Generative AI Conversational Bot for Enhanced Product Review Insights

Nikita Sateesh Chigateri
Department of Computer Science
Western University
London, ON N6A 3K7
nchigate@uwo.ca

Raj Tulluri
Department of Computer Science
Western University
London, ON N6A 3K7
rtulluri@uwo.ca

Abstract

In the digital era, online shopping has become a predominant mode of commerce, with consumers relying heavily on product reviews to make informed purchasing decisions. However, the sheer volume of available reviews poses a significant challenge, often overwhelming potential buyers and obscuring valuable insights. This project introduces an innovative solution to this problem: a conversational bot designed to distil and articulate key insights from product reviews, thus facilitating a more natural, context-aware, and engaging conversation harnessing the power of AI through two main components: a RASA-based intent classification model and a GPT-2 driven conversational agent. The former processes and categorizes user inquiries into distinct intents, covering the spectrum of potential buyer queries. Additionally, a semantic analysis model is employed to ascertain the sentiment of each review, enabling the bot to distinguish between positive and negative feedback in its responses. The dataset employed for training and evaluating our system comprises Amazon reviews pertaining to beauty and personal care products. This corpus enables our model to understand and generate human-like responses across a wide range of topics within the beauty and personal care domain, addressing queries, offering recommendations, and providing advice with a level of relevance and personalization. One of the project's novel challenges is maintaining conversational context, allowing the bot to handle follow-up questions with coherence and relevance. By offering a streamlined, interactive way to navigate product reviews, our project stands at the intersection of AI, NLP, and e-commerce, contributing valuable insights to the field and promising to enhance the online shopping experience for consumers.

1 Introduction

The digital transformation of commerce has revolutionized the way products are bought and sold, with e-commerce emerging as a dominant force in the global market. Over the past decade, the proliferation of online platforms has facilitated unprecedented access to a vast array of products, ranging from everyday necessities to specialized goods. This shift has not only expanded consumer choices but also intensified competition among retailers. As a result, e-commerce platforms continuously seek innovative ways to enhance the shopping experience and foster customer loyalty. One significant factor contributing to this evolution is the strategic use of technology to leverage consumer data, thereby personalizing the shopping journey and making it more convenient and responsive to individual preferences. Within this digital marketplace, customer reviews have become a critical element of the online shopping experience, influencing consumer behavior and business outcomes alike. Reviews provide a platform for consumers to share their experiences, offering insights that go beyond product descriptions provided by sellers.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

The study by Mo et al. [2015] delves into the significant impact of online reviews on consumer purchasing behavior, offering robust empirical evidence to support the assertion that reviews are critical in the e-commerce landscape. Their research systematically analyzes how both the presence and the qualitative aspects of online reviews influence buyers' decisions across various product categories. By employing a combination of quantitative data analysis and behavioral modeling, the study finds that consumers are not only likely to rely heavily on reviews in the absence of physical product verification but also that the credibility and detail contained within these reviews substantially sway their purchasing choices. The study concludes that reviews act as a pivotal source of consumer trust and risk mitigation, essentially serving as a proxy for quality assurance in online shopping environments, underscoring the dual role of reviews in enhancing transparency and fostering informed consumer decisions.

To mitigate these issues, this study proposes an innovative conversational AI system tailored to enhance the online shopping experience by efficiently processing and condensing the wealth of information contained in product reviews. At the heart of this system lies a dual-component AI framework that integrates RASA [Bocklisch et al., 2017], an open-source framework for building conversational agents, with a GPT-2 [Radford et al., 2019b] based conversational model trained to generate human-like responses. The intent classification component powered by RASA effectively categorizes user inquiries into predefined intents such as product features, user satisfaction, and comparisons, allowing the system to handle a broad spectrum of customer queries with high precision. Simultaneously, the sentiment analysis module assesses the emotional tone of each review, enabling the AI to distinguish between positive and negative feedback and tailor its responses accordingly. By dynamically generating context-aware, personalized responses, this AI system not only alleviates the burden of navigating through excessive review data but also delivers a more engaging and user-friendly shopping experience. Through this approach, the project aims to transform how consumers interact with online reviews, making the process more strategic and less overwhelming.

The paper begins with a Related works section that surveys existing literature, focusing on the role of conversational AI in analyzing customer feedback and its impact on purchasing behavior. The Methodology section is divided into several subsections, each detailing a different aspect of our experimental approach: we describe the Dataset used for training and testing, elaborate on the configuration and role of RASA in intent classification, discuss the implementation of the GPT-2 model for generating human-like textual responses, and outline our Training & Testing procedures, including model optimization and evaluation metrics. Following this, the Results section presents the outcomes of our experiments, providing quantitative and qualitative analyses to assess the effectiveness of the conversational agent. Finally, the Conclusion summarizes the key findings, discusses the implications of our work. Additionally, the Future Works section outlines prospective research directions, including the expansion into multilingual capabilities, integration of multi-modal data, and enhancement of learning architectures, which aim to refine the system's responsiveness and adaptability in real-world settings.

