Natural Language Processing and Information Retrieval (CSEN1076)

Abdelraheman Khaled Ali Asran Ahmed

Question Answering System for Product Recommendation

German University in Cairo

Contents

1 Introduction and Motivation

In our increasingly digital world, online shopping has become an integral part of our daily lives. Customers can now access a wide range of products from across the world with just a few clicks, making it simpler than ever to get exactly what they're looking for. But this wealth of options also brings with it a new set of challenges.

As users navigate e-commerce platforms like Amazon, they are often faced with a daunting task: deciding which product to purchase. With countless options available, it can be overwhelming to sift through product listings, compare features, and ultimately make a decision. This is where the concept of product recommendation comes into play.

Imagine you're in the market for a new laptop accessory, such as a mouse or a laptop stand. You may have specific requirements in mind, such as compatibility with your device or a certain price range. However, finding the perfect product that meets all your criteria can be like finding a needle in a haystack.

Many e-commerce platforms have implemented AI-driven conversational assistants, such as Alexa and AliMe, to address this challenge to help users navigate the online shopping experience. These assistants aim to provide personalized recommendations and answer user queries, guiding them towards products that best suit their needs.

One fundamental aspect of these conversational assistants is Product Question Answering (PQA), which involves automatically answering user-generated questions about specific products using natural language processing techniques. Unlike traditional question answering systems, PQA focuses on subjective opinions and preferences, making it a unique and challenging problem to solve.

This project aims to develop a Question Answering System for Product Recommendation, leveraging natural language processing (NLP) techniques. The system will analyze product descriptions, reviews, and other relevant data to provide personalized recommendations to users navigating e-commerce platforms like Amazon. The project is structured into three main phases: understanding the problem domain and conducting a literature review, building a neural network model, and fine-tuning the model with pre-trained models. The project's ultimate goal is to enhance the online shopping experience for users by simplifying the decision-making process and improving customer satisfaction.

2 Literature Review

Product Question Answering (PQA) is a specialized form of question answering that focuses on providing answers to queries related to products in the context of e-commerce and online shopping platforms. PQA is essential in the e-commerce space because it gives customers individualized and pertinent product information, which improves their shopping experience. Customers always have many questions and concerns about features, compatibility, user experiences, and other topics as they browse through huge selections of products and choices. PQA systems let users make informed decisions by automatically producing answers to questions submitted by users.

PQA systems can be categorized into several types based on the methodology employed to generate answers. These include:

1. **Opinion-based PQA**: Focuses on providing yes-no type answers to subjective questions by aggregating opinions from user-generated content such as product reviews.
2. **Extraction-based PQA**: Aims to extract specific text spans from supporting documents, such as product reviews or specifications, to serve as answers to user queries.
3. **Retrieval-based PQA**: Involves selecting the most relevant supporting documents from a candidate pool to answer user queries, typically by ranking documents based on their relevance to the query.
4. **Generation-based PQA**: Generates natural language responses to user queries based on the information available in supporting documents, often leveraging techniques such as sequence-to-sequence models.

Each type of PQA system has its own strengths and limitations, and the choice of approach depends on factors such as the nature of the questions, the availability of data, and the desired user experience.

2.1 Opinion-driven Product Query Resolution (PQR)

Opinion-driven PQR investigations concentrate on addressing binary inquiries prevalent within e-commerce platforms, where users frequently seek subjective viewpoints on products to guide their purchase choices. The primary aim of opinion-driven PQR is to deliver binary responses (Yes or No) to user-raised queries, typically derived from subjective expressions found in ancillary materials such as product reviews.

2.1.1 Datasets and Evaluation Mechanisms

Opinion-driven PQR systems frequently undergo evaluation using datasets like the Amazon Product Dataset, housing millions of answered queries and product reviews spanning diverse categories. Evaluation metrics such as Acc@k are commonly adopted to gauge performance, prioritizing the top k queries based on prediction confidence.

