# Natural Language Processing Report 1

Mostafa Sherif and Mahmoud Moamen

March 14, 2024

**Abstract**

This report presents an in-depth analysis of Amazon's 500 Bestsellers in Laptop Gear dataset for the year 2024, focusing on extracting valuable insights into consumer preferences, sentiments, and market trends within the laptop accessory market. The study begins with an exploration of recent literature, encompassing a comprehensive review of existing research and methodologies employed in understanding similar e-commerce datasets. Following this, the report delves into the data analysis phase, where various statistical techniques and visualization methods are applied to uncover patterns, correlations, and key drivers influencing consumer behavior and product popularity. The report also presents actionable insights for businesses and stakeholders, shedding light on strategic opportunities, challenges, and potential areas for improvement within the laptop gear market. Additionally, the report discusses the limitations inherent in the dataset and analysis methodology, providing a balanced perspective on the findings presented. Overall, this report serves as a valuable resource for decision-makers seeking to navigate the dynamic landscape of e-commerce.

## 1 Introduction

In the contemporary marketplace, the role of data-driven decision-making has become increasingly pivotal for businesses striving to maintain competitiveness and meet evolving consumer demands. Amidst the burgeoning landscape of e-commerce platforms, Amazon stands as a cornerstone, offering a vast array of products ranging from essential electronics to niche accessories. Within this spectrum, the realm of laptop gear emerges as a focal point, reflecting both utility and personal expression in the digital age.

### 1.1 Motivation

The motivation behind this analysis lies in the recognition of the immense value encapsulated within consumer data, particularly within the context of Amazon's Bestsellers in Laptop Gear dataset for the year 2024. By harnessing the insights embedded within this dataset, we embark on a journey to unearth the underlying trends, sentiments, and preferences of consumers towards an assortment of laptop accessories and peripherals.

At the heart of this endeavor lies the pursuit of actionable intelligence - the extraction of insights that can steer strategic decision-making, foster product innovation, and cultivate a customer-centric approach within the competitive e-commerce landscape. Understanding the nuanced interplay between consumer sentiment, product attributes, and market dynamics holds the key to unlocking untapped opportunities, fostering brand loyalty, and driving sustainable growth in an ever-evolving digital ecosystem.

Through analysis and interpretation, we endeavor to empower businesses and stakeholders with the tools and knowledge needed to navigate the complex terrain of the laptop accessory market with confidence and foresight.

# 2 Literature Review

This section will discuss three main points were taken as criteria for searching which are Sentiment Analysis for E-Commerce Products using Natural Language Processing, Sentiment Analysis for Customers' Reviews and Dataset Research on E-Commerce products which shows the recent work done for utilizing natural language processing in the E-Commerce specially the sentiment analysis that would help in the upcoming milestones related for the Amazon dataset.

## 2.1 Sentiment Analysis for E-Commerce Products using Natural Language Processing

The paper titled "Natural Language Processing for Sentiment Analysis in E-Commerce Products" (1) explores the application of sentiment analysis in evaluating consumer sentiments towards products and services, particularly within the e-commerce domain. In recent years, the growth of e-commerce has been exponential, with more consumers turning to online platforms for their shopping needs. This trend underscores the importance of understanding customer opinions and preferences for enhancing online businesses. Sentiment analysis, a branch of natural language processing (NLP), offers a valuable tool for extracting insights from textual data, such as customer reviews and social media posts.

The authors of the paper propose the utilization of PySpark and Spark NLP to address challenges related to real-time data collection and analysis. Traditional sentiment analysis methods often struggle with the large volumes of data generated by e-commerce platforms. PySpark and Spark NLP provide scalable solutions for processing such data efficiently, enabling businesses to derive actionable insights in a timely manner.

In the introduction, the authors underscore the growing significance of online shopping and the pivotal role sentiment analysis plays in comprehending consumer perspectives. They highlight the multifaceted nature of consumer feedback, which encompasses textual reviews, star ratings, and emoticons. Sentiment analysis techniques aim to categorize this diverse range of data into meaningful insights that businesses can use to improve their products and services.

