



Article

Enhancing Product Design through AI-Driven Sentiment Analysis of Amazon Reviews Using BERT

Mahammad Khalid Shaik Vadla ¹, Mahima Agumbe Suresh ²  and Vimal K. Viswanathan ^{1,*} ¹ Mechanical Engineering Department, San Jose State University, San Jose, CA 95192, USA; khalidvadla@gmail.com² Computer Engineering Department, San Jose State University, San Jose, CA 95192, USA; mahima.agumbesuresh@sjsu.edu

* Correspondence: vimal.viswanathan@sjsu.edu; Tel.: +1-(408)-924-3841

Abstract: Understanding customer emotions and preferences is paramount for success in the dynamic product design landscape. This paper presents a study to develop a prediction pipeline to detect the aspect and perform sentiment analysis on review data. The pre-trained Bidirectional Encoder Representation from Transformers (BERT) model and the Text-to-Text Transfer Transformer (T5) are deployed to predict customer emotions. These models were trained on synthetically generated and manually labeled datasets to detect the specific features from review data, then sentiment analysis was performed to classify the data into positive, negative, and neutral reviews concerning their aspects. This research focused on eco-friendly products to analyze the customer emotions in this category. The BERT and T5 models were finely tuned for the aspect detection job and achieved 92% and 91% accuracy, respectively. The best-performing model will be selected, calculating the evaluation metrics precision, recall, F1-score, and computational efficiency. In these calculations, the BERT model outperforms T5 and is chosen as a classifier for the prediction pipeline to predict the aspect. By detecting aspects and sentiments of input data using the pre-trained BERT model, our study demonstrates its capability to comprehend and analyze customer reviews effectively. These findings can empower product designers and research developers with data-driven insights to shape exceptional products that resonate with customer expectations.

Keywords: BERT; T5; natural language processing; content analysis; customer requirements



Citation: Shaik Vadla, M.K.; Suresh, M.A.; Viswanathan, V.K. Enhancing Product Design through AI-Driven Sentiment Analysis of Amazon Reviews Using BERT. *Algorithms* **2024**, *17*, 59. <https://doi.org/10.3390/a17020059>

Academic Editors: Frank Werner and Fabrizio Marozzo

Received: 30 November 2023

Revised: 11 January 2024

Accepted: 25 January 2024

Published: 30 January 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In the research and development cycle of a product, product design firms consider many aspects of developing their product as the best in the market to bring the best sales and profits and to give intense competition in terms of design, quality, and efficiency of the product in their respective industries. Collecting customer feedback review data to extract customer expectations on the products they purchase is one of the most significant aspects every industry has considered in recent years, especially in product design. Predicting consumer expectations on products and driving customer satisfaction is incredibly significant for any industry to succeed in an open market. The product development project collects customer requirements through different programs such as interviews, consumer surveys, and detailed market monitoring [1]. These collected customer data are analyzed and taken as feedback to work on the product design to improve its quality as per customer needs [2]. This kind of automated or manual analysis to extract the customer expectations on a product will help the manufacturing, retail, and e-commerce industries improve their research and development in product design [3].

The growth of individuals' willingness to purchase eco-friendly products is evidence that supporting the demand for eco-friendly products is extortionate and there is growth of ecologically favorable consumers [4]. A report by BBMG [5], a branding and social

impact consultancy in New York City, reveals that 70% of Americans know the significant role of eco-friendly products in controlling the climate crisis due to carbon footprints, and 51% of them are ready to pay even more for eco-friendly products due to their benefits. High sales of any product result in the decline of its manufacturing costs. To promote the sales of eco-friendly products, companies must consider various aspects such as consumer satisfaction and customer feedback on products.

Content analysis is a popular qualitative research technique used for analyzing data in textual form, such as surveys, interview transcripts, and online reviews of products and services. Traditionally, it is a labor-intensive manual qualitative research method that is categorized into four attributes such as “objectiveness, systematicity, quantitiveness, and manifestness” [6]. This is an easily understood analysis process that can be adopted even for those new to consumer-based product analysis [7].

2. Background

This section of the paper will explore different approaches to performing content analysis on customer feedback and prior approaches that predict and identify the sentiments of online review text data. Conventional content analysis is performed in a study to interpret the meaning of text data [8]. In this analysis, coding categories are derived directly from the text data. Human experts read the data word by word to derive the codes by first capturing the exact words from the text that highlight the key thoughts or concepts. Then, these codes are categorized into providing the meaning of the text [9]. Another successful analysis is performed on the raw data collected from transcribed interviews. This qualitative content analysis works on four principles (condensation, code, category, and theme). The text data are shortened while preserving the core meaning of the data. Most exact condensed text is labeled into codes. The codes that are related to each other are gathered to form a category. The theme is given for each category by doing further manual analysis to express the underlying meaning of the category [10]. In a variety of inductive approaches to analyzing qualitative data, thematic content analysis is performed in this research by analyzing the data transcripts, figuring out the identical themes within the data, and combining them. Data were generated from an interview with a child in a public health study to understand children’s knowledge of healthy foods. Researchers extracted all the words and phrases from the interviews and then, by performing thematic content analysis on these data, generated a list of categories that children like about the food [11].

However, qualitative data derived from surveys, interviews, and written open questions and pictures are expressed in words. Only statistical analysis generated meaning for the given data, which is not sufficient to predict customer opinions; other methods were required to analyze the data with great accuracy [7]. In conventional market research, customer feedback and its attributes are obtained through extended interviews and surveys, leading to an increase in the research time and the cost of the research project [12].

Manual data collection models are beneficial for indicating the significance of different product features concerning customer needs. However, these models need prominent skilled employees; only customer information is extracted from small categories. Hence, this type of data collection causes economic hardship and consumes much time in the product development cycle [13]. Due to the shifting of shopping habits, such as purchasing from in-store to online, there is a requirement for advanced methods to collect online data. When a customer purchases a product from an online store, they will provide feedback on that product in the form of a customer review. Online reviews became more prominent as people consider these comments before purchasing. Gathering the text review information manually will take considerable time, and sometimes manual data extraction will mislead the correct information. As online shopping is gaining popularity with each passing day and driving a more significant number of customers to purchase through online shopping, many researchers have attempted to leverage the customer requirement data available in the form of online product reviews. If one can extract useful information on customer

needs from these online reviews, manual methods such as resource-intensive customer surveys can be avoided in the product design cycle [14].

