big data and
cognitive computing

MDPI

Article

Improving Clothing Product Quality and Reducing Waste Based on Consumer Review Using RoBERTa and BERTopic Language Model

Andry Alamsyah * **and Nadhif Ditertian Girawan**

School of Economics and Business, Telkom University, Bandung 40257, Indonesia
* Correspondence: andrya@telkomuniversity.ac.id

Abstract: The disposability of clothing has emerged as a critical concern, precipitating waste accumulation due to product quality degradation. Such consequences exert significant pressure on resources and challenge sustainability efforts. In response, this research focuses on empowering clothing companies to elevate product excellence by harnessing consumer feedback. Beyond insights, this research extends to sustainability by providing suggestions on refining product quality by improving material handling, gradually mitigating waste production, and cultivating longevity, therefore decreasing discarded clothes. Managing a vast influx of diverse reviews necessitates sophisticated natural language processing (NLP) techniques. Our study introduces a Robustly optimized BERT Pretraining Approach (RoBERTa) model calibrated for multilabel classification and BERTopic for topic modeling. The model adeptly distills vital themes from consumer reviews, exhibiting astounding accuracy in projecting concerns across various dimensions of clothing quality. NLP's potential lies in endowing companies with insights into consumer review, augmented by the BERTopic to facilitate immersive exploration of harvested review topics. This research presents a thorough case for integrating machine learning to foster sustainability and waste reduction. The contribution of this research is notable for its integration of RoBERTa and BERTopic in multilabel classification tasks and topic modeling in the fashion industry. The results indicate that the RoBERTa model exhibits remarkable performance, as demonstrated by its macro-averaged F1 score of 0.87 and micro-averaged F1 score of 0.87. Likewise, BERTopic achieves a coherence score of 0.67, meaning the model can form an insightful topic.

Keywords: big data; multilabel classification; natural language processing; sustainability

check for
updates

Citation: Alamsyah, A.; Girawan, N.D. Improving Clothing Product Quality and Reducing Waste Based on Consumer Review Using RoBERTa and BERTopic Language Model. *Big Data Cogn. Comput.* **2023**, 7, 168. https://doi.org/10.3390/bdcc7040168

Academic Editor: Jenq-Haur Wang

Received: 23 September 2023
Revised: 18 October 2023
Accepted: 23 October 2023
Published: 25 October 2023

1. Introduction

The fashion industry is responsible for creating approximately 40 million tons of textile waste worldwide, most of which ends up in landfill or incinerated [1]. The issue of clothing waste is prevalent in developed and developing countries [2]. A major contributor to the accumulation of clothing waste is the tendency of most consumers to dispose of clothes that no longer serve their purpose, compounded by the poor quality of clothing products and their short lifespan [3]. The pressure on clothing companies to prioritize the scale and speed of production over product quality has exacerbated the situation in response to heightened consumer demand in a highly competitive market. This phenomenon, known as "fast fashion", entails the mass production of clothing items quickly and in large volumes, resulting in a decline in product quality for consumers [4].

The shift of clothing companies towards e-commerce platforms has also contributed to the speed and scale of production. By embracing e-commerce, clothing companies have gained access to new market segments and an opportunity to increase their profits [5]. The fashion industry has the largest B2C e-commerce market segment, and its global size is estimated to reach USD 752.5 billion by 2020. This market is expected to grow further by 9.1% per annum and reach a total market size of USD 1164.7 billion by the end of the year.

The United States market was valued at USD 155.6 billion in 2020, with one of its most prominent players being Amazon [6].

Customers with knowledge of a particular product, such as clothing, are valuable feedback resources [3]. We can collect user-generated content (UGC), such as consumer reviews from social media, online forum discussions, and review sections on e-commerce websites. Clothing companies can leverage consumer reviews as an invaluable data source to gather product feedback and ideas for further product development [7].