2 Related Works

This section examines pivotal research that informs the development of our conversational AI system. As the integration of artificial intelligence within e-commerce platforms becomes increasingly sophisticated, understanding the landscape of existing technologies and methodologies is crucial. This section reviews three seminal papers that have significantly contributions. Each study provides essential insights into the capabilities and limitations of current systems, serving as a foundation upon which our project builds.

Yang et al. [2018] introduce an innovative approach to abstractive review summarization. This study addresses the complex challenge of generating concise yet informative summaries from extensive product reviews by incorporating aspect-based and sentiment-aware techniques. The methodology employs a multi-factor attention mechanism using an encoder-decoder framework that interactively learns from the context, sentiment, and aspect words within reviews. The encoder effectively captures different facets of information in the reviews, and the decoder integrates these insights to produce coherent and context-rich summaries. The relevance of their work to our project lies in their sophisticated use of sentiment analysis and aspect extraction, which align closely with our intent classification and sentiment analysis tasks. From their research, we adopt the notion of integrating sentiment analysis into our processing pipeline to tailor responses that are contextually and emotionally congruent with the users' inputs. However there are limitations in their research which we aim to address. Their model primarily focuses on text data and does not account for the conversational aspects of AI interactions, such as maintaining dialogue flow and responding to specific user queries in real-time.

In their seminal work, Cui et al. [2017] explore the development of "Superagent," a dedicated customer service chat bot designed for e-commerce platforms. The study's methodology revolves around building a robust chat bot system capable of handling a wide range of customer inquiries, from product details to order status and return policies. The Superagent utilizes a combination of rule-based systems and machine learning models to interpret customer inputs and generate appropriate responses. Their approach to automating customer support through a chat bot closely aligns with our project's aim to enhance the shopping experience through conversational AI. We draw inspiration from their methodology, particularly their use of a mixed approach combining rule-based and machine learning techniques, which informs our design for intent recognition and response generation in our AI system. While effective in handling predefined queries, the Superagent lacks deeper contextual understanding and the ability to manage extended conversational states, which are crucial for maintaining engaging and coherent interactions over time.

Bhawiyuga et al. [2017] address the design and implementation of an e-commerce chat robot aimed at automating customer service interactions. Their approach combines predefined scripts for common queries with a basic understanding mechanism to interpret less straightforward customer inputs. The core of their chat robot's functionality relies on a comprehensive set of scripted responses designed to cover a wide range of frequently asked questions (FAQs) and to handle customer queries that extend beyond straightforward FAQs, the system incorporates a basic Natural Language Understanding (NLU) component. This component is designed to parse and understand customer inputs using keyword extraction and pattern matching techniques. This aligns with our project's objective to enhance user experience through AI-driven conversational agents. From their study, we adopt the idea of automating responses to frequently asked questions, which informs our approach to integrating robust intent classification mechanisms within our system. Their methodology, however, primarily focuses on handling predefined queries and lacks the depth of contextual understanding and adaptive conversational ability that our project aims to achieve.

In conclusion, the review of these foundational studies highlights significant strides in the development of conversational AI and chat bots within the e-commerce domain. Each of the discussed works—Yang et al. [2018], Cui et al. [2017] and Bhawiyuga et al. [2017] —brings valuable insights into different aspects of automated customer interaction systems, from advanced review summarization to basic and complex query handling. While these studies lay a robust groundwork, they also reveal gaps particularly in areas such as deep contextual understanding, dynamic conversational capabilities, and personalized engagement based on sentiment analysis. Our research aims to bridge these gaps by integrating sophisticated intent classification, sentiment analysis, and conversational capabilities by exploiting GPT-2.

3 Methodology

This section explicates each component of the project, from data acquisition and pre-processing to the training of models for intent classification, culminating in the integration of these elements within a conversational agent powered by GPT-2. The detailed description provided here encompasses the selection and preparation of the dataset, the specific configurations and customizations applied to the RASA framework for intent recognition, the utilization of the GPT-2 model for generating nuanced responses, and the methodologies adopted for training and testing the system's efficacy.