In recent years, deep learning concepts such as convolutional neural networks (CNN) and NLP have made breakthroughs in many fields, such as object recognition, data mining, and image processing. The researchers have used these data mining and natural language processing concepts to develop algorithms to collect online data. Enhancements in the applications of deep learning and the contribution of engineers to the field of AI have made the product designers' jobs more simple by enabling them to evaluate the sentiments of customers on a product, observe the various product trends in the market, examine the design of new successful products, and to recognize where the contemporary design opportunities evolve [13].

By using core deep learning concepts such as sentiment analysis to predict the sentiment of the data, it is straightforward to extract customer opinions in a fraction of a second. In the literature, there are several studies that used NLP methods to generate data-driven decisions. For example, Kim et al. [15] worked on research that conducted an experiment on how to deploy life cycle assessment combined with data mining and the interpretation of online reviews to enhance the design of more sustainable products. In this approach, customer reviews are extracted using the NLP pipeline to divide these reviews into two clusters. Then, ABSA is performed on those clusters that potentially contain relevant sustainability-related data by identifying the product features connected with sustainability-related comments. These classified customer opinions will be helpful to obtain meaningful sustainable design insights into eco-friendly products in complement to the Life Cycle Assessment (LCA) results.

A research study by Saidani et al. [16] aimed to improve industrial ecology and the circular economy by enhancing eco-design practices. Text mining analysis is conducted on the definitions of circular economy, industrial economy, and eco-design. Textalyser and Wordle are the online tools used to perform text mining to identify further and discuss common themes or disparities. In another study [12], online smartphone reviews and manual product documents distributed by manufacturers were collected as data. These data are preprocessed and lemmatized using an NLP toolkit that provides the semantic characteristics of each word. Lemmatized form and parts of speech (POS) tagging are implemented in this study. Using word2vec, all the words from online review data and product manuals are embedded into vectors. These embedded phrases were clustered, and each cluster was labeled with the most frequent word. Product designers select important product features after this cluster labeling. This sub-feature analysis will improve the design implication to the embodiment design. However, this sub-feature analysis focused on smartphone data; it can be utilized to improve the design of eco-friendly products. A customer network is developed based on customer attributes. Customer attributes were extracted from the customer reviews, and product features were extracted from the online data of 58 mobile phones. In this method, customer attributes were divided into positive and negative features. For example, a consumer satisfied with the product's feature (F) and another customer complaining about it were classified. The similarity between the customer opinions is measured by two indices: 1. cosine similarity and 2. topic similarity. This method develops a customer network if the customer meets the criterion threshold. Each cluster represents a segment, and by segmentation and analysis, the segment's characterization features and sentiments are predicted. This approach suggests that product designers focus on customer-oriented products [16]. Mokadam et al. [14], a study on eco-friendly product reviews, applied ABSA to determine customer needs. Twelve aspects were generated using key terms and keywords in the raw text data, each with one sentiment category. They trained an aspect detection model that learns sentence similarities between manually generated aspects and the review text to predict the aspect's presence in the review. The intensity of the sentiment expressed in a sentence is generated by VADER, a rule-based model for sentiment analysis, which outputs results in terms of positive and negative comments.

Previously, methodologies in natural language processing leaned heavily on word embeddings, such as word2vec and GloVe, for training models on vast datasets. Yet, these embeddings have limitations tied to their reliance on various language models and occasional struggles in extracting nuanced word contexts. To overcome these drawbacks, the NLP research landscape witnessed the emergence of expansive language models such as BERT, GPT, LLAMA, BART, and T5.

These large-scale language models have redefined the approach to tackling specific tasks within NLP, ranging from named entity recognition to question answering and sentiment analysis to text classification. Their advent represents a shift toward more comprehensive and versatile models capable of grasping intricate linguistic nuances, thereby advancing the efficacy of natural language understanding and analysis.

GPT

The Generative Pre-Trained Transformer (GPT) signifies a series of neural language models developed by OpenAI, functioning as generative models adept at sequential text generation, akin to human-like progression, word by word after receiving a initial text prompt. Its autoregressive text generation approach uses the transformer architecture, employing decoder blocks.

The GPT series, comprising models pre-trained on a combination of five datasets encompassing Common Crawl, WebText2, Books1, Books2, and Wikipedia, is distinguished by an impressive parameter count of 175 billion. Like BERT, GPT models exhibit the capability for fine-tuning specific datasets, enabling adaptation of their learned representations. This adaptability empowers GPT to excel in diverse downstream tasks, including text completion, summarization, question answering, and dialogue generation. This flexibility positions GPT as a versatile tool in natural language processing applications [17].

LLAMA

LLAMA (Language Model for Language Model Meta AI) is a comprehensive framework developed explicitly for research purposes, aiming to evaluate and compare existing language models across diverse natural language processing (NLP) tasks. Its creation assisted AI and NLP researchers in conducting standardized assessments of various large language models, comprehensively analyzing their capabilities, strengths, and weaknesses.

Trained across a spectrum of parameters spanning from 7 billion to 65 billion, LLAMA incorporates an array of evaluation metrics and tasks. These metrics encompass facets such as generalization, robustness, reasoning abilities, language understanding, and generation capabilities of language models. By encompassing such a wide range of assessments, LLAMA offers a holistic approach to evaluating and comparing the performance and functionalities of language models within the NLP domain [18].

BARD

Bayesian Adaptive Representations for Dialogue (BARD) is a sizable language model crafted by Google, primarily targeting computational efficiency within transformer-based NLP models. Designed to minimize computational demands, BARD underwent training on an estimated data volume exceeding a terabyte. Through tokenization, it dissects prompts and adeptly grasps semantic meanings and inter-word relationships via embeddings.

BARD's notable attributes are its cost effectiveness and adaptability, representing a model well suited for various tasks. Moreover, its unique capabilities, such as accessing the internet and leveraging current online information, position BARD ahead compared to ChatGPT chatbot. This amalgamation of computational efficiency and access to real-time information endows BARD with distinct advantages in dialogue-based language models.

T5

The Text-to-Text Transfer Transformer (T5), an encoder–decoder model introduced by Google, explicitly targets text-to-text tasks. Trained on a blend of supervised and unsupervised datasets, T5 operates via the encoder–decoder transformer architecture, akin

to other transformer-based models. T5 has two sides, just like a coin. The encoder analyzes your text, understanding its meaning and relationships between words. The decoder then crafts a response, weaving words together to form something new—such as a summary, a translation, or even dataset labeling. Its spectrum of tasks encompasses text summarization, text classification, and language translation, processing textual inputs to generate textual outputs.