Product development aims to create durable items, minimizing waste and resource depletion. Analyzing e-commerce reviews is a valuable approach, but it requires advanced techniques. Multilabel classification is a necessary machine-learning method that assigns multiple labels to reviews, going beyond single-category classification. The method allows a detailed evaluation of products, considering various characteristics and customer feedback dimensions. This analytical approach is crucial for evolving products that meet diverse consumer needs. By decoding multifaceted customer feedback, development teams gain insights into material handling. This effort leads to the optimized use of material, efficient handling, reduced waste, and a clear integration of sustainability in product development.

In the context of clothing, it can be useful to utilize machine learning to gain insight to improve product quality. Research by [8] employs sentiment analysis on e-commerce product reviews, using data collection programs to gather comments and a sentiment classifier to categorize feedback automatically. Furthermore, research by [9] utilizes supervised machine learning to extend existing life-cycle assessment studies and create a tailored model for assessing clothing products' environmental sustainability throughout their life cycles. In addition, classifying clothing dimensions and knowing the topic from a review can enhance the understanding of the market and pave the way to more sustainable business.

The deep-learning model BERT (Bidirectional Encoder Representations from Transformers) stands out for its capability to comprehend sentences contextually, making it an ideal choice for label prediction on multilabel classification tasks [10]. The BERT model has been utilized in research in diverse areas, such as detecting fake news on the Internet [11] and identifying one's behavior based on the nature of extraversion and neuroticism of personality dimensions [12]. In this study, we have chosen the RoBERTa model, which is trained on a larger dataset and can contextualize words better than BERT [13]. The BERT model can also be implemented on topic modeling tasks called BERTopic. We will use RoBERTa, and BERTopic to map product quality problems in the fashion industry, especially clothing, to help clothing companies produce quality products to reduce clothing waste.

Despite the advancements in applying deep-learning methodologies to the fashion industry, a discernible gap persists in utilizing robust language models for classifying clothing quality. Previous research endeavors have predominantly focused on employing deep learning for categorizing types of clothing products [14]. Furthermore, research conducted by Dirting et al. (2022) innovatively leveraged the BERT model for a multilabel classification task to discern hate speech [15]. The specific application of language models like BERT and its derivatives to classify and evaluate clothing quality remains absent from the existing literature. The RoBERTa model, an optimized version of BERT, showcases potential in this domain due to its enhanced training and preprocessing technique. It could arguably facilitate a more nuanced understanding and classification of textual descriptions and reviews related to clothing quality. This research contribution integrates innovative machine-learning methodologies, particularly multilabel classification and utilizing advanced models such as RoBERTa and BERTopic, within the scope of fashion industry analysis. Section 2 covers theories related to clothing waste management, the significance of quality-focused product development, and the application of machine learning for consumer review insights. Section 3 shows the research framework, data preparation, and model evaluation metrics. Section 4 discusses model results and their potential for enhancing clothing product quality. Finally, Section 5 provides research implications, limitations, and future directions.

2. Materials and Methods

2.1. Management of Waste

Managing waste in the fashion industry is essential. Prevention is the first and most crucial step in managing waste, followed by product reuse, recycling, and recovery [16]. The very last option is disposal. Furthermore, waste prevention can be done by promoting longer product lifespans [17].

Waste is a human creation that loses its function or no longer carries out its purpose [18]. In the business context, it is interpreted as transforming all raw resources into outputs in the form of goods or services that no longer have value. In that sense, natural resources must be processed into products according to market demand to minimize waste production; by creating products that meet consumer demand and preferences, companies can reduce the amount of waste generated from unsold or unused items.

Companies must prioritize product development to meet consumers' ever-changing demands and preferences for high-quality clothing. By continuously improving and innovating their product offerings, businesses can differentiate themselves from competitors and create unique value propositions [19]. This effort benefits the environment by reducing waste, driving customer satisfaction, and improving profitability. By adopting a product development approach, clothing companies can stay ahead of the curve, meet consumer expectations, and achieve sustainable growth in the long term.