2.1.2 Approaches

In their pioneering work, McAuley and Yang (2016) introduced Mixtures of Opinions for Question Answering (Moqa), a system built upon the Mixtures of Experts (MoEs) model. Moqa treats each review as an "expert" providing a binary prediction, either "Yes" or "No", in response to a given product-related question. The confidence of each review's prediction is weighted based on its relevance to the question, allowing Moqa to aggregate opinions effectively.

Subsequent advancements by Wan and McAuley (2016) and Yu and Lam (2018b) further refined Moqa to address the nuances of ambiguity and subjectivity inherent in product-related questions. Wan and McAuley's improvements focused on enhancing Moqa's ability to handle uncertain or ambiguous answers, ensuring robust performance across diverse question types. Meanwhile, Yu and Lam's contributions involved incorporating aspect-specific embeddings into Moqa, enabling the model to capture nuanced relationships between questions and reviews. These enhancements underscored the ongoing efforts to optimize opinion-based PQA systems for real-world applications.

In parallel, recent studies by Fan et al. (2019) and Zhang et al. (2019) have explored the potential of neural network architectures and pre-trained language models in advancing opinion-based PQA. By leveraging neural networks, such as BiLSTM, and pre-trained language models like BERT, these approaches aim to learn more sophisticated feature representations from the available data. By capturing the intricate interplay between questions and reviews, neural network-based models have demonstrated superior performance compared to traditional methods, showcasing their potential to revolutionize opinion-based PQA.

A notable innovation in opinion-based PQA comes from Rozen et al. (2021) with their introduction of Similarity-Based Answer Prediction (SimBA). SimBA introduces a novel approach to leveraging existing data by exploiting similarities in resolved questions about similar products. By identifying relevant answers from past interactions, SimBA offers a promising avenue for improving the accuracy and efficiency of opinion-based PQA systems. This innovative method highlights the importance of harnessing the wealth of available data to enhance the performance of PQA systems in real-world scenarios.

2.1.3 Limitations

Opinion-driven PQR strategies offer a straightforward approach to tackling a substantial portion of product-related queries using relatively uncomplicated methodologies. However, these strategies may lack intricate query-specific details, focusing primarily on presenting the overall opinion polarity without deeper insights.

2.2 Extraction-based PQA

Extraction-based Product Question Answering (PQA) operates akin to traditional extraction-based QA methods, aiming to extract specific spans of text from supporting documents as answers to product-related questions. The objective is to identify a sequence of tokens within a given document that correctly addresses the question at hand.

2.2.1 Datasets and Evaluation Mechanisms

Several datasets have been curated specifically for extraction-based PQA research, each with its unique characteristics and evaluation protocols. Xu et al. (2019) introduced the ReviewRC dataset, constructed from SemEval-2016 Task5 reviews, while Gupta et al. (2019) developed the AmazonQA dataset, derived from the Amazon dataset, distinguishing between answerable and unanswerable questions based on review content. Bjerva et al. (2020) contributed the SubjQA dataset, focusing on the relationship between subjectivity and PQA across various domains. Evaluation metrics such as Exact Match (EM) and F1 scores are commonly employed to assess model performance.

2.2.2 Approaches

Methodologically, researchers have explored diverse approaches to address the challenges inherent in extraction-based PQA. Xu et al. (2019) utilized pretraining objectives like masked language modeling and next-sentence prediction to enhance BERT's performance on both general MRC and e-commerce review datasets. In contrast, Gupta et al. (2019) integrated information retrieval techniques to filter irrelevant reviews and built an answerability classifier to handle unanswerable questions, employing the R-Net model for span-based QA. Bjerva et al. (2020) introduced a subjectivity-aware QA model, leveraging multi-task learning to jointly tackle extraction-based PQA and subjectivity classification.

2.2.3 Limitations

Despite the potential of extraction-based PQA to provide precise answers, its practicality in real-world applications may be limited due to its less user-friendly nature and potential loss of additional contextual information, as discussed by McAuley and Yang (2016) and Deng et al. (2022). Consequently, the focus on extraction-based PQA in recent years has been relatively limited compared to other PQA paradigms.

2.3 Retrieval-based PQA

Retrieval-based Product Question Answering (PQA) approaches frame the task as a sentence selection problem, aiming to retrieve the most suitable answer from a pool of candidate sentences to effectively address the given question.