Diversifying its capabilities, T5 spans various iterations, including the small, base, and large models, and T5-3b and T5-11b, each trained with parameter counts ranging from 3 billion to 11 billion. This expansive pre-training contributes to T5's adaptability and ease of fine tuning for domain-specific NLP tasks, marking it as a versatile and customizable model for various text-based applications [18].

BERT

In natural language processing (NLP), Google's groundbreaking Bidirectional Encoder Representations from Transformers (BERT) model has emerged as a pivotal framework. BERT's advancement lies in its pre-training phase, where it ingests an extensive corpus comprising 2500 million unlabeled words from Wikipedia and an additional 800 million from the Book Corpus. This pre-training methodology involves unsupervised learning tasks that entail predicting masked words within sentences and discerning logical coherence between pairs of sentences. The essence of BERT's innovation is its bidirectional context understanding; it comprehends context from both the left and right sides of words within sentences. Leveraging its pre-training on colossal text datasets, BERT exhibits adaptability by fine-tuning domain-specific datasets for distinct tasks. This fine-tuning process involves initializing the BERT model with pre-learned parameters tailored explicitly to labeled downstream tasks, such as sentiment analysis.

BERT's significance in the NLP landscape is underscored by its robust pre-training on extensive textual resources, enabling it to grasp nuanced linguistic contexts. Moreover, its adaptability via fine tuning on domain-specific datasets is a testament to its versatility across various NLP applications, positioning BERT as a cornerstone framework in language understanding and analysis.

Several recent studies have showcased the efficacy of BERT in diverse applications within natural language processing (NLP). Zhang et al. [19] developed the BERT-IAN model, specifically enhancing aspect-based sentiment analysis on datasets related to restaurants and laptops. Their model, utilizing a BERT-pretrained framework, separately encodes aspects and contexts. Through a transformer encoder, it adeptly learns the attention mechanisms of these aspects and contexts, achieving sentiment polarity accuracies of 0.80 for laptops and 0.74 for restaurants. Tiwari et al. [20] proposed a multiclass classifier leveraging the BERT model on SemEval 2016 benchmark datasets. Their approach effectively predicted sentiment and aspects within reviews, employing multiple classifiers combining sentiment and aspect classifiers using BERT. In a study by Shi et al. [21], the utilization of BERT in market segmentation analysis based on customer expectations of product features was explored. By deploying a BERT classification model on apparel product data, the analysis focused on understanding user attention toward specific product features such as pattern, price, cases, brand, fabric, and size. This analysis provided valuable insights for companies adapting to the apparel market. Tong et al. [22], applied a BERT-based approach to comprehending contextual elements (Task, User, Environment) from online reviews. Context of Use (COU) is pivotal in successful User Experience (UX) analysis. Their study extracted COU elements using the BERT model on a scientific calculator dataset, elucidating the tasks associated with the product and the diverse user types and environments in which these calculators are utilized.

The literature found in [12,14,16] revealed that the ABSA and NLP computational mechanisms can still be exploited for text mining and automation of content analysis. Collectively [20–22], these studies underscore the versatility and effectiveness of BERT in various NLP domains, from aspect-based sentiment analysis to understanding customer reviews. Based on the advantages of BERT, we propose an aspect-based sentiment analysis

approach to automate the extraction of customer expectations from online reviews to provide valuable data-driven insight to product designers.

The objective of this paper is to construct an NLP pipeline that automates the extraction of customer expectations from Amazon online reviews across various product categories. This will involve implementing aspect-based sentiment analysis to predict and categorize the aspects discussed within the reviews, while simultaneously classifying their sentiments.

3. Material and Methods

3.1. Manual Content Analysis

“Environmentally friendly products” was selected as a category for the analysis. This category was selected because of its niche nature and increasing popularity in online retail platforms. Once the product selection is completed by following the given flowchart criteria in Figure 1, reviews of the products are collected. These products must meet another criterion: each contains at least ten reviews on Amazon.com. Vendor-verified purchases are used to avoid biased reviews. Reviews that only describe the characteristics and features of products, such as design, quality, price, etc., were considered for this study. Data about customer emotions such as package and delivery issues are avoided because this model aims to discuss the product features. Reviews are selected based on how well they reflect the measures and are gathered to reflect the overall rating distribution. Once the data are gathered, content analysis is performed to determine the trends and aspects within the reviews. This is a supervised approach performed by human experts. This analysis begins with review collection to maintain the distribution of the ratings; at least ten reviews are collected semi-randomly for each product that is picked for the analysis. For example, a reusable home cleaning kit is picked from Amazon.com, and reviews are selected randomly to match the average distribution to the ratings of that product given by Amazon. These product reviews were distributed by 60% 5-star ratings, 14% 4-star ratings, 13% 3-star ratings, 5% 2-star ratings, and 6% 1-star ratings. According to this rating distribution ratio, high priority was given to 5-star ratings, and four 5-star rating reviews occupied a place in the ten reviews, followed by two 4-star reviews, two 3-star reviews, one 2-star review, and one 1-star review, to balance the portion.

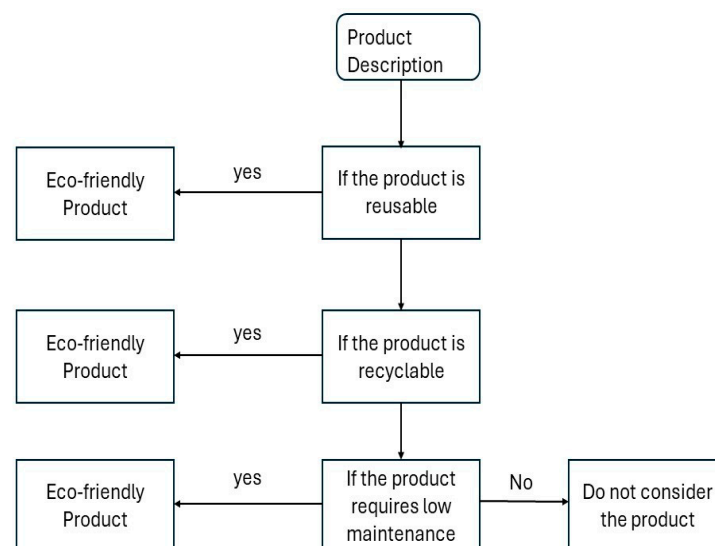


Figure 1. Product selection flowchart used for the manual content analysis.