2.2. Product Development in Fashion

Understanding the role of product development in reducing fashion industry waste, we examine research by Goworek et al. (2020), which highlights clothing as a significant waste contributor. Improving clothing durability involves enhancing knowledge, skills, processes, and infrastructure [20]. Product development should focus on longevity and suggest product development in fashion addresses market opportunities from globalization and information technology use, considering trends, culture, and advertising to meet consumer demands. Brands prioritize sustainable products as consumers seek sustainable fashion.

Involving consumers in product development is crucial but challenging due to diverse feedback and qualitative aspects. Language models can help analyze user-generated content (UGC) to reduce product failure rates and create consumer-friendly products [21]. Consumer reviews are a data source for feedback and an avenue for product development ideas.

2.3. Quality Management in Fashion

A study by Kumar et al. states that quality management is vital in guiding product development toward customer-focused and quality-driven outcomes. It helps management decision-makers by suggesting the dimensions of dynamic and quality-management capabilities that significantly impact new product development performance [22]. Quality management consists of several processes and sub-processes. It includes the organizational structure, roles, responsibilities, resources, and infrastructure necessary to achieve quality goals [23].

In the realm of the garments industry, Shen and Chen's (2020) research outlines a three-step quality-management process [24]. The initial phase involves manufacturer department quality assessment to ensure a minimal error rate in production. Subsequently, production quality management is achieved by disseminating information regarding activities like inventory, materials, and work in progress. The concluding stage, quality inspection and assurance, engages production management to evaluate the completed product. Incorporating consumer perspectives at this juncture is advised, as their perceptions and expectations define quality [18], upholding the organization's commitment to excellence.

In the sense of improved material handling, quality management plays a significant role. By monitoring and analyzing material handling procedures for bottlenecks and inefficiencies, ongoing enhancements are possible. This mutual dedication to improvement ensures that quality management and material handling can adapt effectively to changing demands [25]. Processes are streamlined, and unnecessary steps are removed. Similarly, efficiency is bolstered in material handling by optimizing material movement, minimizing transfers, and reducing handling time. When manufacturers apply the process optimization principles from quality management to material handling, they achieve a smoother and more efficient flow of materials [26].

2.4. Clothing Quality Dimensions for Implementing Quality Management

Product quality improvement is crucial to reducing waste by aligning with consumer preferences. Companies must consider the consumer's perspective when defining and measuring product quality to achieve this. One way to accomplish this is by analyzing customer reviews on company e-commerce platforms. By understanding quality in terms of its dimensions, businesses can better manage quality and enhance the development process [27]. Quality is a fundamental concept that applies to both tangible and intangible products. In the context of clothing, quality refers to the cycle of wear post-purchase, including factors such as clothing care and durability [28]. They also identified five key dimensions of clothing product quality in Table 1.

Table 1. Quality dimension of clothing product.

Dimension	Description
Materials	The clothing material properties encompass thickness, weight, stretch, and flexibility to support human movement [27]
Construction	The construct factors such as stitch position, type, clothing piece, style, and interlining are relevant [27]
Color	Clothing colors can reflect the wearer's mood, influence body perception, and derive from designers' surroundings for appealing combinations [27,28]
Finishing	Finishing enhances clothing characteristics, aesthetic value, and service life, affecting appearance, shine, softness, drape, density, and usability. It can be permanent or temporary [19]
Durability	A resistance feature to movement, wear, and washing includes abrasion, pilling, and stiffness [29]

2.5. Machine-Learning Approach for Predicting Quality Dimensions

The machine-learning approach is essential for efficient and accurate classification when dealing with large-scale data, as some datasets contain certain features or words that might be strongly correlated with specific classes or outcomes of interest, which is too complex to handle manually. By utilizing powerful classification algorithms, organizations can categorize data into distinct classes, enabling the development of effective prediction models based on the relationships between different variables and their respective labels [30]. Thus, it considerably reduces the time required to classify vast amounts of data and allows for more informed decision-making based on meaningful insights [31]. The use of machine learning for data classification has become increasingly prevalent in today's data-driven world, and businesses that adopt this approach can gain a significant competitive advantage by making better use of their data resources.