3.1 Dataset

In this study, we utilize a subset of the comprehensive Amazon US Customer Reviews Dataset, available on Kaggle, which includes a wide range of product categories. Specifically, we have chosen to focus on reviews from the Health & Personal Care and Beauty categories. These categories were selected because the products they encompass often require careful consideration by consumers due to their direct impact on health and personal well-being. Reviews in these sectors are not only abundant but also rich in detailed consumer feedback, reflecting a diverse array of user experiences and sentiments. This depth and variety are crucial for training our models to accurately understand and respond to complex user inquiries and emotions.

The Amazon US Customer Reviews Dataset is structured with several key columns crucial for our analysis. The attributes product id and product title serve to identify specific products, facilitating the aggregation and categorization of reviews. Most notably, the star rating and review body fields provide direct insights into the consumer's sentiment. The verified purchase flag and total votes count are used to assess the authenticity and community-endorsed relevance of each review, respectively. To ensure the reliability and quality of the data, stringent filtering criteria were applied. Initially, reviews stemming from unverified purchases were excluded to mitigate the inclusion of biased or non-authentic feedback. Additionally, only products with a minimum of 100 reviews were considered to guarantee a substantial volume of data per product. Finally, reviews garnering fewer than 10 helpful votes were omitted to focus on feedback that has been community-validated as useful. These filtering steps are critical to refining the dataset, ensuring that the conversational AI is trained on high-quality and trustworthy data.

Table 1: Review Counts Before and After Data Cleaning

Category	Initial Review Count	Final Review Count
Health & Personal care	5,331,449	127,099
Beauty	5,115,666	71,031

To quantify the impact of our data filtering criteria on the dataset, we present Table 1, which summarizes the initial and final counts of reviews for the Health and Beauty product categories.

The preprocessing of review data is a crucial step in preparing the input for our conversational AI system, ensuring that the text is clean and standardized for effective analysis and model training. Initially, the review subject and review body are combined to form a single text field for each entry. Subsequent cleaning operations are methodically applied to enhance data quality and uniformity:

- **Normalization:** All texts are converted to lowercase to standardize the input, mitigating any case sensitivity issues that might affect the analysis.
- **Whitespace Management:** Extraneous whitespaces are eliminated, including trimming spaces at the beginnings and ends of texts and reducing multiple spaces to a single space within the text.
- **HTML Tag Removal:** Raw data often contains HTML tags; these are removed to prevent parsing errors during the natural language processing stages.
- **Special Character Removal:** All special characters are removed from the texts, except for emojis. Emojis are preserved due to their significant emotional content, which is pertinent for sentiment analysis.
- **Sentence Tokenization:** The cleaned reviews are then segmented into individual sentences using a sentence tokenizer.

Segmenting reviews into individual sentences forms a crucial part of our preprocessing methodology. A single review can often encompass a range of emotions and intents, reflecting a complex mix of consumer feedback that might include both praises and criticisms within the same text. By breaking down reviews into sentences, our system can analyze and respond to each specific sentiment or intent expressed, rather than treating the review as a homogenous block of text. This sentence-level analysis allows the AI to identify nuanced emotional cues and detailed aspects of consumer feedback more effectively.

3.2 Sentiment Analysis

Sentiment analysis is a critical component of natural language processing that involves analyzing textual expressions to discern the underlying sentiments. This computational technique is widely used to interpret the polarity of content within user-generated texts, such as online reviews, social media posts, and customer feedback, categorizing them into positive, negative, or neutral sentiments. Effective sentiment analysis enables businesses and researchers to gauge public opinion, and enhance customer interaction, especially in environments like e-commerce where customer sentiment can significantly influence purchasing behavior.

BERT (Bidirectional Encoder Representations from Transformers) is a groundbreaking model in the field of natural language processing developed by Google [Devlin et al., 2019]. Central to its architecture is the transformer, an attention mechanism that learns contextual relations between words (or sub-words) in a text. Unlike previous models that processed text sequentially (either left-to-right or right-to-left), BERT reads the entire sequence of words at once, thereby allowing it to capture the context from both directions simultaneously. This bidirectional nature is one of the core innovations of BERT, making it particularly effective for tasks that rely on a deep understanding of context.

In this study, we utilize a derivative of the BERT architecture. DistilBERT [Sanh et al., 2019] is designed to deliver a similar performance to BERT while being more efficient. This model architecture achieves a substantial reduction in size and increase in speed by distilling the knowledge of BERT. The chosen DistilBERT model was trained on a comprehensive dataset which includes a balanced distribution of sentiments across various languages, sourced from multiple international domains including Amazon product reviews. This training dataset is ideal due to its diversity and volume, which ensure the model's robust performance across different sentiment analysis scenarios as well as it's ability to understand patterns of how reviews are written by customers. This model assigns a sentiment score and sentiment label (positive, negative or neutral) to each sentence of each review segmented in the previous step. This information is added to the dataset to further enhance contextual capabilities of the intent recognition model and conversational chat bot.

