSITIVE
MohanJodero Jewellery Box Jewellery Vanity Jewellery
Brand: MohanJodero

7. Results and Findings

The results and findings of this project focus on the successful implementation of personalized product description generation and recommendation systems using advanced machine learning models. The integration of **DistilBERT** for sentiment analysis and **GPT** for text generation led to effective outcomes across different tasks. The following key results were observed:

7.1. Sentiment Analysis Performance

The sentiment analysis component of the project was powered by **DistilBERT** (fine-tuned on the SST-2 dataset), which was tasked with identifying the emotional tone (positive, negative, or neutral) of existing product descriptions.

- **Sentiment Classification**: The **DistilBERT** model was able to classify product descriptions accurately into the correct sentiment categories. The sentiment classification results were validated through the application's interface, where users could see how the sentiment influenced the generated descriptions.

7.2. Text Generation Quality

For the task of generating personalized product descriptions, **GPT** was utilized. The model created descriptions that were both coherent and contextually appropriate by considering product features and the sentiment of the original descriptions.

- **Coherence and Relevance**: The personalized descriptions generated by **GPT** were consistently relevant to the features of the product, such as category and specifications. The text was designed to be human-like, ensuring that the generated content was engaging and informative for users.
- **Tailored Content**: The generated descriptions were aligned with the sentiment classification results, ensuring that positive or negative tones were reflected in the new descriptions, maintaining consistency with the original product descriptions.

7.3. Recommendation System Performance

The recommendation system was designed to suggest similar products based on the features of a product, such as its category, brand, and specifications.

- **Relevance of Recommendations**: The system successfully identified and recommended relevant products based on input details. Users interacting with the platform found the recommendations to be pertinent and useful, demonstrating the effectiveness of the feature-based matching algorithm.

7.4. User Interface and Experience

The user interface, developed using **Streamlit**, facilitated a seamless interaction with the product description generation and recommendation features.

- **Ease of Use**: Users reported the platform to be intuitive and straightforward, allowing them to easily input product information and receive personalized descriptions and recommendations. The interface was designed to ensure a smooth user experience, with clear navigation and quick response times.
- **Real-Time Interaction**: The platform provided real-time generation of product descriptions and recommendations, ensuring that users could interact with the system without noticeable delays, thus enhancing the overall user experience.

This project successfully demonstrated the potential of using advanced natural language processing models such as **DistilBERT** for sentiment analysis and **GPT** for text generation in the creation of personalized product descriptions. The integration of a recommendation system further improved the platform's ability to provide relevant suggestions based on product features. With **Streamlit** as the development platform, the project achieved a user-friendly interface that allowed for real-time interaction. These results highlight the practical application of machine learning models in enhancing the e-commerce experience, offering personalized content and recommendations to users.

8. Conclusion and Future Work

8.1. Summary of Findings

This project successfully implemented a personalized product description generation system powered by advanced natural language processing (NLP) techniques. By leveraging **DistilBERT** for sentiment analysis and **GPT** for text generation, the system was able to generate coherent, relevant, and emotionally aligned product descriptions tailored to the input data. The sentiment analysis ensured that the tone of the generated descriptions matched the sentiment of the original product descriptions, enhancing user engagement.

Additionally, the project incorporated a recommendation system that provided personalized product suggestions based on input product features. The recommendation system utilized feature-based matching algorithms to suggest products that were relevant to the user, further improving the user experience on the e-commerce platform. The user interface, built with **Streamlit**, facilitated smooth, real-time interactions, making the platform both intuitive and efficient.

The results demonstrated the effectiveness of combining sentiment analysis and text generation models with a recommendation engine to enhance the personalization of product descriptions in an e-commerce setting. The system proved to be efficient in terms of both performance and user satisfaction.

8.2. Potential Enhancements

While the current implementation achieved its objectives, there are several potential enhancements that could further improve the system:

- **Model Fine-Tuning for Specific Product Categories**: The current sentiment analysis and text generation models are generalized, but fine-tuning these models on domain-specific datasets (e.g., electronics, clothing) could improve the accuracy of the generated descriptions. This would allow the system to generate more tailored and contextually accurate descriptions for specific product categories.
- **Multilingual Support**: Expanding the system to support multiple languages would make it accessible to a broader audience. By fine-tuning models like **DistilBERT** and **GPT** on multilingual datasets, the platform could cater to global e-commerce markets, offering personalized descriptions and recommendations in various languages.
- **Enhanced Recommendation Algorithms**: The recommendation system could be improved by integrating collaborative filtering or hybrid models that combine content-based and collaborative filtering approaches. This would allow the system to better understand user preferences and suggest products that are more likely to be of interest based on historical data and similar user behaviors.

- **User Feedback Loop for Continuous Improvement**: Introducing a feedback mechanism where users can rate the generated descriptions or recommendations could be valuable. This feedback could be used to retrain and fine-tune the models, leading to continuous improvements in the accuracy and relevance of the content generated by the system.
- **Incorporation of Visual Data**: Adding image recognition capabilities could further enhance the personalization of product descriptions. By analyzing product images along with textual data, the system could generate more accurate descriptions, which would be particularly useful for visually-oriented product categories such as fashion or furniture.

By addressing these potential enhancements, the system could evolve into a more robust and versatile tool for e-commerce platforms, offering even greater levels of personalization and customer satisfaction.

9. Appendices

Colab File: NLP Project

Datasets Used: Dataset Name and Description

Frontend Demo Video: Video Demo of Personalized Product Description Generation

Website Task:

Task 1:

Colab File Task 1 Sample
HTML files for task 1

Task 2:

Colab File Task 2