Researchers have developed a variety of algorithms to classify quality dimensions in different domains. One such approach was presented by Xie and Burstein (2011). They used machine learning to provide an adaptive attribute-based system for evaluating the quality of online information sources [30]. The research identifies Support Vector Machines (SVM) as the suitable classification method for addressing the specific learning problems of the study—notably, the achieved prediction for performance in online healthcare resources. The result ranges from 73% to 90% accuracy, confirming the feasibility of using ML techniques to generate value suggestions for describing resource quality attributes. Furthermore, the paper highlights the practical applicability of the proposed approach through a preliminary usability test with domain experts, yielding promising results and enabling informed decision-making. Another study by Liu and Chen uses the Long Short-Term Memory (LSTM) model to forecast the quality scores of service providers. The utilization of LSTM network-based sensitivity analysis, combined with improvement expenses, to sort the subdimensions of the service quality model. The results of LSTM prediction accurately reflect customer requirements, as demonstrated by the analysis of online reviews of hotels [32]. These innovative approaches to quality classification highlight the potential of advanced ML techniques to provide valuable insights and drive informed decision-making in various industries.

2.6. Multilabel Classification

Dealing with real-world data can be challenging, as a single review or document may contain multiple semantic aspects simultaneously. Researchers have developed a solution to address the issue whereby each data point is assigned suitable multiple labels representing its unique semantics [33]. This type of classification is known as multilabel classification, where each document is labeled to multiple classes, unlike single-label classification, where each document is labeled to only one class. Multilabel classification is commonly applied to text classification tasks, where each document is associated with more than one topic or theme. By leveraging the power of multilabel classification, researchers can accurately capture the complexity and nuance of real-world data, unlocking new insights and driving more informed decision-making in various industries.

Multilabel classification has emerged as a powerful tool for analyzing complex e-commerce data. By synchronizing information from these two sources, the researchers could more accurately separate categories of items and gain a deeper understanding of their properties. Similarly, Deniz et al. (2022) recognized the potential of multilabel classification for in-depth product analysis, particularly in the context of the abundant textual data produced in e-commerce reviews [34]. To evaluate the effectiveness of their model, the researchers employed three different datasets, each with a different number of labels, and used MicroF1 and MicroR as evaluation metrics. These metrics differ from those used in single-label classification, highlighting the unique challenges and opportunities presented by multilabel classification in the e-commerce domain. By leveraging the power of multilabel classification, researchers can unlock new insights and drive more informed decision-making, empowering businesses to meet the ever-evolving demands of the modern marketplace.

To evaluate the model performance, we need to employ suitable evaluation metrics. It is crucial for obtaining precise and significant outcomes in multilabel classification scenarios. One study provides a comprehensive comparison of evaluation metrics for both single-label and multilabel classification, highlighting the need for different metrics in the latter case to avoid misleading conclusions [35]. The researchers suggest that micro-average and macro-average measurements are particularly useful for representing model performance, as they are closely linked. In Table 2, the performance evaluation of a multilabel classification model includes metrics such as MicroP, MacroP, MicroR, MacroR, MicroF1, and MacroF1 [34].

The applications of multilabel classification used in this research are elaborated in research by Wei et al. [36]. The paper introduces a multilabel text classification model that employs multi-level constraint augmentation and label association attention to enhance low-frequency label prediction in cases with limited samples. The model incorporates a data augmentation approach involving multi-level constraints to mitigate label category imbalance and ensure systematic sample generation. This augmentation process takes into account historical generation data, original sample text information, and sample topics to guide text generation. The primary challenges addressed in this context are the imbalance in the number of distinct label categories and the difficulty in distinguishing closely related labels. The authors present a multilabel text classification model rooted in multi-level constraint augmentation and label association attention mechanisms to tackle these issues. In addition to the real-world use case, Lin et al.'s research [37] introduces a multi-task, multilabel emotion classification model comprising three key components:

a general representation module, an emotion representation module, and an adversarial classifier. The model incorporates emotion descriptors to capture inter-emotion correlations and employs adversarial training to regulate the injection of excessive emotion-related information into the shared layer. Two datasets, one in Indonesian and the other in English, were constructed for the multilabel emotion classification task, comprising 4207 and 26,019 samples, respectively. The proposed approach outperforms existing state-of-the-art baselines in multilabel emotion classification. It achieves macro-average F1 scores of 50.21%, 41.33%, and 40.24% on the Chinese, English, and Indonesian datasets, respectively.

Table 2. Multilabel classification evaluation metrics.

Formula	Description
$\mathrm{MicroP} = \frac{\Sigma_{C_i \in C} \mathrm{TPs}(c_i)}{\Sigma_{C_i \in C} \mathrm{TPs}(c_i) + \mathrm{FPs}(c_i)}$	Micro-averaged Precision measures the overall Precision of all classes (c_i) by calculating the sum of True Positives (TP) and dividing it by the combined sum of each class's TP and False Positives (FP).
$\mathrm{MacroP} = \frac{\Sigma_{C_i \in C} \mathrm{P}(D, c_i)}{\lvert C \rvert}$	Macro-averaged Precision calculates the average Precision of all classes (c_i) by determining each class's Precision and then averaging the values.
$\mathrm{MicroR} = \frac{\Sigma_{C_i \in C} \mathrm{TPs}(c_i)}{\Sigma_{C_i \in C} \mathrm{TPs}(c_i) + \mathrm{FNs}(c_i)}$	Micro-averaged Recall calculates the average Recall of all classes (c_i) by summing True Positives (TP) and False Negatives (FN) across all classes and then dividing the total TP by the combined sum of TP and FN.
$\mathrm{MacroR} = \frac{\Sigma_{C_i \in C} \mathrm{R}(D, c_i)}{\lvert C \rvert}$	Macro-averaged Recall calculates the overall average Recall by determining the Recall value for each class (c) and then averaging all the Recall values.
$\mathrm{MicroF1} = 2 \cdot \frac{\mathrm{MicroP} \cdot \mathrm{MicroR}}{\mathrm{MicroP} + \mathrm{MicroR}}$	Micro-averaged F1-score aggregates F1-scores of all classes. Calculate it by multiplying MicroP and MicroR, then multiply the result by two and divide by the sum of MicroP and MicroR.
$\mathrm{MacroF1} = \frac{1}{N} \sum_{i=0}^{N} \mathrm{F1}$	Macro-averaged F1-score calculates the average of all F1 scores by determining each label's F1 score and then averaging them.

2.7. Natural Language Processing and Language Model

Developing models that understand human language is crucial to classify textual data effectively. Natural Language Processing (NLP) provides a collection of computational techniques for automatically analyzing and representing language, making it an essential tool for overcoming this challenge. The NLP objectives range from natural language translation and information retrieval to text summarization, question answering, topic modeling, and opinion mining [29]. One essential technique used in NLP is language modeling, which involves learning and determining the probability of a word based on training data. The ultimate goal is to predict the token in each sequence [32] accurately.

Conventional classifiers like SVM and Naïve Bayes require extensive preprocessing for language modeling, such as removing missing data, lowercasing, tokenization, and lemmatization. However, advanced models like BERT, which is pre-trained, and LSTM, which is an RNN, both eliminate the need for preprocessing steps. The models can be fine-tuned for specific tasks using different data and leveraging the knowledge [38]. Transformers architecture has revolutionized NLP by effectively capturing long-range dependencies in language models [39]. Open-source libraries like Transformers provide a common API for various state-of-the-art architectures and offer a selection of pre-trained models, making them widely accessible [40].

