

# Question Answering System that would help the user decide on a product to buy.

By: Dina Tamer and Malak Amr

March 14, 2024

## 1 Introduction

Question answering is a critical NLP problem and a long-standing artificial intelligence milestone. QA systems allow a user to express a question in natural language and get an immediate and brief response. QA systems are now found in search engines and phone conversational interfaces, and they're fairly good at answering simple snippets of information. On more hard questions, however, these normally only go as far as returning a list of snippets that we, the users, must then browse through to find the answer to our question.

In response to this problem, our project endeavors to introduce a groundbreaking Question Answering System (QAS) empowered by the latest advancements in Natural Language Processing (NLP). Our system is designed to revolutionize the way consumers interact with e-commerce platforms, offering a personalized and intuitive approach to product discovery. By harnessing the power of NLP, our QAS aims to understand and interpret user queries in natural language, enabling seamless communication between the user and the system.

## 2 Motivation

The motivation behind our project stems from a deep-seated desire to alleviate the pain points commonly associated with online shopping, particularly the overwhelming nature of product selection. Traditional search engines and recommendation systems, while effective to some extent, often fall short in providing tailored recommendations that align with the unique preferences and requirements of individual users. Recognizing this gap, we seek to leverage the capabilities of NLP to develop a QAS that not only understands the nuances of user queries but also extracts relevant information from product descriptions, reviews, and specifications to deliver concise and informative responses.

Our objectives are multi-faceted, aiming to address various challenges inherent in the process of product selection. Firstly, we strive to develop a robust and efficient QAS capable of accurately interpreting user queries and providing relevant answers in real-time. Secondly, we aim to enhance the adaptability of our system by integrating machine learning techniques, allowing it to evolve and improve its understanding of user preferences over time. Lastly, we seek to empower consumers by offering a seamless and personalized shopping experience that simplifies the decision-making process and fosters confidence in their purchasing decisions.

Throughout this paper, we will delve into the methodology and architecture of our QAS, outlining the techniques and algorithms employed for natural language understanding, information extraction, and machine learning integration. We will also present the results of our system's performance evaluation, demonstrating its effectiveness in assisting users with product selection. By offering insights into the implications of our work and potential applications in the realm of e-commerce optimization, we aim to contribute to the advancement of consumer-centric technologies and pave the way for a more seamless and enjoyable online shopping experience.

### 3 Literature review

The literature review section is a crucial component of research, offering a comprehensive overview of existing knowledge and methodologies relevant to the development of a Question Answering System (QAS) aimed at aiding users in product purchasing decisions. In the realm of QAS, various techniques and approaches have been explored to enhance the user experience and decision-making process. From traditional rule-based systems to advanced machine learning models, researchers have delved into diverse strategies to improve the accuracy, efficiency, and user-friendliness of QAS for product recommendation. This literature review will traverse the landscape of methodologies, encompassing natural language processing (NLP), machine learning algorithms, sentiment analysis, and user modeling. By examining prior works, we aim to identify the strengths and limitations of each approach, paving the way for the synthesis of a novel QAS that excels in assisting users in making informed decisions when choosing products to purchase.

The paper [1] focuses on addressing the challenge of providing quick responses to user questions in e-commerce websites by automatically identifying plausible answers from product reviews. The motivation behind the study is to enhance user experience by reducing the waiting time for responses and improving the efficiency and trustworthiness of online shopping. To tackle this problem, the researchers proposed a novel multi-task deep learning method named QAR-net, which incorporates carefully designed attention mechanisms. This method leverages large-scale user-generated QA data and manually labeled review data to train an end-to-end deep model for answer identification in review data. The model consists of three sub-networks: Q-subnet, A-subnet, and R-subnet, which utilize Bidirectional Gated Recurrent Unit (Bi-GRU) for feature extraction and attention techniques for generating high-level embeddings of texts. Experiments conducted on data collected from Amazon demonstrated the effectiveness of QAR-net, showing that it outperformed baseline methods. The results of the implementation indicated that the proposed method successfully identified plausible answers from product reviews for user questions, showcasing its potential to improve the user experience in e-commerce platforms. In conclusion, the researchers found that QAR-net, the multi-task attentive model they developed, was effective in leveraging user-generated QA data to enhance question-answer matching in e-commerce settings. The study highlighted the importance of utilizing deep learning techniques and attention mechanisms to automate the process of extracting answers from reviews, ultimately providing users with timely and relevant information during their online shopping experience.