3.3 Intent Classifier

Intent classification is a fundamental task in NLU that involves identifying the purpose or intent behind a user's input. It plays a critical role in building conversational AI systems, such as chat bots or virtual assistants, enabling these technologies to understand and respond appropriately to user requests. By classifying an input into predefined categories, the system can determine the most relevant actions or responses to provide, based on the user's intentions. Intent classification helps bridge the gap between human language and machine response, making it a key component in automating and improving communication between users and AI-driven systems.

RASA [Bocklisch et al., 2017] is an open-source framework for building conversational AI applications, such as chat bots and virtual assistants. It is designed to provide developers with the tools necessary to build sophisticated AI-driven conversational interfaces that can understand and respond to user input effectively. Within the RASA ecosystem, the DIETClassifier (Dual Intent and Entity Transformer) is a prominent component specifically engineered for intent recognition and entity extraction, which are crucial for understanding the user's intent and extracting relevant information from user inputs. The DIETClassifier [Bunk et al., 2020] leverages a transformer architecture (as shown in Figure 1), which allows it to handle complex language understanding tasks by considering the context of words and their relationships within a sentence. This capability makes it exceptionally good at distinguishing between different intents and entities, even when the linguistic expressions are subtle or complex. As a lightweight and scalable model, DIET requires significantly less computational power and memory. The key advantage of using DIET classifier is its ability to train on relatively less data for the chosen intents. It offers a faster and less expensive alternative to intent classification models requiring vast amounts of data to train.

For the training of the DIET classifier, a specialized pipeline was constructed. The pipeline begins with the WhitespaceTokenizer, which segments text into tokens based on whitespace, this is followed by two instances of the CountVectorsFeaturizer; the first operates at the word level, and the second is configured to analyze character n-grams from 1 to 4 enhancing the model's ability to understand and capture sub-word information which can be critical for recognizing intents in morphologically rich words.

In this project, we addressed a total of 17 distinct intents (as shown in Table 2) reflecting the varied types of inquiries and responses typical in customer reviews of beauty, health and personal care products. To train the DIET classifier, we meticulously hand-labeled 200 sentences for each intent, drawn from actual review texts, to serve as training data. An additional set of 100 sentences per intent was reserved for testing the classifier's performance. The training process involved experimenting

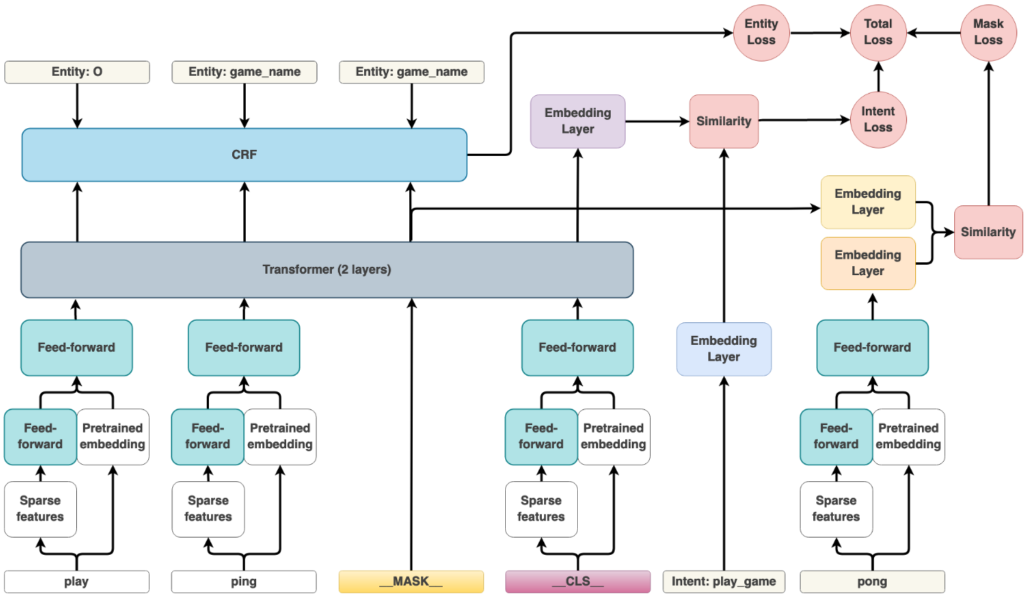

Figure 1: A schematic representation of the DIET architecture. The phrase ”play ping pong” has the intent play game and entity game name with value ”ping pong”. Weights of the feed-forward layers are shared across tokens.