2.7.1. Transformer Architecture

Transformers are a groundbreaking deep-learning architecture that has revolutionized natural language processing (NLP) and other sequential data tasks. At the heart of this architecture lies the encoder, a crucial component responsible for processing the input sequence and creating meaningful representations of the data. The encoder reads the sentence word by word, paying special attention to how each word relates to others, figuring out their importance and how they connect. It looks at the whole sentence, considering the context and meaning of each word based on all the others. This way, it captures long-distance relationships and understands the bigger picture. Using this contextual understanding, the encoder creates a special representation for each word, considering how it fits into the whole sentence. These representations are like little summaries of each word, incorporating all the relevant information around it [39].

The decoder uses these smart summaries to generate a meaningful response or translation. It knows how to use the encoded information to come up with a well-formed and contextually appropriate answer. This encoding and decoding process allows Transformers to excel at understanding complex language patterns, making them incredibly useful for tasks like translation, summarization, and more. The encoder's ability to grasp the overall meaning of a sentence by considering all the words together sets Transformers apart and makes them powerful in dealing with language tasks. Transformers employ a multilayer stack of encoders, refining representations through layer normalization and preserving input information with residual connections. By considering the entire sequence simultaneously, the encoder models dependencies globally, distinguishing them from traditional linear processing in models like RNNs. With the contextualized representations feeding the decoder, Transformers become proficient in various NLP tasks, efficiently processing sequential data and capturing word relationships. This remarkable capability makes them widely applicable and beneficial for many people, providing a grip on understanding complex language and enabling more accurate and contextually meaningful outcomes in various real-world applications [39].

The Transformer model is considered better than conventional neural networks for numerous reasons. First, it is based on self-attention, which allows it to extract important information from the inputs, leading to improved performance in various tasks such as weather forecasting [41]. Second, the Transformer model incorporates multi-head self-attention and encoder-decoder attention, enhancing its ability to capture spatial and temporal features, resulting in better performance than recurrent neural networks [42]. Additionally, the Transformer model has been shown to be effective in developing reliable protection schemes for transformers by fusing multiple features and improving generalizability [43]. Furthermore, when combined with a convolutional neural network, the Transformer model has demonstrated high accuracy and robustness in classifying electrocardiogram signals, making it suitable for industrial applications [44]. Finally, the Transformer model has been successfully used as a computationally efficient homogenization surrogate model, enabling accurate predictions of material response in composite microstructures [45].

2.7.2. BERT

A multilayer bidirectional Transformer encoder involves pretraining and fine-tuning [46]. Pretraining uses a large corpus from BooksCorpus with 800 million words. Fine-tuning adapts BERT to a specific dataset by replacing the output layer and modifying the original weights. BERT's 12 hidden layers, 768 hidden sizes, 12 attention heads, and 110 M parameters enable multilabel text classification [13]. The fine-tuning objective is to adapt the BERT model to the entered dataset to produce an accurate model [47]. Researchers have used BERT to improve classification performance compared to other language models. Table 3 shows the performance comparison between BERT and other models.

Table 3. BERT comparison.

Case	Dataset	Performance
Using BERT, Char-CNN, Graph-CNN, LSTM, and Bi-LSTM as tools to predict the number of users of online food delivery services by Biswas et al. (2021) [48]	5680 Facebook comments	F1 score BERT 92.5%, Bi-LSTM 84.3%, Graph-CNN 82.9%, Char-CNN 75.8%, dan LSTM 72.6%.
BERT, K-Star, and FFNN for real estate investment models on online textual information [49]	5 million property records from Airbnb and Zillow	F1 score of BERT is 92%, FFNN is 82%, dan K-Star is 81%.
BERT model to find helpful and unhelpful customer reviews [50]	Yelp open dataset from 12 October 2004 to 14 November 2018	F1-Score of BERT is 71%, SVM is 67%, NB is 62%, and k-NN is 59%.
BERT, k-NN, SVM, and Naïve Bayes to perform sentiment analysis with symmetrical structures to obtain features at the sentence level of agricultural products [51]	6152 data from social media, news websites, e-commerce websites, and offline surveys.	BERT has an accuracy rate of 70% and an F1-score of 71%, while SVM has an accuracy of 67.9% and an F1-score of 67.8%, Naïve Bayes has an accuracy rate of 61.7% and an F1-score of 62.8%, and k-NN has an accuracy rate of 59.6% and an F1-score of 59.1%.