A survey paper on Product Question Answering system In [2] provides a clear framework for understanding the different approaches and methodologies employed in addressing product-related questions. This categorization allows for a detailed analysis of the methods and evaluation protocols used in each problem setting, offering insights into the strengths and limitations of various techniques in the context of PQA. Furthermore, the paper delves into the unique challenges that characterize PQA and differentiate it from traditional Question Answering (QA) systems. By focusing on the subjectivity, reliability, multi-type resources, and low-resource issues specific to PQA in E-Commerce platforms, the paper sheds light on the complexities involved in effectively answering product-related queries. Understanding these challenges is crucial for developing robust solutions that can handle the diverse nature of user-generated content and the dynamic landscape of E-Commerce platforms. By systematically analyzing existing research efforts and discussing potential future directions, the paper aims to provide a comprehensive overview of the state-of-the-art in PQA and guide researchers towards addressing key challenges in this domain.

In [3], they introduced a comprehensive framework for addressing user queries on e-commerce product pages, with a specific emphasis on handling both factoid and non-factoid questions. The framework is designed to leverage a combination of deep learning-based distributional semantics and structured semantics imposed by a domain-specific ontology. It encompasses various components such as question category classification, question-answer annotators, deep learning-based sentence embedding, and question-answer matching models. The proposed system aims to address the challenges of identifying question intent, product attribute-value extraction, semantic matching, and maintaining high precision in providing accurate answers. Additionally, the system is designed to overcome the scarcity of training data and other resources typically encountered in domain-specific question answering systems. The paper also includes an evaluation of different components of the framework and an ablation study to analyze their contribution to the system’s performance. The evaluation demonstrates that the proposed system achieves a 66% higher precision compared to a baseline model, indicating its effectiveness in addressing user queries on e-commerce product pages. Overall, the paper provides a detailed and systematic approach to developing a question-answering system for e-commerce product pages, highlighting the interplay of various components and their impact on system performance.

The article [4] delves into the intricate task of generating accurate and meaningful answers for product-related questions within e-commerce platforms. The researchers introduce the innovative Meaningful Product Answer Generator (MPAG) model as a solution to the challenges faced in this domain. The methodological approach employed in the study involves a comprehensive framework comprising a review clustering algorithm, a review reasoning module utilizing a write-read memory architecture, an attributes encoder based on a key-value memory network, a prototype reader for extracting answer skeletons, and an answer generator incorporating an editing gate to fuse reasoning results with product attributes dynamically. The evaluation of the MPAG model is conducted using various metrics, including BLEU for lexical unit overlapping, embedding-based metrics for assessing semantic similarity, and the distinct metric for evaluating answer diversity. The dataset utilized in the research consists of a substantial collection of information from 469,953 products spanning 38 categories, with each product having an average of 59.1 reviews and 9.0 attributes associated with it. The results of the experiments showcase the superior performance of the MPAG model compared to existing baselines, demonstrating its capability to consistently generate specific and proper answers for product-related questions in e-commerce settings. The conclusion emphasizes the model’s effectiveness in providing reasonable explanations for the generated answers, mitigating the safe answer problem, and enhancing reasoning abilities in generating meaningful responses, thereby establishing the MPAG model as a valuable tool for e-commerce question-answering tasks.