Then, the content analysis is performed on the gathered review data to determine the trends and aspects within the reviews. Once the content analysis is performed on the data, the data analysis, called the quantification process, begins. In this process, the data are classified into bins and similar keywords. Product reviews are classified into aspects that reflect identical customer review data sentiments during this analysis. Reviews with

similar sentiments are filtered, the aspects with similar sentiments are gathered, and these data are categorized. The established data and their categories are finalized with identical keywords, and the aspects that define them are similar. In analyzing large datasets, these lists of categories of relevant words and aspects are used as a manual guide for the human operator while performing the analysis to extract the sentiments of the data. The aspect list of the 65 environmentally friendly products was prepared at the end of the successful analysis, and a labeled dataset with 1443 reviews was generated. This dataset is used to tune both BERT and T5 models. Table 1 shows the list of aspects and keywords identified in the analysis.

Table 1. User sentiment categories (aspect) and related keywords.

Number	Designation	Aspect	Example of Keywords
1		Aesthetics	Crisp, Beautiful, Wrinkled
2		Ease of Reprocessing	Wash, Clean, Charge
3		Durability	Wear, Died, Break, Resistant
4		Use Efficiency	Time, Fast, long
5		Performance	Hold, Well, Glitch
6		Adaptability	Versatile, Outside, Suitable
7		Ergonomics	Comfortable, Easy, Awkward
8		Ease of Storage	Store, Fold, Small
9		Ease of Use	Use, Easy, Convenient
10		Interference	Loud, Taste, Smell
11		Safety	Safe, Drop, Burn
12		Price	Expensive, Cheap, Cost

3.2. Synthetic Data

Data are crucial in achieving great accuracy in any existing NLP model. The performance and efficiency of the model depends on clean and preprocessed data. Training an NLP classifier requires a large amount of labeled data, and preparing massive, labeled datasets is challenging and time consuming. For a successful NLP model, the data quality plays a significant role. GPT-3.5 is used to generate synthetic data. By using three different prompts, synthetic data are generated using ChatGPT.

Prompt-1

In the first stage, a prompt is given to GPT to generate review data related to eco-friendly products and label the review with one of the 12 aspects extracted from manual analysis.

The prompt used to generate the data is here “I need to create some reviews for these cases: ‘Adaptability’, ‘Durability’, ‘Ease of Use’, ‘Ergonomics’, ‘Interference’, ‘Performance’, ‘Use Efficiency’, ‘Aesthetics’, ‘Ease of Reprocessing’, ‘Ease of Storage’, ‘Price’, ‘Safety’ as the aspects.

The reviews can belong to various products from eco-friendly product category such as phone, air frier, beauty products, home appliances, etc. I don’t have time to collect the reviews from the Amazon website so I will need you to synthesize them for me. While creating these reviews please consider the fact the reviewers have different backgrounds and not every review will be a proper review. I need reviews that look realistic. considering this please create some reviews for the case of “Efficiency” make sure the reviews don’t overlap with other cases mentioned in the beginning. Please don’t give obvious looking review with the keyword in it. It should convey that the review is about the case not explicitly mentioning it. Also, I just need the review not the product detail generate some examples from the standpoint of someone who is a adult person or a middle aged person what I mean is you should consider a wide demographics. keep a mix of positive negative and neutral. give me 100 examples. don’t put serial number and don’t put the examples in quotes”.

Using this prompt, almost 4000 reviews were generated for the 12 aspects.

Prompt-2

To prevent repetition and duplicate reviews, GPT is guided to focus on specific aspects by adjusting the prompts. Keywords extracted from manual content analysis, as seen in Table 1, are provided for each aspect, modifying the prompt accordingly. This approach encourages GPT to create diverse, unique data while avoiding similar reviews and eliminating duplicates. About 2000 reviews were generated using these tailored prompts. Additionally, eco-friendly product names such as trash bags, travel kits, yoga gear, bottles, glassware, pet toys, socks, portable campfires, desks, and tables were incorporated into these tailored prompts to further enrich the generated data.

Prompt-3

GPT is guided by real-time review phrases related to different aspects to diversify and enhance generated content. These phrases, extracted from 400 reviews across 12 aspects, provide valuable learning material for GPT to understand user expression patterns in online reviews. This customized prompt, tailored to each aspect, enables GPT to capture and emulate user emotions effectively, yielding over 1200 reviews that aim to be more realistic and user specific. This approach enriches learning for those seeking to understand how sentiments are conveyed in online reviews.

3.2.1. Aspect Detection Training Using BERT

Once we have generated data, we split them into two parts: 80% for training and 20% for testing. We've created three datasets to improve our BERT model to detect the aspects of the reviews. Then we will choose the best performing model with the best results to detect the aspect once it is trained it on the below datasets.

Synthetic Data (SD): we have divided the generated data into an 80:20 ratio for training and testing purposes.

Synthetic Data and Manual Analysis (SD and MA): synthetic data are used to train the model and manual analysis data to test its performance.

Combined Data (COMB): this dataset combines both synthetic and manual data in a balanced way. For training, we have gathered 80% of reviews from both datasets. We have reserved 20% of the reviews from each dataset to test the model and combine them.

These three datasets were used to train the BERT model and accuracy results of the model for the datasets shown in Table 2.

Table 2. Accuracy of the BERT model for three datasets.

Datasets	Reviews	Accuracy
SD	7217	73
SD and MA	8660	58
COMB	8660	91

The BERT-based model, initialized for sequence classification, effectively identifies aspects within review data. The datasets for training and testing undergo preparation, involving the removal of duplicate rows and handling missing values exclusively within both datasets. Each aspect in both datasets is numerically encoded (0–12) by mapping the original aspects to facilitate model comprehension. *Tokenization* is pivotal in utilizing the BERT tokenizer on textual review data. This transformation converts the text into numerical representations. Employing unique token IDs, the tokenizer assigns each word in the text, enabling the BERT model to recognize them. The textual reviews are transformed into tokenized sequences, incorporating padding, truncation, and a maximum sequence length of 512.

These tokenized sequences are organized within a labeled dictionary, constituting the dataset for training and validation purposes. A trainer initialized with the BERT model and training arguments oversees the training process on these datasets. The model learns to predict aspects within textual data by associating labels with their respective numerical values (0–12) and aligning similar words in the review through mapped token sequence IDs. The model undergoes pre-training on the training dataset and subsequent evaluation on the test dataset.

Evaluation metrics, computed through accuracy, precision, recall, and F-1 score, gauge the model's performance by evaluating the training data with test data. This BERT model, pre-trained on feature-specific data for aspect prediction, includes the saved BERT tokenizer for future utilization in training and detecting aspects within real-time review data, a crucial step in the subsequent stages of this research.