2.3.1 Datasets and Evaluation Mechanisms

The absence of ground-truth question-review (QR) pairs poses a challenge for retrieval-based PQA evaluation. While efforts have been made to annotate additional QR pairs into existing datasets, such as the Amazon dataset, original datasets can still be employed for evaluation. Standard ranking metrics like mean average precision (MAP), mean reciprocal rank (MRR), and normalized discounted cumulative gain (NDCG) are commonly used to assess model performance.

2.3.2 Approaches

Initial efforts, like the SuperAgent chatbot by Cui et al. (2017), employed multiple ranking modules to select answers from various data sources within product pages. Subsequent work by Kulkarni et al. (2019) introduced a pipeline system involving question classification and ensemble matching models for answer ranking. Recent approaches tend to focus on end-to-end models trained on limited sources, addressing the need for extensive annotated data.

To mitigate the low-resource challenge, transfer learning frameworks have been proposed, leveraging shared knowledge from large-scale datasets like Quora and MultiNLI. Mittal et al. (2021) introduced a distillation-based training algorithm using QA pairs retrieved by syntactic matching systems. Additionally, distant supervision paradigms and multi-task deep learning methods have been explored to train models using both user-generated QA data and manually labeled QR pairs.

Efforts to improve interpretability include identifying important keywords within questions and associating relevant words from QA pairs. Pre-trained

language models like BERT have been employed to obtain weak supervision signals from community QA pairs for measuring relevance between questions and heterogeneous information.

2.3.3 Limitations

Retrieval-based approaches offer complete and informative sentences as answers but may lack precision in addressing specific questions due to the general nature of supporting documents like reviews, which are not tailored for answering individual queries.

2.4 Generation-based PQA

Generation-based Product Question Answering (PQA) draws inspiration from sequence-to-sequence (Seq2seq) models applied in other natural language generation tasks. These models aim to automatically generate natural language sentences as answers to product-related questions.

2.4.1 Datasets Evaluation Mechanisms

The Amazon dataset and JD dataset are commonly used for generative PQA. The JD dataset, originating from one of China's largest e-commerce platforms, encompasses a vast array of products and categories, each associated with QA pairs, reviews, and product attributes. Evaluation of generation-based methods involves both automatic metrics like ROUGE, BLEU, Embedding-based Similarity, BertScore, and BleuRT, as well as human evaluation protocols assessing aspects such as fluency, consistency, and helpfulness.

2.4.2 Approaches

Generation-based PQA typically involves retrieving relevant documents as a preprocessing step before model development. To address noise in retrieved documents, approaches like Wasserstein distance-based adversarial learning and attention-based weighting strategies have been employed. Furthermore, considering the subjective nature of many product-related questions, some methods incorporate opinion mining into answer generation. Cross-passage hierarchical memory networks and heterogeneous graph neural networks are proposed to leverage information from diverse resources for answer generation. Additionally, efforts have been made to tackle issues like safe answer generation and benchmarking PQA over semi-structured data.

2.4.3 Limitations

Generation-based methods offer natural answers tailored to specific questions but face challenges like hallucination, factual inconsistency, and a lack of robust automatic evaluation protocols.

2.5 Conclusion

In conclusion, each type of PQA system has its strengths and limitations, and the choice of approach depends on factors such as the nature of the questions, data availability, and desired user experience. Future research efforts should aim to address the challenges inherent in each approach while further enhancing the accuracy and efficiency of PQA systems in real-world scenarios.

3 Dataset Overview

3.1 Introduction to the Dataset

The dataset that will be used for our analysis is "Amazon's 500 Bestsellers in Laptop Gear 2024". This carefully selected dataset offers a comprehensive view of the current trends and preferences within the laptop accessory market on Amazon. It provides insights into a wide array of products, ranging from essential electronic peripherals to novelty decals, reflecting the diverse interests and demands of consumers in the digital marketplace.

3.2 Source and Size of the Dataset

The dataset originates from Amazon and comprises a collection of the top 500 best-selling laptop accessories. It is expertly curated to capture the pulse of the market, offering a comprehensive snapshot of products resonating with customers on the platform.