## 4 Methodology

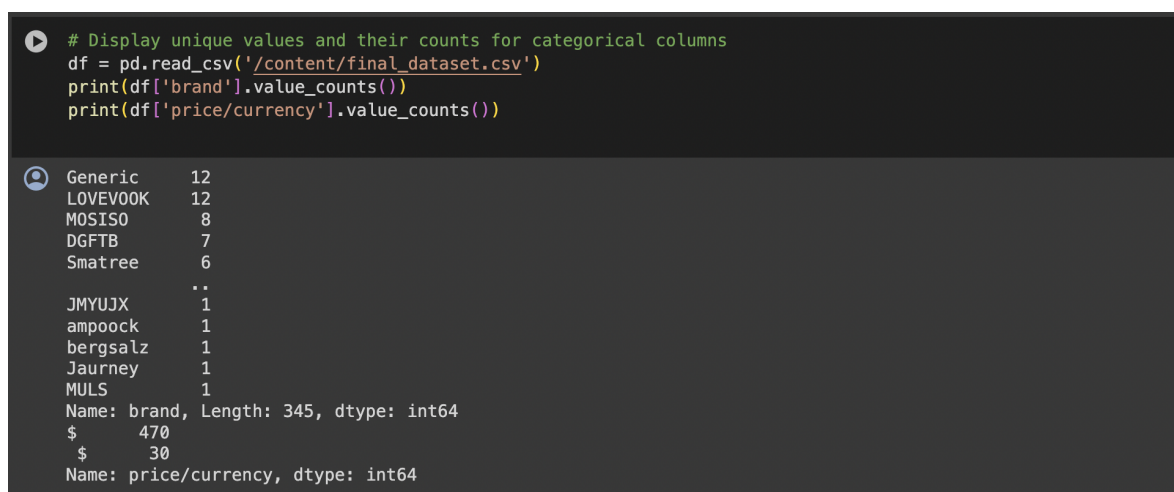
In the pursuit of creating an informed and data-driven decision-making tool for product selection, this study employs a comprehensive methodology centered around the analysis of Amazon's 500 Best-sellers in Laptop Gear for the year 2024. The dataset, meticulously curated from Kaggle, encompasses essential product attributes such as title, brand, description, currency, stars, and reviews count. Our methodology unfolds through a series of data analyses, each designed to extract valuable insights. Initial exploratory data analysis sheds light on the distribution of brands, star ratings, and reviews count, providing a holistic view of the dataset's landscape. Sentiment analysis, utilizing Natural Language Processing techniques, further enriches our understanding by categorizing product descriptions into positive, neutral, and negative sentiments. Subsequently, brand-centric analyses unravel the top-performing brands based on average star ratings, sentiment distribution, and total reviews count. These analytical layers collectively contribute to a nuanced understanding of the dataset, forming the foundation for robust decision support in the realm of consumer product selection. The detailed methodology outlined herein underscores our commitment to transparency, rigor, and the generation of actionable insights for end-users.

### 4.1 Data Preprocessing

Before conducting the data analysis, some data preprocessing had to be done on the dataset, as there were null values in several rows which might have caused inaccurate results. In preprocessing the dataset, several steps were undertaken to enhance data quality and completeness. Initially, missing descriptions were filled by leveraging the corresponding titles, as the titles section contain some sort of description of the product. Next, missing values in the price/value column were imputed using a K-nearest neighbors (KNN) approach, where the model was trained on the entire dataset to predict missing prices based on other relevant features. Similarly, missing stars ratings were imputed using KNN imputation, ensuring that the imputed values were rounded to two decimal places for consistency. Additionally, missing reviewsCount values were imputed using KNN imputation and subsequently rounded to the nearest integer. Finally, any missing entries in the price/currency column were filled with the string '\$'. These preprocessing steps collectively contributed to a more comprehensive and reliable dataset suitable for further analysis and modeling.

### 4.2 Data Analysis

After applying the data preprocessing, We started our analysis by extracting the main keywords in the products' descriptions. Figure 1 shows the output of this analysis.



```
# Display unique values and their counts for categorical columns
df = pd.read_csv('/content/final_dataset.csv')
print(df['brand'].value_counts())
print(df['price/currency'].value_counts())
```

brand	count
Generic	12
LOVEV00K	12
MOSISO	8
DGFTB	7
Smatree	6
..	..
JMYUJX	1
ampoock	1
bergsalz	1
Journey	1
MULS	1

Name: brand, Length: 345, dtype: int64

price/currency	count
\$	470
\$	30

Name: price/currency, dtype: int64

Figure 1: Descriptions Keywords

Then we checked the star ratings of all the products in the amazon dataset and here are the results. Figure 2 shows that the average star ratings of all products was 4.5.