3.2.2. Aspect Detection Training Using T5

The Text-to-Text Transfer Transformer (T5) is initialized to detect the aspects. The COMB dataset is used for training and testing following a 80:20 ratio. To ensure data cleanliness, the text data undergo preprocessing steps such as lowercasing, stripping, and duplicate removal. The SimpleT5 and T5-Base library is employed to initialize and fine tune a T5 model for aspect analysis tasks. T5 operates on a text-to-text framework, where the input and output are in text format. For aspect detection, the input text is the review. This review is split into smaller units, called tokens, the model's essential input elements. These tokens are often words or characters. Each token is converted into high-dimensional numerical embeddings representing the token's semantic meaning within the text context (aspect). T5 employs a transformer architecture comprising encoder and decoder layers. The encoder understands the input text's context while the decoder generates the output. Transformers use self-attention to weigh the importance of each token about others, capturing long-range dependencies and context within the text. During training, the T5 model learns from input–output pairs of review data, where the input contains review text and the output includes aspects. Fine tuning adjusts the model's weights to specifically detect aspects within review data, refining its ability to identify and extract meaningful elements. When we provide a real-time review, the fine-tuned T5 model predicts or identifies aspects within the text based on its learned representations and associations. Metrics such as accuracy, precision, recall, and F1-score are computed to evaluate the effectiveness and performance of the model.

The best-performing model among T5-small and T5-base is saved for future use and to compare the results with the BERT model.

3.3. Aspect Detection Model

An NLP pipeline was explicitly developed for Aspect-based Sentiment Analysis, focusing on online review data obtained from Amazon through web scraping scripts. The BERT model, pre-trained and adept at understanding sentence similarities, was utilized to predict the presence of aspects within the reviews. This model was fine tuned to learn the contextual relationships between labels and the textual content, thereby inferring the existence of aspects within the reviews. The input review data underwent thorough text preprocessing to ensure optimal analysis, including removing punctuation, emojis, and stop words. This preparation step aimed to refine the text for more effective processing by the model. The pre-trained BERT model for sequence classification and its accompanying BERT tokenizer were incorporated within the prediction pipeline as shown in Figure 2. These components, essential to the pipeline, enabled the identification of aspects within new review data about a particular product. The pipeline's classification model predicted the reviews' aspects by matching tokenized words with the original aspects, leveraging their corresponding numerical values. By comparing and aligning these values, the model accurately inferred the presence and nature of the aspects within the review texts.

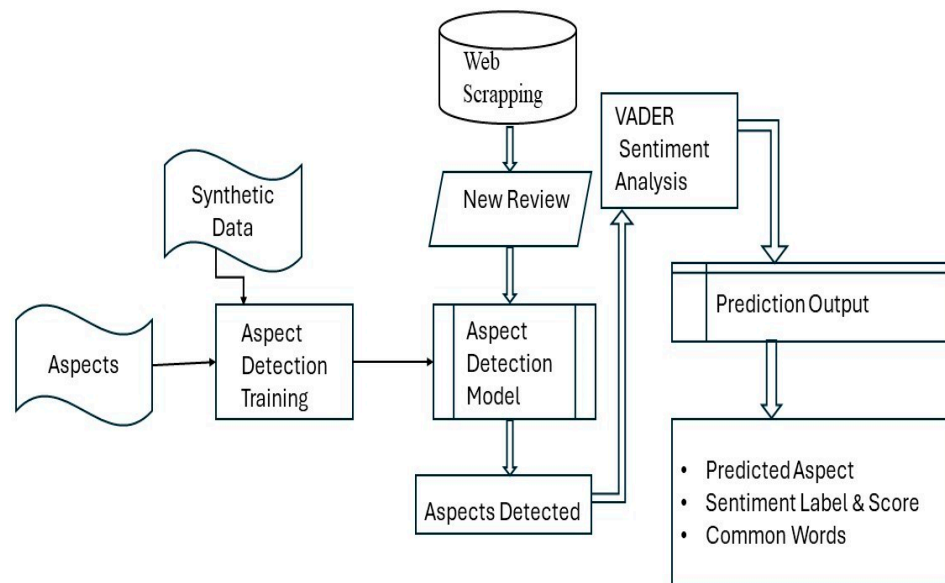


Figure 2. Block diagram of the prediction pipeline; the double arrow lines refer to the steps at the time of prediction.

3.4. Sentiment Analysis

The review datasets undergo normalization processes such as tokenization, stemming, and lemmatization to automate the classification of reviews into positive, negative, or neutral sentiments. These techniques aim to standardize textual data before sentiment analysis is performed. VADER, a lexicon and rule-based tool renowned for predicting sentiments on social media and e-commerce platforms, analyzes sentiments within Amazon reviews. Leveraging its rule-based approach, VADER is an optimal choice for sentiment prediction in this research.

The model assigns a compound score, ranging from -1 to $+1$, to evaluate the sentiment of each sentence. Sentences scoring ≥ 0.05 are labeled as positive; those with scores ≤ -0.05 are deemed harmful; while those within the range of -0.05 to 0.05 are categorized as neutral. The compound score effectively encapsulates the sentiment polarity of the sentence. One of the compelling facets of the VADER model lies in its ability to predict sentiment scores and label the sentiment (positive, negative, or neutral) based on these scores. This dual functionality enhances its utility in automating the sentiment analysis process for Amazon reviews.

Table 3 shows the output of the NLP pipeline predicting the aspects and determining the sentiment of sentences by extracting the familiar words of each review. In this table, the data shown here are picked randomly from the 142 predicted outputs for a single product. In this review, data aspects are distributed in different ratios. Sometimes, 12 aspects can only be present in the given review datasets if the product ratings are high. We can extract the sentiments and predict the aspect of any product following the same pattern to find the customer expectations.

Table 3. Aspect prediction and sentiment analysis results of product data from Amazon.

Predicted Aspect	Sentiment Label	Sentiment Score	Common Words
Adaptability	Positive	0.9403	oven, mode, pan
Aesthetics	Positive	0.4215	get, you, photo
Ease of Reprocessing	Positive	0.9538	easy, use, coming
Interference	Positive	0.8759	Fryer, air, love
Adaptability	Negative	−0.0816	well, thought, control

Table 3. Cont.