Table 2: List of Intents for DIET Classifier Training

Intent Number	Intent Description
1	express_satisfaction
2	compare_products
3	instructions
4	usage
5	express_dissatisfaction
6	pricing
7	ask_result
8	features
9	verify_authenticity
10	ask_usage
11	ask_quantity
12	ask_packaging
13	ask_shipping
14	recommendation
15	report_side_effect
16	ask_features
17	return_and_refund

with various hyper-parameters to refine the model's accuracy; this included adjusting the number of transformer layers and the size of the embedding used by the DIET classifier.

Table 3 presents a detailed comparison of the performance of the DIET classifier models trained with varying configurations of transformer layers and embedding sizes. From these configurations, the model with 3 transformer layers and an embedding size of 128 emerged as the most effective, achieving an accuracy of 88.22%, with precision and recall closely aligned at 88.93% and 88.25%, respectively. To visually represent the performance distribution across all intents, a histogram has been included as shown in Figure 2. This histogram effectively illustrates the accuracy and recall rates for each intent, providing a clear graphical representation of the classifier's capabilities and the consistency of its performance across different types of user queries.

Table 3: Comparison of DIET classifier Performance with Different Hyper-parameters

Model	Transformer Layers	Embedding Size	Accuracy	Precision	Recall
1	3	512	83.96%	85.01%	83.95%
2	3	256	84.37%	85.32%	84.37%
3	3	128	88.22%	88.93%	88.25%
4	2	128	76.02%	75.88%	76.02%

3.4 GPT

Generative Pre-trained Transformer 2 (GPT-2) is an advanced model developed by Radford et al. [2019a], built upon the transformer architecture, which has revolutionized the field of natural language processing. At its core, GPT-2 employs a stacked transformer decoder architecture, comprising multiple layers of self-attention mechanisms and fully connected neural networks. Each layer of the model processes input sequences entirely in parallel, which significantly enhances the computational efficiency. The self-attention mechanism within these layers allows GPT-2 to weigh the importance of each word in a sentence, regardless of its position, enabling the model to generate coherent and contextually relevant text based on the learned relationships between all words. GPT-2 was pre-trained on a diverse internet corpus using unsupervised learning, specifically through the task of predicting the next word in a sentence, which allows it to generate high-quality text across a wide range of topics and styles.

The choice to utilize GPT-2 in this project stems from its robust language generation capabilities, which are essential for creating a conversational AI that can generate human-like responses. Furthermore, GPT-2's ability to maintain context over extended text makes it particularly well-suited for dialog systems, where continuity and relevance of conversation are paramount. By leveraging GPT-2, the project aims to enhance the interactivity and user engagement of the AI system, ensuring that the generated responses are not only accurate but also varied and engaging, mimicking a natural conversational experience.

To prepare the training dataset for the GPT-2 model, a meticulous process was followed that began with the extraction of intents from individual sentences of product reviews using the DIET Classifier. Each sentence within a review was labelled with its respective intent. After all sentences were processed for intent, they were recombined into their original review format. This recombination allowed us to compile a comprehensive list of all intents present within each complete review, reflecting the full scope of user interactions and inquiries about the product. Subsequently, for each identified intent within a review, a corresponding question was generated. This was achieved by randomly selecting a question from a pre-defined list tailored to match specific intents. This approach ensured that the questions were contextually relevant and varied, enhancing the training data's diversity and richness.

The final training data was structured in a specific format conducive to training the conversational model: each entry comprised the context (the product name), a relevant question based on the detected intent, and the full text of the review as the answer. This example illustrates how each training data entry is formatted for the GPT-2 model:

Example 1

```
Context: plant therapy black pepper essential oil. 100% pure,
undiluted, therapeutic grade. 10 ml (1/3 oz)
Question: In what situations is this product most useful?
Answer: 5 stars great eo. very pleased with how fast i received this. this has
a very pleasant smell, not overpowering, which i liked. i used it in a
blend i made for pain and it worked great! i will keep this on hand
from now on
```

Example 2

```
Context: thunder ridge emu products pure emu oil, 8 ounce
Question: What are the most common uses for this product?
Answer: 5 stars wonderful stuff for preventing stretch marks. I used emu oil
```