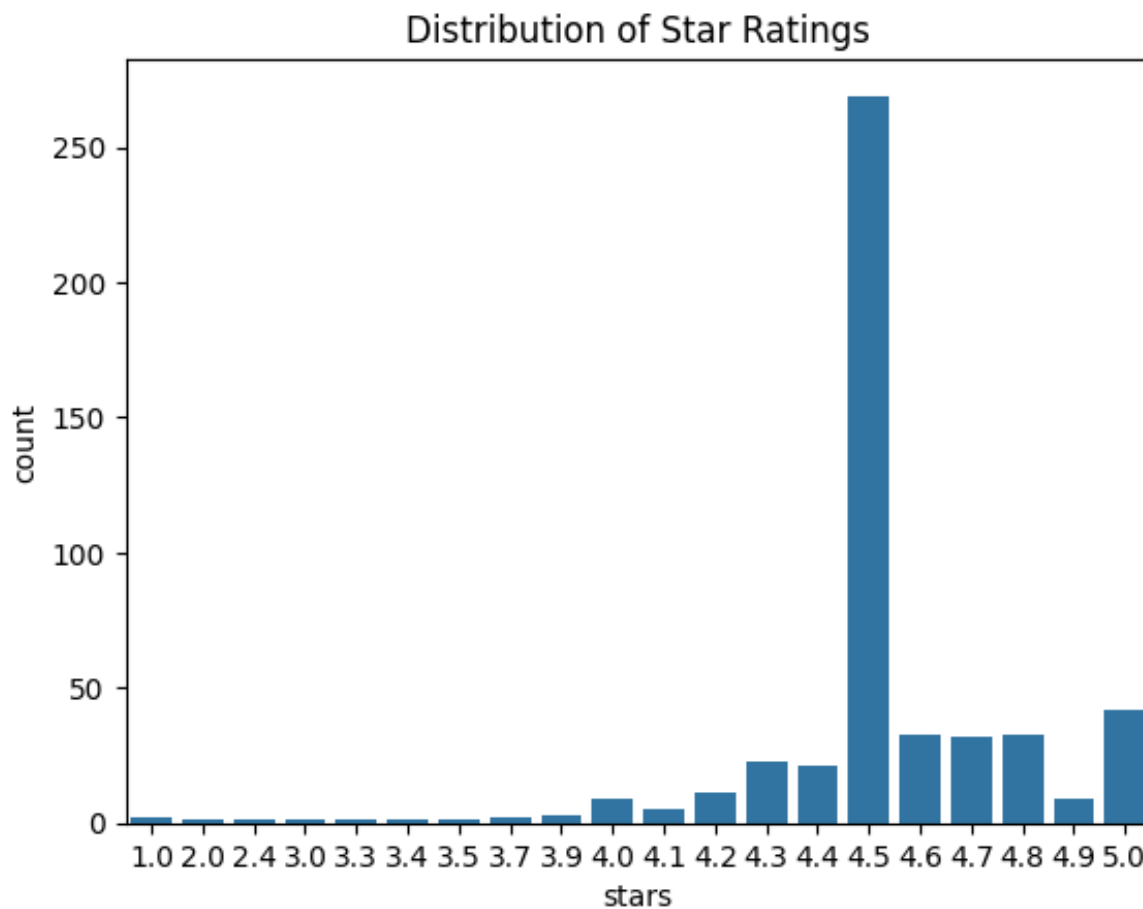


Figure 2: Star Ratings

Then we imported wordcloud to generate a word cloud of the most frequently used words in the description column. Figure 3 shows the output word cloud.



Figure 5 is another display of this sentiment analysis.