Predicted Aspect	Sentiment Label	Sentiment Score	Common Words
Durability	Negative	−0.5423	like, bought, march
Adaptability	Neutral	0	used, bread, bake
Ease of Use	Positive	0.6597	like, easy, use
Price	Positive	0.9442	toaster, oven, make
Performance	Negative	−0.2091	control, not work

4. Results and Discussion

The BERT-based classification model is pre-trained comparatively on a low amount of data. Hence, to train the model on three datasets, the training and validation experiments run on the Google collab using the T4 GPU. Once the tokenizer and sequence classification models are trained, these models catch 386 MB, and both are saved in the drive to recall and deploy them in the NLP prediction pipeline to predict the aspects of the input review data. T5-small and T5-base both are deployed and fine tuned on the Google collab using T4 GPU.

Accuracy: the proportion of correctly classified reviews among the total reviews.

Precision: the accuracy of positive predictions, considering both true positives and false positives.

Recall: the proportion of correctly predicted positive reviews out of all actual positives.

F1 Score: the means of precision and recall, providing a balanced measure.

To evaluate the performance of a large language model, we required calculate the accuracy, precision, recall, and F-score. The evaluation metrics for the BERT model on COMB dataset are resulted in Table 4 and the evaluation metric results for T5-Base and T5-small can be seen in Tables 5 and 6. In the below tables, the accuracy for the three models is compared and can be seen as 92% for BERT, 91% for T5-base, and 88%- T5-small. Three models show almost equal results in terms of accuracy.

Table 4. Evaluation metrics for the BERT model trained on COMB dataset.

	Precision	Recall	F1-Score	Support
Adaptability	0.94	0.89	0.92	142
Aesthetics	0.92	0.94	0.93	130
Durability	0.88	0.92	0.90	183
Ease of Reprocessing	0.93	0.95	0.94	121
Ease of Storage	0.97	0.95	0.96	125
Ease of Use	0.93	0.89	0.91	133
Ergonomics	0.90	0.86	0.88	139
Interference	0.88	0.92	0.90	130
Performance	0.85	0.85	0.85	231
Price	1.00	0.95	0.97	128
Safety	0.90	0.96	0.93	125
Use Efficiency	0.97	0.95	0.96	146
Accuracy			0.92	1733
Macro Avg	0.92	0.92	0.92	1733
Weighted Avg	0.92	0.92	0.92	1733

Table 5. Evaluation metrics for the T5-base model trained on COMB dataset.

	Precision	Recall	F1-Score	Support
Adaptability	0.89	0.89	0.89	142
Aesthetics	0.87	0.95	0.91	130
Durability	0.97	0.93	0.95	183
Ease of Reprocessing	0.94	0.93	0.93	121
Ease of Storage	0.94	0.95	0.95	125
Ease of Use	0.89	0.89	0.89	133
Ergonomics	0.91	0.80	0.85	139
Interference	0.90	0.94	0.92	130
Performance	0.85	0.86	0.86	231
Price	0.95	0.97	0.96	128
Safety	0.91	0.97	0.94	125
Use Efficiency	0.96	0.92	0.94	146
Accuracy			0.91	1733
Macro Avg	0.92	0.92	0.92	1733
Weighted Avg	0.91	0.91	0.91	1733

Table 6. Evaluation metrics for the T5-small model trained on COMB dataset.

	Precision	Recall	F1-Score	Support
Adaptability	0.82	0.89	0.86	142
Aesthetics	0.82	0.78	0.80	130
Durability	0.86	0.93	0.89	183
Ease of Reprocessing	0.86	0.93	0.89	121
Ease of Storage	0.92	0.94	0.93	125
Ease of Use	0.91	0.87	0.89	133
Ergonomics	0.82	0.80	0.81	139
Interference	0.84	0.92	0.88	130
Performance	0.90	0.78	0.84	231
Price	0.98	0.95	0.96	128
Safety	0.97	0.90	0.93	125
Use Efficiency	0.93	0.96	0.95	146
Accuracy			0.88	1733
Macro Avg	0.89	0.89	0.89	1733
Weighted Avg	0.88	0.88	0.88	1733

In Figure 3, evaluation metrics of the three aspect detection models are compared. BERT model comparatively efficient and performs better than T5-Base and small.

In this paper in the prediction pipeline, we utilized the fine-tuned BERT model as the final NLP model to detect the aspect for sentiment analysis. According to the results, BERT outperformed the T5 model. But this is not the only reason to choose BERT over T5. In the training stage, T5 models, excluding the 12 aspect labels they generated as new labels

in Table 7, we can see duplicate aspects such as attraction, color, charge. To fix this, we used cosine similarity to map the duplicate labels to original labels. Sentence transformer, a library, is used to map these labels to original labels. This affects the computational efficiency of the T5 model, and the BERT model obtained an advantage over these models.

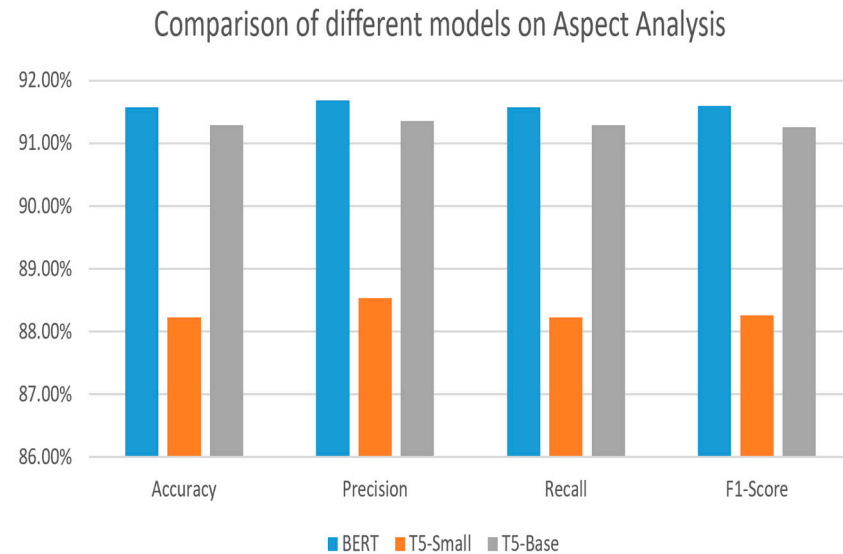


Figure 3. Comparison of evaluation metrics for three models on COMB dataset.

Table 7. T5-base model generating duplicate aspects.