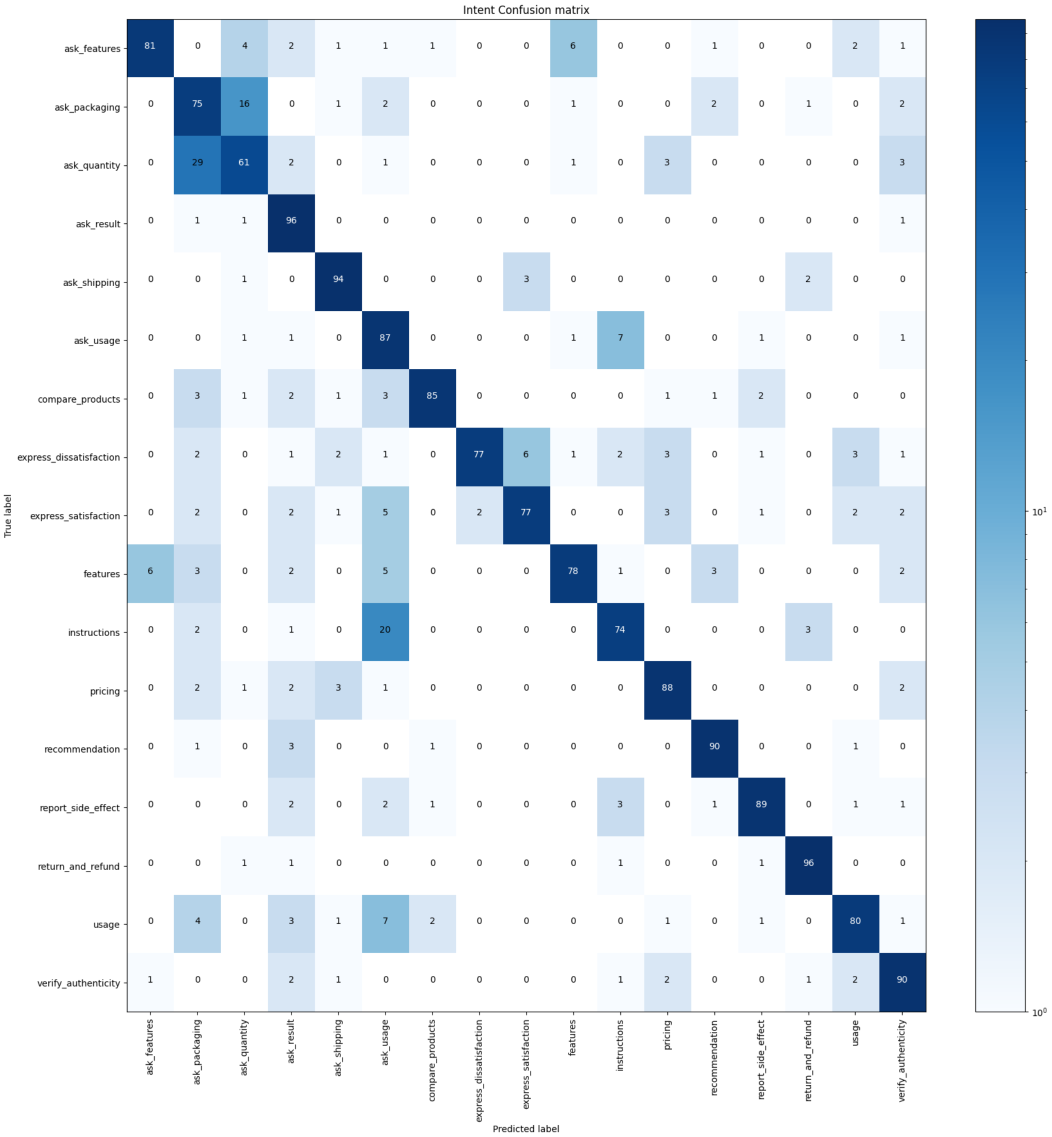

Figure 2: Histogram of DIET classifier with 3 transformer layers and an embedding size of 128 dimensions.

`with my first pregnancy and didn’t get any stretch marks. I am prone to stretch marks as I have gotten them from just gaining weight. I plan to use it again with my second pregnancy.`

For the fine-tuning of the GPT-2 model, specifically the distilgpt2 variant, a structured and iterative training approach was adopted to adapt the model to domain-specific needs. Initially, a GPT2Tokenizer was employed to pre-process the training data, ensuring that all texts were tokenized uniformly up to a maximum length of 128 tokens. The training was conducted over several iterations, with checkpoints being saved periodically to capture the model’s state at strategic points. This approach allowed for continuous evaluation and adjustment of the model’s parameters to optimize performance. Each iteration of training helped refine the model’s ability to understand and generate contextually relevant and coherent responses, resulting in a robust, domain-adapted conversational agent.