The paper delves into previous studies on sentiment analysis, highlighting diverse techniques and methodologies employed in prior research endeavors. Traditional approaches include rule-based systems, machine learning algorithms, and hybrid models combining both. While these methods have shown promise, they often struggle with the nuances of human language and context, leading to suboptimal results.

The proposed system architecture comprises two primary phases: data collection and real-time analytics. The authors delineate the methodology, encompassing data preprocessing, sentiment sentence extraction, and feature selection through Spark NLP and PySpark. By leveraging these tools, businesses can streamline the sentiment analysis process and extract meaningful insights from large volumes of textual data.

The results and discussion segment present the outcomes of the proposed system, including tokenization, preprocessing, the visualization of common keywords, sentiment classification utilizing WordCloud visualization, and an analysis of frequency distribution based on reviewer age and sentiment scores. These analyses provide valuable insights into consumer preferences and sentiments, enabling businesses to tailor their offerings to better meet customer needs.

In conclusion, the authors summarize the key findings of their study and emphasize the efficacy of the Spark NLP technique in sentiment analysis. They suggest future research avenues, such as investigating deep learning approaches for sentiment analysis and relationship analysis. The growing availability of data and advances in NLP technology offer exciting opportunities for further research in this field.

## 2.2 Sentiment Analysis for Customers' Reviews

The paper titled "Sentiment Analysis of E-commerce Customer Reviews Based on Natural Language Processing" (2) by Xiaoxin Lin explores the application of machine learning algorithms to analyze customer sentiment in e-commerce clothing reviews. As the e-commerce landscape continues to evolve, businesses are increasingly turning to customer reviews to gauge product satisfaction and identify areas for improvement. Sentiment analysis techniques offer a systematic approach for analyzing these reviews and extracting valuable insights.

The study aims to understand the correlation between review features and product recommendations using natural language processing (NLP). Five popular machine learning algorithms, namely Logistic Regression, Support Vector Machine (SVM), Random Forest, XGBoost, and LightGBM, are employed for sentiment analysis. These algorithms offer different trade-offs in terms of computational complexity, model interpretability, and predictive performance.

The research is based on a dataset obtained from Kaggle consisting of Women's E-Commerce Clothing Reviews, with various features such as title, rating, recommendation indicator, and positive feedback count. The text data undergoes vectorization using the TF-IDF algorithm to enable machine learning analysis. TF-IDF, short for Term Frequency-Inverse Document Frequency, is a common technique for converting textual data into numerical representations suitable for machine learning algorithms.

The paper discusses each algorithm in detail, explaining their theoretical foundations, parameter settings, and optimization techniques. Logistic Regression, SVM, Random Forest, XGBoost, and LightGBM are evaluated based on metrics like accuracy, precision, recall, F1 score, and Area Under Curve (AUC). These metrics provide a comprehensive evaluation of each algorithm's performance across different dimensions.

Results show that LightGBM achieves the best performance among the algorithms, with the highest accuracy and AUC value. Ridge Regression, Linear Kernel SVM, and XGBoost also demonstrate competitive performance. Conversely, SVM with the RBF kernel exhibits the lowest accuracy. Comparison with other studies using similar datasets reveals that the addition of XGBoost and LightGBM algorithms enhances prediction accuracy.

The paper concludes by emphasizing the potential for further refinement of sentiment analysis through advanced NLP techniques and improved model training. It suggests strategies such as refining text preprocessing, distinguishing between rating and recommendation options, and exploring additional machine learning algorithms to enhance accuracy and comprehension of customer sentiment in e-commerce reviews. By leveraging these strategies, businesses can gain deeper insights into customer preferences and enhance the overall shopping experience.