	Precision	Recall	F1-Score	Support
Adaptability	0.89	0.89	0.89	142
Aesthetic	0.00	0.00	0.00	0
Aesthetics	0.87	0.92	0.90	130
Attraction	0.00	0.00	0.00	0
Charge	0.00	0.00	0.00	0
Colors	0.00	0.00	0.00	0
Durability	0.97	0.93	0.95	183
Ergonomics	0.91	0.80	0.85	139
Integration	0.00	0.00	0.00	0
Interference	0.90	0.94	0.92	130
Maintenance	0.00	0.00	0.00	0
Performance	0.85	0.86	0.86	231
Price	0.95	0.97	0.96	128
Reprocessability	0.94	0.93	0.93	121
Safety	0.91	0.97	0.94	125
Storability	0.94	0.95	0.95	125
Usability	0.89	0.89	0.89	133
Use Efficiency	0.96	0.92	0.94	146
Accuracy			0.91	1733
Macro avg	0.61	0.61	0.61	1733
Weighted avg	0.91	0.91	0.91	1733

The aspects are predicted, and the distribution of the aspects in the training dataset can be seen in Figure 4.

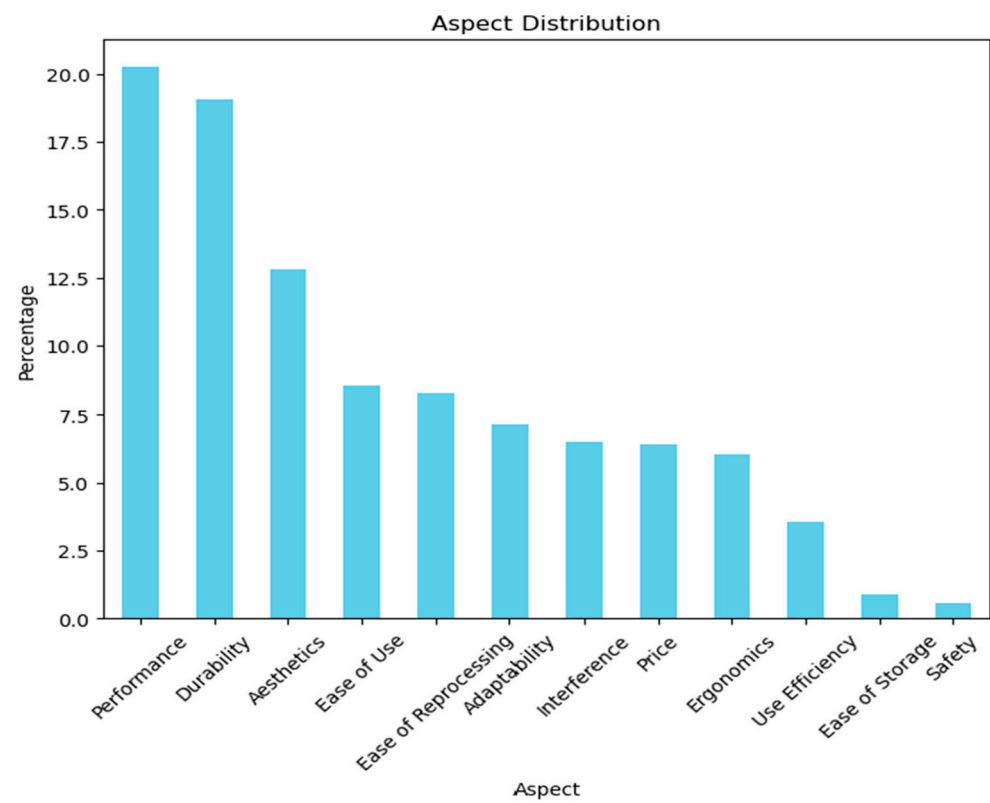


Figure 4. Aspect distribution in the training data.

The distribution of the customer ratings in the input data can be seen in Figure 5. Most portions are occupied with five- and 4-star ratings, and the other rating distribution is varied.

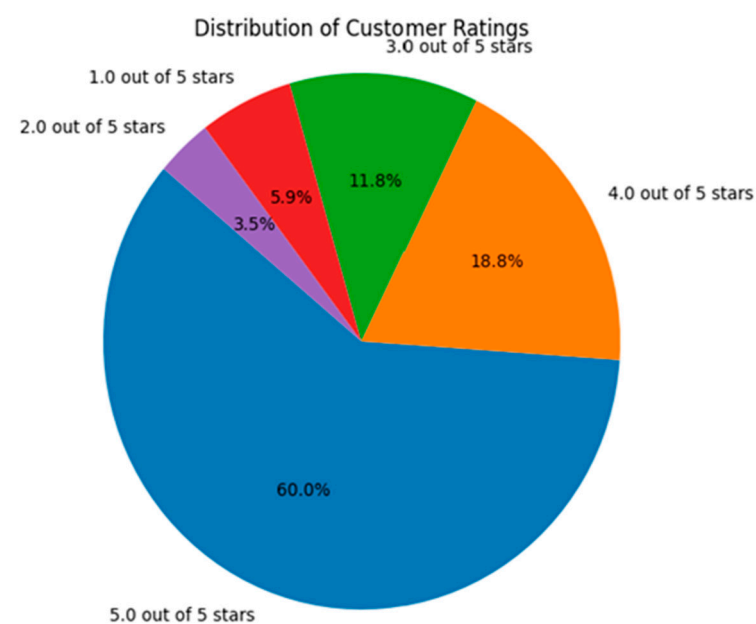


Figure 5. Distribution of customer ratings in the input data.

VADER is performed on the preprocessed data, and the prediction results are obtained, as shown in Table 3. Once the sentences are classified into positive and negative sentiments,

and T5 pre-trained models, remarkably achieved a 92% and 91% accuracy rate in predicting aspects within textual data. The exceptional performance primarily owes credit to BERT, an advanced NLP framework developed by Google, which, being pre-trained on extensive datasets, simplifies the user's task by obviating the need for extensive training. Tailoring the classifier to domain-specific requirements further amplified its efficacy, enabling precise aspect prediction and review classification. Accurately predicting aspects facilitated data-driven decisions for designers, empowering them with insights gleaned from classified reviews. The proposed approach minimizes time consumption and reduces the need for human resources to manually extract customer needs from online reviews. Its versatility allows application across diverse product categories sourced from Amazon, utilizing raw review data fed into the prediction pipeline.

However, the model is trained on a dataset generated using keywords and aspects from eco-friendly product reviews; for future work, there is scope to test the hypothesis using content analysis for a different category of products to create new aspects and sample key words. Exploring the potential of this model's prediction results to derive design insights presents a great avenue for future research. Envisioning its evolution into a search engine for designers, enabling effortless retrieval of customer opinions with a single click, signifies an exciting frontier and the ideal objective of this paper.

In conclusion, leveraging BERT and T5, the presented methodology stands as a powerful, efficient, and adaptable model for automating the extraction of customer needs from online reviews for a category of products. Its potential for providing design insights and its evolution into a designer-centric search engine paves the way for further impactful research in this domain.