4 Results

In the evaluation phase of the project, the GPT-2 model was tested to assess its capability to generate contextually appropriate and coherent responses based on a structured prompt format. Each prompt consisted of three parts: a contextual lead-in which was the product name, a directed question related to the review's intent, and an open-ended segment where the model was expected to generate a continuation, specifically the answer. The effectiveness of the generated responses was quantitatively measured using a separated testing set, which was not part of the training data, ensuring the evaluation's integrity and robustness.

To evaluate the performance of the GPT-2 model in generating text, several established metrics were employed, each providing insights into different aspects of text quality. The BLEU (Bilingual Evaluation Understudy) score [Papineni et al., 2002], primarily used in machine translation, quantifies the similarity between the generated text and a set of reference texts by measuring the precision of n-grams. BLEU scores range from 0 to 1, where 1 indicates a perfect match with the reference, though they are often expressed as a percentage. METEOR (Metric for Evaluation of Translation with Explicit ORdering), another metric developed for translation [Banerjee and Lavie, 2005], extends beyond BLEU by incorporating synonymy and stemming, allowing for a more nuanced comparison and typically provides a score also ranging from 0 to 1, where higher values represent better alignment with human judgment. Lastly, the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric [Lin, 2004], commonly used for summarization tasks, assesses the overlap of n-grams, word sequences, and word pairs between the generated text and references. ROUGE scores are similarly bounded between 0 and 1, where 1 signifies that all reference n-grams are present in the generated text, indicating higher recall. These metrics collectively enable a comprehensive analysis of the linguistic quality, relevance, and fluency of the responses generated by the AI system.

Table 4: Performance Metrics of the GPT-2 Model Across Datasets

Dataset	**Metric**	**Score**	**Recall**	**Precision**	**F1-Score**
Beauty	**BLEU**	31.02%	-	-	-
	METEOR	16.60%	-	-	-
	ROUGE-1	-	28.21%	29.85%	28.66%
	ROUGE-2	-	10.093%	10.085%	10.019%
	ROUGE-L	-	7.91%	9.61%	8.40%
Health & Personal Care	**BLEU**	36.99%	-	-	-
	METEOR	16.38%	-	-	-
	ROUGE-1	-	28.44%	29.76%	28.75%
	ROUGE-2	-	10.059%	10.031%	10.099%
	ROUGE-L	-	8.15%	9.57%	8.53%

The evaluation results from the GPT-2 model across the Beauty and Health & Personal Care datasets, as summarized in the Table 4, indicate significant performance in the generation of text relative to the human-crafted references. The BLEU scores for the Health & Personal Care dataset at 36.99% and for the Beauty dataset at 31.02% signify a reasonably high degree of lexical similarity to the reference texts. These scores suggest that the model is effectively capturing key phrases and terms from the reference data. The METEOR scores, which are 16.60% for Beauty and 16.38% for Health & Personal Care, while moderate, highlight a critical aspect of the model's performance: its difficulty in achieving a human-like structural and semantic alignment with the reference texts. METEOR assesses not only the exact matches but also synonymy and the overall order of words, providing a more rigorous examination of language understanding. The diverse nature of product reviews, where customers may express similar sentiments in varied linguistic styles or using different terminologies, poses a significant challenge. This diversity can lead to lower METEOR scores, as the model may not always capture or replicate ways in which humans might phrase their thoughts. The ROUGE scores provide a detailed view of the model's capability to recall and precisely match the structure of the reference texts. ROUGE-1 scores, nearly 28.66% for Beauty and 28.75% for Health & Personal Care, suggest that the model is fairly competent at capturing essential content words, or uni-grams. However, ROUGE-2 and ROUGE-L scores, which are crucial for evaluating the coherence and order of longer textual units, show lower performance. Particularly, the ROUGE-2 scores for Beauty and

Health & Personal Care reflect some ability to form bi-gram structures, but the figures remain modest, highlighting a potential area for improvement in generating more complex sentence structures.

Example 1

```
Context: plant therapy black pepper essential oil. 100% pure,
undiluted, therapeutic grade. 10 ml (1/3 oz)
Question: In what situations is this product most useful?
Generated Answer: this oil good for pain it helps with pain good smells good making mixtures for
```

Example 2

```
Context: thunder ridge emu products pure emu oil, 8 ounce
Question: What are the most common uses for this product?
Generated Answer: really good for skin oil it is good for preventing
used for pregnancy good skin and marks helps with skin pregnancy.
```

Example 1 demonstrates the model's limited linguistic diversity when tasked with describing the uses of black pepper essential oil. The response, while relevant, repetitively focuses on the oil's efficacy for pain relief and its pleasant aroma. This redundancy highlights the model's challenges in producing linguistically varied and rich text, which is evident from the low ROUGE-2 scores. Similarly, Example 2 reveals the model's struggle with generating nuanced text. Although the response correctly identifies common uses of emu oil for preventing stretch marks during pregnancy, it lacks depth and variety in expression, repeatedly cycling through a restricted vocabulary focused on skin benefits. This pattern aligns with the modest METEOR and ROUGE-L scores, underscoring the model's difficulties in expanding beyond basic repetitive structures to more sophisticated and varied linguistic constructions.