### 2.3 Dataset Research on E-Commerce products

The paper "Dataset of Natural Language Queries for E-Commerce" (3) introduces a comprehensive dataset containing 3,540 natural language queries pertinent to e-commerce, particularly focusing on product searches for laptops and jackets. In the era of big data, datasets play a crucial role in driving research and innovation across various domains. However, existing datasets often lack detailed natural language information, limiting their utility for specific applications such as product search in e-commerce.

The dataset is obtained through controlled experiments involving participants with varied domain knowledge. It includes annotations for vague terms and key product features, specifically focusing on laptop queries. The dataset is positioned as a valuable resource for advancing research in natural language processing and interactive information retrieval within the domain of product search.

Furthermore, the paper explores various potential applications of the dataset. It discusses how the dataset can facilitate the development of spelling correction models to enhance natural language processing tasks. Additionally, it highlights the dataset's utility in addressing vocabulary mismatch issues by identifying vague expressions in queries. The dataset's annotations also aid in attribute mapping, enabling the matching of unstructured query information with structured product attributes. Moreover, the dataset can be leveraged for product query classification, contributing to the development of more sophisticated algorithms.

In conclusion, the paper emphasizes the dataset's significance in advancing research in natural language processing and e-commerce. It outlines future plans to expand the dataset, including additional annotations, clean versions of queries, and the incorporation of more product domains to enhance the generalizability of models derived from it. By continuously updating and improving the dataset, researchers can ensure its relevance and usefulness in addressing emerging challenges in e-commerce and natural language processing.

# 3 Data Analysis: Transforming, Visualizing, and Providing Insights

To start our analysis, we started by importing essential libraries such as pandas, numpy, matplotlib, seaborn, and nltk. These libraries provided robust tools for data manipulation, visualization, and natural language processing. The dataset was then loaded into a pandas DataFrame, enabling efficient data handling and exploration. The uploaded data was then printed in order to validate that the import was successful.

N.B.: The Dataset provided with a notebook (4) that helped as a reference in some cases

## 3.1 Data Transformation

Data transformation is a fundamental step in the data analysis process, encompassing various techniques aimed at preparing raw data for analysis and interpretation. This phase involves converting data into a structured format conducive to analytical processing, ensuring its quality, consistency, and suitability for further exploration. Common data transformation techniques include handling missing values, standardizing data formats, normalization, and encoding categorical variables. By employing appropriate data transformation methods, analysts can mitigate data quality issues, enhance the effectiveness of analytical models, and derive actionable insights that drive informed decision-making. Ultimately, data transformation plays a pivotal role in unlocking the full potential of datasets, enabling stakeholders to extract valuable insights and derive tangible business value from their data assets.

### 3.1.1 Handling Missing Values

An essential aspect of data pre-processing is handling missing values. Through a systematic assessment, the presence of null values in the dataset was identified. This critical step provided insights into the completeness of the dataset and guided subsequent data preprocessing steps to ensure data quality and reliability. To address missing values in key columns such as 'price/value', 'stars', and 'reviewsCount', an imputation technique leveraging the KNNImputer was employed. This technique facilitated the estimation of missing values based on the values of neighboring data points, thereby preserving the integrity of the dataset. Additionally, the 'price/currency' column underwent standardization, enhancing data consistency and facilitating further analysis. Finally, the 'description' column had a lot of missing values. However, after careful studying, it was concluded that the 'title' and 'description' columns were usually very similar. As a result, rather than dropping the empty 'description' rows, a new column 'text' was created by concatenating the 'title' and 'description' columns, facilitating textual analysis.

### 3.1.2 Normalization and Tokenization

In the process of normalization and tokenization, we standardized and prepared the textual data from the 'text' column for further analysis. This involved several key steps aimed at enhancing the quality and structure of the text data to facilitate computational analysis. Firstly, we employed lemmatization to reduce words to their base or root form, ensuring consistency in the representation of words. Additionally, we removed stopwords and punctuation from the text, eliminating irrelevant or redundant information that could distort analysis results. Finally, as the columns 'title' and 'description' were similar, duplicated words were removed for every row. By standardizing the text data through normalization and tokenization, we created a structured and uniform 'tokens' column suitable for sentiment analysis and textual exploration. This preprocessing step was essential in laying the groundwork for subsequent analytical tasks, enabling us to derive meaningful insights from the textual content of the dataset.