Author Contributions: Conceptualization, V.K.V. and M.A.S.; methodology, M.K.S.V., V.K.V., and M.A.S.; software, M.K.S.V.; validation, M.K.S.V., V.K.V., and M.A.S.; formal analysis, M.K.S.V.; investigation, M.K.S.V.; resources, V.K.V.; data curation, M.K.S.V.; writing—original draft preparation, M.K.S.V.; writing—review and editing, V.K.V. and M.A.S.; visualization, M.K.S.V.; supervision, V.K.V. and M.A.S.; project administration, V.K.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Both the data and the code used in this research are publicly available via a GitHub Link: <https://github.com/MahammadKhalid/Enhancing-Product-Design-through-AI-Driven-Sentiment-Analysis-of-Amazon-Reviews-using-BERT> accessed on (16 January 2024).

Acknowledgments: We would like to acknowledge Aashay Mokadam and Hirofumi Sato for their support in conducting this study. We also acknowledge the contributions of Feruza Amirkulova towards the completion of this work.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript.

NLP	Natural Language Processing
ASBA	Aspect Based Sentiment Analysis
BERT	Bidirectional Encoder Representations from Transformers
CNN	Convolutional Neural Networks
AI	Artificial Intelligence
LCA	Life Cycle Assessments
POS	Parts of Speech
T5	Text-to-Text Transfer Transformer
VADER	Valence Aware Dictionary and Sentiment Reasoner
IAN	Interactive Attention Networks
COU	Context of Use
UX	User Experience

References

1. Ulrich, K.T.; Eppinger, S.D. *Product Design and Development*; McGraw Hill: New York, NY, USA, 1992.
2. Bickart, B.; Schlindler, R.M. Internet Forums as Influential Sources of Consumer Information. *J. Interact. Mark.* **2001**, *15*, 31–40. [\[CrossRef\]](#)
3. Hu, M.; Liu, B. Mining and Summarizing Customer Reviews. In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA, USA, 22–25 August 2004.
4. Bemporad, R.; Baranowski, M. *Conscious Consumers Are Changing the Rules of Marketing. Are You Ready?*; Food Marketing Institute: Arlington, VI, USA, 2007; pp. 24–27.
5. Laroche, M.; Bergeron, J.; Barbaro-Forleo, G. Targeting consumers who are willing to pay more for environmentally friendly products. *J. Consum. Mark.* **2001**, *18*, 503–520. [\[CrossRef\]](#)
6. Neuendorf, A.K. *The Content Analysis Guidebook*; Sage: Los Angeles, CA, USA, 2016.
7. Bengtsson, M. How to plan and perform a qualitative study using content analysis. *NursingPlus Open* **2016**, *2*, 8–14. [\[CrossRef\]](#)
8. Kondracki, N.L.; Wellman, N.S. Content analysis: Review of methods and their applications in nutrition education. *J. Nutr. Educ. Behav.* **2002**, *34*, 224–230. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Hsieh, H.F.; Shannon, S.E. Three approaches to qualitative content analysis. *Qual. Health Res.* **2005**, *15*, 1277–1288. [\[CrossRef\]](#) [\[PubMed\]](#)
10. Erlingsson, C.; Brysiewicz, P. A hands-on guide to doing content analysis. *Afr. J. Emerg. Med.* **2017**, *7*, 93–99. [\[CrossRef\]](#) [\[PubMed\]](#)
11. Elo, S.; Kyngäs, H. The qualitative content analysis process. *J. Adv. Nurs.* **2008**, *62*, 107–115. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Park, S.; Kim, H.M. Finding Social Networks Among Online Reviewers for Customer Segmentation. *J. Mech. Des.* **2022**, *144*, 121703. [\[CrossRef\]](#)
13. Robert, I.; Liu, A. Application of data analytics for product design: Sentiment analysis of online product reviews. *CIRP J. Manuf. Sci. Technol.* **2018**, *23*, 128–144.
14. Mokadam, A.; Shivakumar, S.; Viswanathan, V.; Suresh, M.A. Online Product Review Analysis to Automate the Extraction of Customer Requirements. In Proceedings of the ASME 2021 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Virtual, 17–19 August 2021.
15. Saidani, M.; Joung, J.; Kim, H.; Yannou, B. Combining life cycle assessment and online customer reviews to design more sustainable products—Case study on a printing machine. *Procedia CIRP* **2022**, *109*, 604–609. [\[CrossRef\]](#)
16. Saidani, M.; Yannou, B.; Leroy, Y.; Cluzel, F.; Kim, H. How circular economy and industrial ecology concepts are intertwined? A bibliometric and text mining analysis. *arXiv* **2020**, arXiv:2007.00927.
17. Dhasmana, G.; Prasanna Kumar, H.R.; Prasad, G. Sequence to Sequence Pre-Trained Model for Natural Language Processing. In Proceedings of the 2023 International Conference on Computer Science and Emerging Technologies (CSET), Bangalore, India, 6–7 November 2023; pp. 1–5. [\[CrossRef\]](#)
18. Oralbekova, D.; Mamyrbayev, O.; Othman, M.; Kassymova, D.; Mukhsina, K. Contemporary Approaches in Evolving Language Models. *Appl. Sci.* **2023**, *13*, 12901. [\[CrossRef\]](#)
19. Zhang, H.; Pan, F.; Dong, J.; Zhou, Y. BERT-IAN Model for Aspect-based Sentiment Analysis. In Proceedings of the 2020 International Conference on Communications, Information System and Computer Engineering (CISCE), Kuala Lumpur, Malaysia, 3–5 July 2020; pp. 250–254.
20. Tiwari, A.; Tewari, K.; Dawar, S.; Singh, A.; Rathee, N. Comparative Analysis on Aspect-based Sentiment using BERT. In Proceedings of the 7th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 21–23 February 2023; pp. 723–727.
21. Shi, P.-Y.; Yu, J.-H. Research on the Identification of User Demands and Data Mining Based on Online Reviews. In Proceedings of the 2022 International Conference on Big Data, Information and Computer Network (BDICN), Sanya, China, 20–22 January 2022; pp. 43–47.
22. Tong, Y.; Liang, Y.; Liu, Y.; Spasić, I.; Hicks, Y. Understanding Context of Use from Online Customer Reviews using BERT. In Proceedings of the 18th IEEE International Conference on Automation Science and Engineering (CASE), Mexico City, Mexico, 20–24 August 2022; pp. 1820–1825.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.