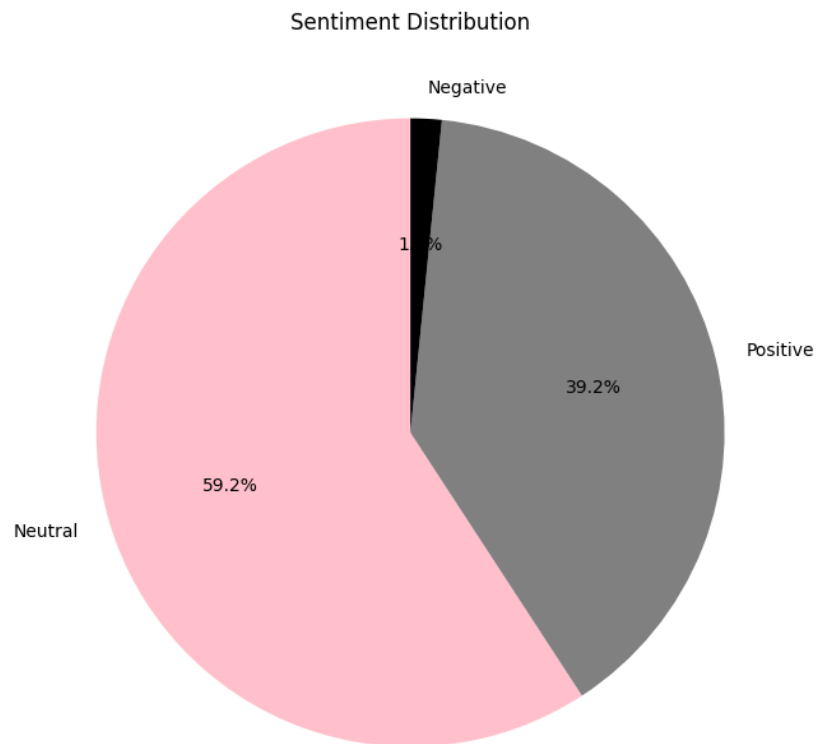


Figure 5: Sentiment Analysis pie chart

Moreover, we conducted a price analysis which shows us the frequency of the prices and the price distribution among all products. Figure 6 shows the output of this analysis.