### 3.1.3 Sentiment Analysis

Leveraging the TextBlob library, sentiment analysis was conducted to measure the polarity of text data. By assigning sentiment scores to each text entry in a 'sentiment_score' and categorizing sentiments as 'Positive', 'Negative', or 'Neutral' into a 'Sentiments_labels' column, valuable insights into the sentiment of the text data assigned to each product were obtained.

### 3.1.4 Categorizing the Products based on Star Ratings

The categorization of data based on star ratings into 'stars_category' provided further granularity to the analysis. By breaking down reviews into 'Bad Review', 'Average Review', or 'High Review' categories, distinct patterns and trends in product ratings were obtained. This categorization enriched the analysis by offering actionable insights into product performance and would later be used to be compared to the sentiment of the provided text.

## 3.2 Data Visualization

Visualization served as a powerful tool for conveying insights derived from the dataset. Through various visualization techniques, complex data patterns and relationships were derived. The visualizations facilitated the interpretation of analysis results and can enable stakeholders to grasp key findings at a glance.

### 3.2.1 Top 5 Brands with Positive/Negative/Neutral Sentiments

This visualization provides insights into sentiment created by different brands. By identifying the brands with the highest count of positive, negative, and neutral sentiments, businesses can understand which brands are resonating positively or negatively with customers. This can inform brand management strategies, product development decisions, and marketing efforts. The tables can be seen in figures 1,2, and 3 respectively.

| | | text |
|---|---|---|
| brand | Sentiments_labels | |
| Generic | Positive | 10 |
| LOVEVOOK | Positive | 5 |
| PEHDPVS | Positive | 5 |
| SlimQ | Positive | 4 |
| Twelve South | Positive | 4 |

Figure 1: Top 5 Brands with Positive Sentiment

| | | text |
|---|---|---|
| brand | Sentiments_labels | |
| LOVEVOOK | Negative | 7 |
| MOSISO | Negative | 6 |
| Smatree | Negative | 6 |
| AMCJJ | Negative | 3 |
| VNINE | Negative | 2 |

Figure 2: Top 5 Brands with Negative Sentiment

| brand | Sentiments_labels | text |
|---|---|---|
| UGXKNAE | Neutral | 5 |
| DGFTB | Neutral | 5 |
| Espacio | Neutral | 4 |
| ANVMSRO | Neutral | 3 |
| dokikalos | Neutral | 3 |

Figure 3: Top 5 Brands with Neutral Sentiment

### 3.2.2 Categorizing Products based on Sentiment Labels

Figure 4 shows the pie chart that was employed to provide a comprehensive overview of the distribution of sentiment labels across the dataset. This visualization offers insights into the prevalence of positive, negative, and neutral sentiment categories, enabling stakeholders to gauge the overall sentiment landscape. Understanding the distribution of sentiment labels helps in assessing overall customer satisfaction levels and identifying areas for improvement. By pinpointing the prevailing sentiment categories, businesses can prioritize initiatives to address customer concerns, enhance product features, and improve the overall customer experience.