3.3 Main Features and Variables

- **Title**: This feature captures the essence of each product, providing a concise description of its main attributes and functionalities.
- **Brand**: Identifies the producer or manufacturer of the product, which is essential for brand impact analysis and understanding market dynamics.
- **Description**: Offers detailed narratives about each product, providing rich textual data ripe for analysis. This feature enables deeper insights into product features, specifications, and unique selling points.
- **Price/Currency**: Indicates the currency in which the price is listed, providing essential information for market analysis and comparison across different regions or platforms.
- **Price/Value**: Provides economic data points that are crucial for market and sales analyses. Understanding price trends and variations can help identify pricing strategies and consumer preferences.
- **Stars**: Reflects customer satisfaction and perceptions of product quality. The star rating system, commonly used on e-commerce platforms like Amazon, offers valuable insights into consumer sentiment and product performance.
- **ReviewsCount**: Measures the level of engagement and popularity among consumers. A higher number of reviews typically indicates greater customer interest and interaction with the product.

3.4 Missing Data Analysis

3.4.1 Analysis Findings

An initial analysis of the dataset was conducted to gain insights and identify potential limitations. This analysis included examining missing values across various columns, which can provide valuable cues regarding data quality and completeness. The plot in Figure 1 illustrates that approximately 245 entries lack descriptions, potentially impacting the comprehensiveness of our analysis as detailed narratives provide valuable insights into product features, functionality, and consumer perceptions. Additionally, there are 30 missing values each in the 'price/currency' and 'price/value' columns, suggesting potential inconsistencies or errors in recording price information. Furthermore, around 241 entries are missing values for both 'stars' and 'reviewsCount' columns, indicating that these products may not have received any reviews yet.

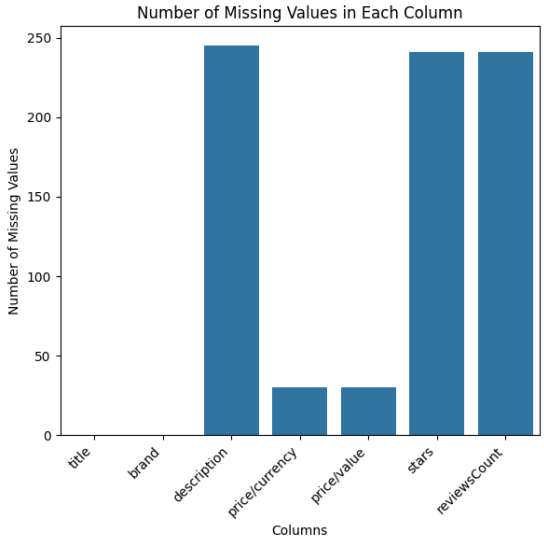


Figure 1: Number of Missing Values

To facilitate further analysis of the dataset, missing values were addressed by substituting them with appropriate placeholders. For the entries lacking descriptions, the placeholder "No description Available" was used. Similarly, missing values in the 'stars' and 'reviewsCount' columns were replaced with zeros. As only 30 rows were missing price information, these entries were dropped from the dataset

3.4.2 Limitations

- **Biased Reviews Data:** The missing values in the 'stars' and 'reviewsCount' columns suggest potential biases in the review data. Products without reviews may have different characteristics or may not have been as popular as those with reviews, leading to skewed perceptions of product satisfaction and popularity.
- **Incomplete Product Information:** The absence of descriptions for a significant number of entries limits the depth of our understanding of product features and attributes. This could introduce bias if certain types of products are more likely to have missing descriptions, affecting analyses that rely on detailed product information.
- **Price Data Reliability:** The missing values in the price-related columns raise concerns about the reliability of pricing data. Incomplete or inconsistent pricing information could lead to inaccurate analyses of pricing trends or comparisons between products.
- **Quality of Data Collection**: Overall, the presence of missing values across multiple columns indicates potential issues in data collection or recording processes. These inconsistencies may introduce biases and adversely impact the performance of models relying on this data.