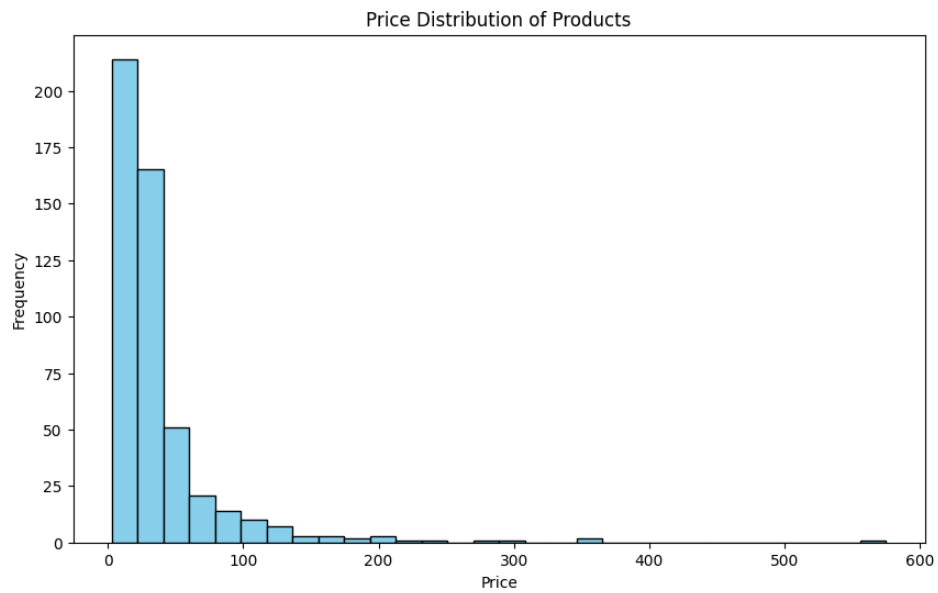


Figure 6: Price Distribution

We also conducted an analysis by grouping the brands and the number of descriptions of each brand. Figure 7 shows the output of this analysis.

```

brand
LOVEV00K      12
Generic       12
MOSISO        8
DGFTB         7
Smatree       6
...           ...
JGTM          1
JESW0         1
Inateck       1
Icicrim       1
wonfurd       1

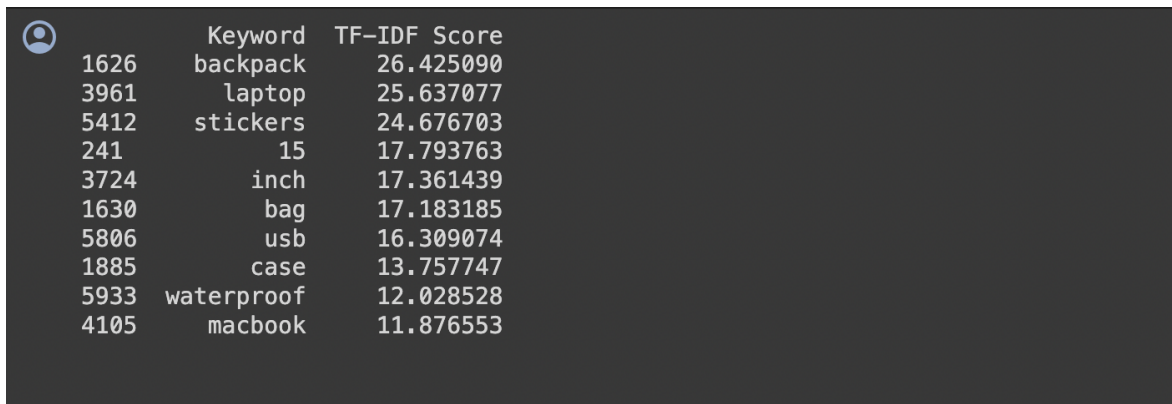
[345 rows x 1 columns]

```

Figure 7: Brand Grouping



lastly, TF-IDF was utilized to identify significant keywords from the dataset. TF-IDF assigns weights to words based on their frequency in individual product descriptions and rarity across all descriptions. This allowed us to extract essential terms that represent the products effectively. By calculating TF-IDF scores for each word, we pinpointed the most important keywords in the dataset. This approach enabled us to highlight key features of the products and gain insights into their characteristics, aiding in subsequent analysis and decision-making processes. Figure 8 shows the result of this analysis.

A terminal window with a dark background and light blue text. It displays a table of TF-IDF scores for various keywords. The table has three columns: a numerical value, a keyword, and the TF-IDF score. The keywords are listed in descending order of their TF-IDF scores.

	Keyword	TF-IDF Score
1626	backpack	26.425090
3961	laptop	25.637077
5412	stickers	24.676703
241	15	17.793763
3724	inch	17.361439
1630	bag	17.183185
5806	usb	16.309074
1885	case	13.757747
5933	waterproof	12.028528
4105	macbook	11.876553

Figure 8: TF-IDF

## References

- [1] L. Chen, Z. Guan, W. Zhao, W. Zhao, X. Wang, Z. Zhao, and H. Sun, “Answer identification from product reviews for user questions by multi-task attentive networks,” in *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-20)*. AAAI Press, 2020.
- [2] Y. Deng, W. Zhang, Q. Yu, and W. Lam, “Product question answering in e-commerce: A survey,” 02 2023.
- [3] A. Kulkarni, K. Mehta, S. Garg, V. Bansal, N. Rasiwasia, and S. Sengamedu, “Productqna: Answering user questions on e-commerce product pages,” in *WWW 2019 Workshop on ECNLP*, 2019. [Online]. Available: <https://www.amazon.science/publications/productqna-answering-user-questions-on-e-commerce-product-pages>
- [4] S. Gao, X. Chen, Z. Ren, D. Zhao, and R. Yan, “Meaningful answer generation of e-commerce question-answering,” *ACM Transactions on Information Systems*, vol. 39, no. 2, pp. 1–26, 2021.