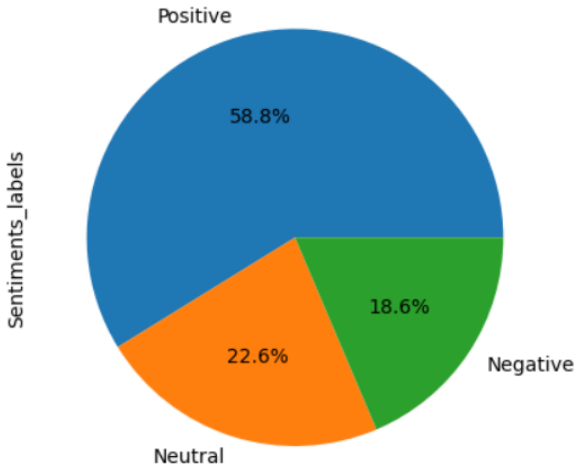


Figure 4: Pie Chart of Sentiment Categorization

### 3.2.3 Grouping the Products based on their Prices and Sentiment Labels

The grouping of data based on the 'sentiment_labels' and 'price/value' columns facilitated an exploration of the relationship between sentiment and product pricing. This visualization allowed businesses to discern how pricing influences customer sentiment and perception of product value. By analyzing the intersection of sentiment labels and price/value ranges, businesses can identify optimal pricing strategies that resonate well with customers. This insight is invaluable for pricing optimization, competitive positioning, and maximizing profitability in the market.

### 3.2.4 Displaying the Words based on their Frequency

A word cloud, shown in figure 5, was generated to visually represent the most frequently occurring words in the text data. This visualization offers a quick and intuitive way to identify prominent themes, keywords, and topics within the dataset. By visualizing the most common words, businesses can gain insights into customer preferences, product features, and emerging trends. This aids in understanding the key drivers of customer sentiment and guiding strategic decision-making to address customer needs effectively.

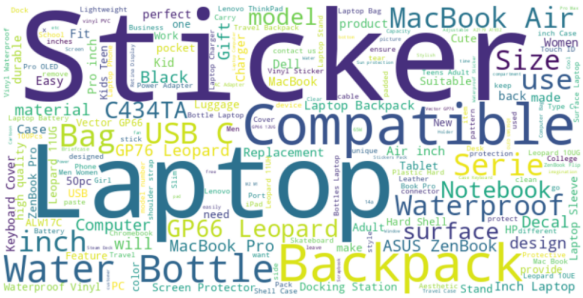


Figure 5: Word Cloud portraying word frequency

### 3.2.5 Grouping of the Products based on their Rating Categories and Sentiment Labels

The grouping of data based on sentiment labels and star rating categories allowed for an examination of the relationship between sentiment and product ratings. This visualization provided insights into how sentiment aligns with product ratings and whether positive or negative sentiment influences star ratings. By analyzing sentiment trends alongside star ratings, businesses can identify factors driving customer satisfaction and dissatisfaction, informing product improvements and customer service initiatives. These visualizations collectively offer a comprehensive view of customer perceptions and behaviors, empowering businesses to make informed decisions and enhance the overall customer experience. As shown in 6, it was clear that positive sentiment in the text provided with a product improves the chance of a good product rating.

| | | text |
|---|---|---|
| Sentiments_labels | stars_category | |
| Positive | High Review | 268 |
| Neutral | High Review | 106 |
| Negative | High Review | 84 |
| Positive | Average Review | 25 |
| Negative | Average Review | 8 |
| Neutral | Average Review | 7 |
| Negative | Bad Review | 1 |
| Positive | Bad Review | 1 |

Figure 6: Table of products grouped by the sentiment labels and stars rating

# References


[1] B. K. Jha, G. Sivasankari, and K. Venugopal, "Sentiment analysis for e-commerce products using natural language processing," *Annals of the Romanian Society for Cell Biology*, pp. 166–175, 2021.

[2] X. Lin, "Sentiment analysis of e-commerce customer reviews based on natural language processing," in *Proceedings of the 2020 2nd international conference on big data and artificial intelligence*, pp. 32–36, 2020.

[3] A. Papenmeier, D. Kern, D. Hienert, A. Sliwa, A. Aker, and N. Fuhr, "Dataset of natural language queries for e-commerce," in *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, pp. 307–311, 2021.

[4] A. KOCADINÇ, "Amazon$_p roduct_c omment, https://www.kaggle.com/code/ahmetkocadinc/amazon-product-comment?kernelsessionid = 162602975,''$