3.5 Numeric Columns Analysis

3.5.1 Exploring Distributions of Numeric Columns

For the analysis of the numeric columns, we will begin by examining the trends within these columns. This exploration aims to uncover the distribution and fluctuations within the numerical data, shedding light on potential patterns. We will then explore the correlation between these numeric columns to understand the relationships and dependencies among different variables. This analysis will provide valuable insights into the underlying patterns and associations within the dataset, facilitating a deeper understanding of the factors influencing product features, pricing, customer satisfaction, and other key metrics.

- **Price/Value Distribution:** The histogram analysis reveals that the prices of laptop accessories vary widely, ranging from $2.99 to $575.00. The distribution is right-skewed, with a larger number of products having lower prices. Common price points cluster around values below $10.00, indicating that lower-priced accessories are more prevalent in the dataset. However, higher-priced accessories are also represented but less frequently.
- **Stars Distribution:** The stars column, representing customer satisfaction ratings, shows a diverse range of ratings from 0.0 to 5.0. The most

common rating is 0.0, showing that many products had missing values in the stars column. However, there are also numerous highly rated products with ratings above 4.0, indicating a mix of satisfaction levels among customers.

- **ReviewsCount Distribution:** The reviewsCount column, indicating the number of reviews received, exhibits a wide range of counts from 0 to 42821. Like the stars column, the most common count is 0, showing that many products had missing values in the reviews column. However, there are products with varying levels of review engagement, with some receiving a significant number of reviews. Overall, while some products have high engagement levels, others lack reviews, potentially impacting the reliability of customer feedback.

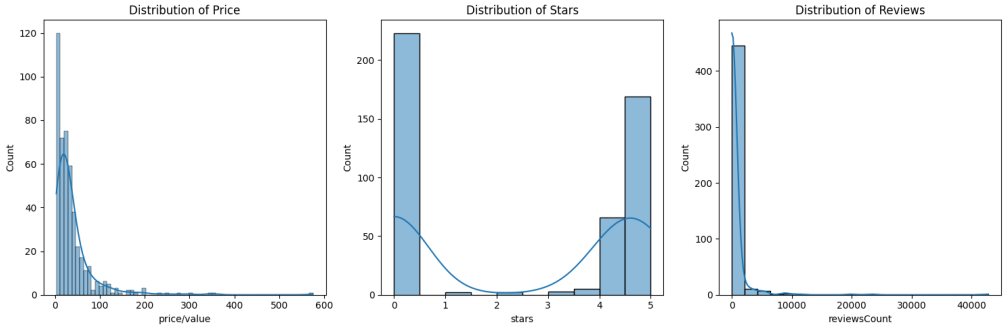


Figure 2: Numeric Columns Distributions

3.5.2 Correlation Analysis

The correlation analysis among the numeric columns in the dataset suggests several insights. Across all pairs of columns (price/value vs. stars, price/value vs. reviews count, stars vs. reviews count), there is more or less no correlation. This indicates that higher prices don't always equate to higher satisfaction, pricing has little impact on customer engagement measured by review counts, and there's no strong relationship between customer satisfaction ratings and the number of reviews.

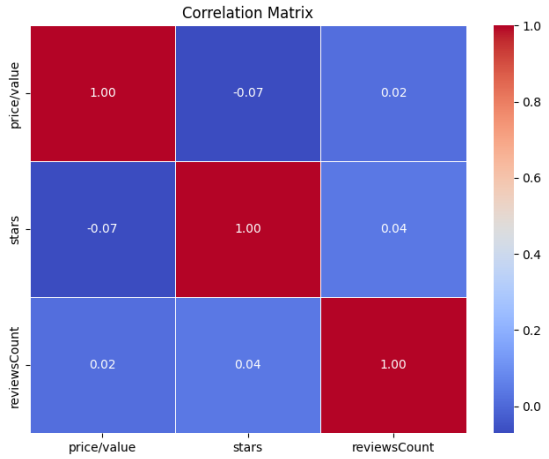


Figure 3: Correlation Matrix

3.5.3 Limitations

The presence of numerous products with zero ratings and reviews may skew perceptions of overall product satisfaction and popularity. Additionally, the right-skewed distribution of prices and review counts indicates potential biases towards lower-priced and less-reviewed products, impacting the generalizability of insights derived from the dataset. The observed correlations, or lack thereof, might stem from the dataset's narrow scope, which only includes the top 500 best-selling products. This selection bias could lead to weaker or negligible correlations since these popular products generally have lower prices and higher customer satisfaction ratings.