These examples and scores obtained collectively illustrate the current limitations of the GPT-2 model in generating detailed and contextually diverse responses, indicating areas for future improvement in model training and data handling.

5 Future Works

Looking forward, significant enhancements are planned to augment the capabilities and effectiveness of the conversational AI system under discussion. A primary objective involves extending the system's ability to maintain larger conversational contexts over extended interaction periods. This enhancement will necessitate refining the model's memory and context management capabilities to handle prolonged dialogue sessions, enabling it to retain pertinent information over time and reference past interactions. This capability is crucial for applications requiring continuous conversations, such as customer support or virtual assistance, where recalling previous exchanges can lead to more personalized and relevant interactions.

Additionally, efforts will be directed toward enhancing the human-like quality of the system's responses. The aim is to evolve the system to generate responses that are not only contextually appropriate but also exhibit coherent and naturally flowing speech. Advanced techniques in natural language generation and deep learning will be explored to fine-tune the nuances of the language used by the AI. By improving the linguistic style and response consistency, the system is expected to engage users in a manner that closely mirrors human conversational patterns, thereby enhancing user experience.

To further enhance the performance and adaptability of the model, future iterations will focus on training with more targeted data that features lesser diversity. This approach aims to refine the model's responses to be more domain-specific and accurate. Additionally, transitioning from using the distil version of GPT-2 to the full GPT-2 model is anticipated to leverage deeper neural networks for richer context capture and generation capabilities, potentially improving the sophistication and quality of interactions.

6 Conclusion

The aim of this project was to utilize advanced artificial intelligence, specifically the GPT-2 model, to enhance the utility of online product reviews by generating context-aware, insightful responses directly from user feedback. The empirical evaluations detailed within the study reveal that while the model exhibits competence in replicating lexical structures, as evidenced by reasonable BLEU scores, it struggles significantly with deeper semantic coherence and linguistic complexity, as highlighted by lower METEOR and ROUGE scores. Particularly, the inadequate performance in ROUGE-2 and ROUGE-L metrics underscores the model's difficulties in crafting extended, contextually cohesive text sequences, which are crucial for simulating the nuanced exchanges typical in effective customer service interactions. Although the project aligns with the initial objective to mitigate consumer overload by synthesizing and responding to extensive product reviews, the results indicate a pivotal need for further refinement in training methodologies and model architecture. Such enhancements are essential to adequately capture the complexity and diversity of human written reviews.

In conclusion, this project represents a significant step towards the integration of AI into consumer-facing applications, promising to augment interaction efficiency and user satisfaction by delivering prompt, relevant responses derived from vast user-generated content. However, the limitations identified call for a continued iterative approach to model development, incorporating more sophisticated natural language processing techniques and richer datasets.

References

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. pages 65–72, 2005.

Adhitya Bhawiyuga, M. Ali Fauzi, Eko Sakti Pramukantoro, and Widhi Yahya. Design of e-commerce chat robot for automatically answering customer question. pages 159–162, 2017. doi: 10.1109/SIET.2017.8304128.

Tom Bocklisch, Joey Faulkner, Nick Pawlowski, and Alan Nichol. Rasa: Open source language understanding and dialogue management. `https://rasa.com`, 2017. Accessed: 2023-04-17.

Tanja Bunk, Johannes E. M. Mosig, Vova Kozlov, Alan Nichol, and Joseph J. Godfrey. Diet: Lightweight language understanding for dialogue systems. *arXiv preprint arXiv:2004.09936*, 2020.

Lei Cui et al. Superagent: A customer service chatbot for e-commerce websites. *Proceedings of ACL, 2017, system demonstrations*, 2017.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. pages 4171–4186, 2019.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. pages 74–81, 2004.

Zan Mo, Yan-Fei Li, Peng Fan, et al. Effect of online reviews on consumer purchase behavior. *Journal of Service Science and Management*, 8(03):419, 2015.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. pages 311–318, 2002.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019a.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019b. Accessed: 2023-04-17.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Min Yang, Qiang Qu, Ying Shen, Qiao Liu, Wei Zhao, and Jia Zhu. Aspect and sentiment aware abstractive review summarization. pages 1110–1120, August 2018. URL `https://aclanthology.org/C18-1095`.