3.6 Textual Analysis of Product Titles, Brands, and Descriptions

This section explores the textual aspects of the dataset, focusing on the 'Title', 'Brand', and 'Description' columns. The goal is to uncover insights into the characteristics, features, and brand associations of the listed laptop accessories.

3.6.1 Brand Analysis

The dataset comprises 327 unique brands, reflecting a diverse range of options available in the laptop accessory market. Among these, the most frequent brands include "Generic" for unbranded products, along with "LOVEVOOK" each appearing 11 times in the dataset.

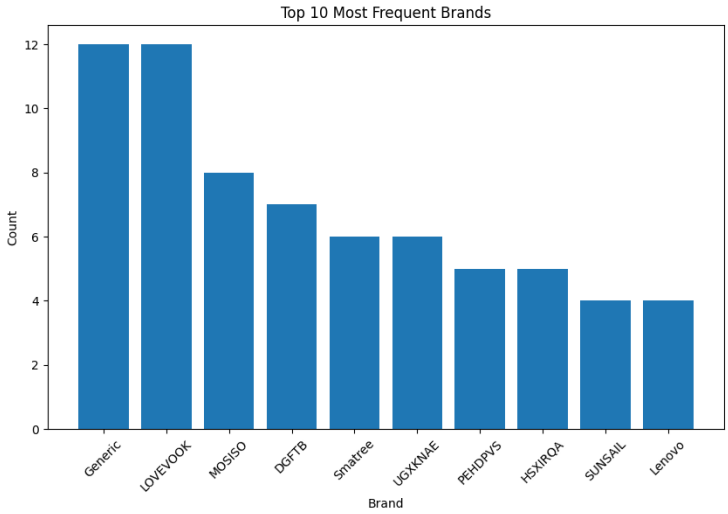


Figure 4: Top 10 Brands

The presence of "Generic" as a frequent brand may indicate a segment of unbranded or generic products. This may suggest a set of budget-friendly options or products sourced from various manufacturers without a distinct brand identity. However, it's unclear how these products compare to branded alternatives in terms of quality. Additionally, recommending products labeled as "Generic" may introduce ambiguity for users relying on the model's recommendations, as they may lack the assurance or familiarity associated with recognized brands. This ambiguity could affect user trust and satisfaction with the recommendation system, highlighting the importance of addressing brand diversity and clarity in product recommendations.

3.6.2 Title and Description Analysis

In this section, the focus is on exploring the title and description columns of the dataset, providing valuable insights into the characteristics and features of laptop accessories.

- **Basic summary statistics:** For titles, we observed a total word count of 13,107, with an average word length of approximately 5.06 characters. The character count for titles is 77,731, indicating concise descriptions. However, the relatively low sentence count of 15 suggests that titles are typically short phrases rather than complete sentences, which is expected given their purpose of providing concise product identifiers.

  On the other hand, the description column exhibits more extensive text content, with a significantly higher word count of 33,050. Despite the

higher word count, the average word length is slightly lower at approximately 4.78 characters, indicating a slightly more varied vocabulary compared to titles. The character count for descriptions is notably higher at 186,068, reflecting the additional detail and information provided in this column. Furthermore, the higher sentence count of 992 suggests that descriptions contain more complete sentences and detailed information about the products.

- **Word count distribution:** This section explores the distribution of word counts across titles and descriptions, shedding light on the common lengths and patterns observed in these textual elements. For titles, there is a prevalence of moderate-length titles, typically ranging from 16 to 29 words. This suggests that concise yet descriptive titles are common among laptop accessories listed on the platform. Longer titles, exceeding 29 words, are less frequent, indicating a preference for brevity in conveying product information.

  In contrast, descriptions exhibit a skew towards shorter lengths, with the majority falling within the range of 2 to 16 words. Notably, many of these concise descriptions correspond to products that lacked detailed descriptions and were replaced with the placeholder "No description available." This introduces a potential bias in the analysis, as a significant portion of products in the dataset lack genuine descriptions. Consequently, the true distribution of description lengths may be skewed, affecting the overall understanding of how product information is presented. Additionally, there is a notable decline in frequency for descriptions with word counts exceeding 16, indicating that longer descriptions are less prevalent.

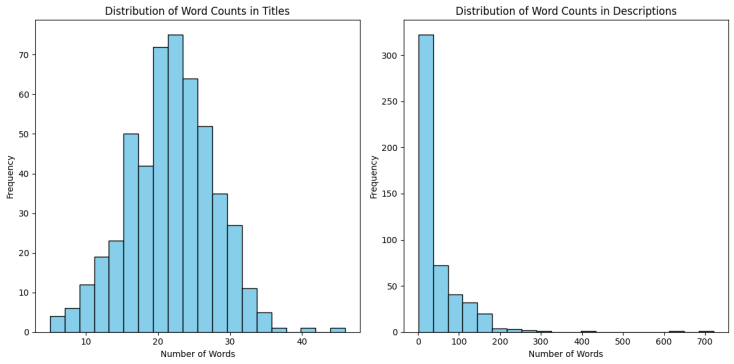
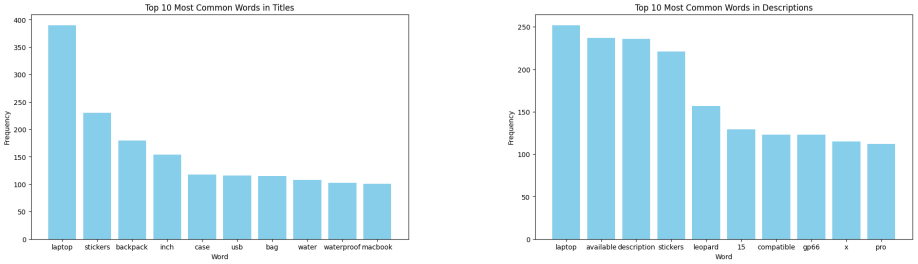


Figure 5: Word Count Distribution for Titles and Descriptions

- **Vocabulary Analysis:** The title vocabulary consists of 2837 unique words. The top 10 most frequent words include terms like "laptop", "stickers", "backpack", "inch", and "case", which are indicative of common laptop accessories. The high frequency of words related to specific products

such as "macbook", "waterproof", and "usb" suggests popular categories within the dataset.

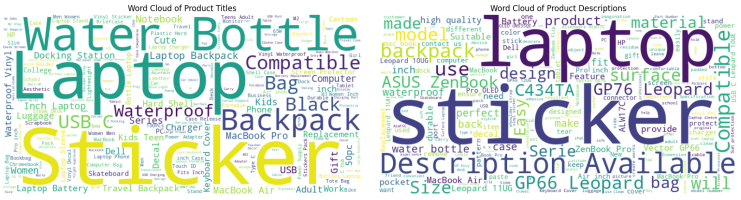
The description vocabulary is larger, comprising 5540 unique words. The top 10 most frequent words include terms like "laptop", "available", "description", "stickers", and "leopard". Interestingly, the word "available" appears frequently, possibly due to its use as a placeholder in descriptions. Other common words like "compatible", "pro", and specific product names like "gp66" and "15" reflect attributes and specifications mentioned in the descriptions



(a) Top 10 words in titles

(b) Top 10 words in descriptions

Figure 6: Comparison of top 10 words in titles and descriptions



(a) Title Word Cloud

(b) Description Word Cloud

Figure 7: Word Clouds of Titles and Descriptions

The most frequent words in both titles and descriptions are related to product features, types, and specifications commonly associated with laptop accessories. Certain words like "available" in descriptions may not contribute much semantic meaning but are included due to their use as placeholders. The vocabulary size in descriptions is larger than that in titles, indicating a greater diversity of language used to describe products and their attributes