2.7.3. RoBERTa

A Robustly Optimized BERT Pretraining Approach (RoBERTa) is a pre-trained BERT model with a large corpus in English, including BookCorpus, English Wikipedia, CC-News, openWebText, and Stories. The total data size from five sources reaches 160 GB. RoBERTa uses dynamic masking to solve the problem of static masking, which avoids the same mask during pretraining on each epoch. RoBERTa has 12 hidden layers, 768 hidden sizes, 12 attention heads, and 125 M parameters. In addition, RoBERTa does not use NSP, as it is useless for pretraining. Facebook teams claimed RoBERTa could produce better output than BERT [46].

RoBERTa has been utilized in numerous studies. The research conducted by Malik et al. (2023) combined utilization of multilingual and translation-based methodologies, as investigated in this study, offers a promising avenue for addressing the intricate task of detecting hope speech across various languages [52]. This approach facilitates the classification of content in diverse linguistic contexts. Another study by You et al. (2022) proposed ASK-RoBERTa. The research represents a noteworthy development in the field of aspect-based sentiment classification (ABSC) [53]. It is a sentiment knowledge-adaptive pretraining model tailored for this specific task. Additionally, the study encompasses the formulation of term and sentiment mining rules, which are constructed through a rigorous process involving part-of-speech tagging and sentence dependency grammar analysis. Furthermore, RoBERTa endeavors to confront the formidable task of discerning spurious information pertaining to the COVID-19 pandemic. The empirical findings showcased in this study substantiate the efficacy of employing pre-trained language models, notably BERT and RoBERTa, in identifying and mitigating deceptive news articles and content concerning the COVID-19 crisis [54].

The improvement made by the Facebook team to BERT has implications for RoBERTa's performance [55]. The performance comparison of BERT and RoBERTa on several studies is shown in Table 4.

Among various deep-learning models, including XLNet and Electra, BERT stands out as the top performer in emotion recognition tasks. Specifically, RoBERTa achieves the highest F1-score, demonstrating its superior performance in this context. In comparison, Electra attains an F1-score of 0.33, XLNet achieves 0.48, and RoBERTa reaches 0.49. These results underscore RoBERTa's effectiveness and ability to outperform other models in the emotion recognition task [55].

Table 4. BERT and RoBERTa comparison.

Case	Dataset	Performance
Assessing BERT, DistillBERT, RoBERTa, XLNet, and ELECTRA for recognizing emotions from 28 emotion labels in text [55]	GoEmotion: 58,000 Reddit comments	RoBERTa F1-score is 49%, and BERT is 46%
Apply RoBERTa, ALBERT, and DistillBERT to detect spam or fake reviews [56]	1.4 million Yelp restaurant and hotel reviews	BERT F1-score is 65%, and RoBERTa is 68%
Modify BERT, XLNet, and RoBERTa for clickbait detection using data expansion, pruning, and augmentation [57]	Webis Clickbait Corpus 2017: 40,967 data; Kaggle clickbait detector: 18,397	BERT F1-score is 68%, while RoBERTa is 69%
Verifying facts using BERT, RoBERTa, and Electra [58]	Fact Extraction and Verification (FEVER) dataset: 1000 data	RoBERTa's F1-score is superior, with 95%, and BERT 94%

2.7.4. Topic Modeling with BERTopic

Topic modeling identifies sentence subject matter by grouping words representing content [59]. This unsupervised learning method scans a corpus and groups similar words, enabling the interpretation of dataset meaning [60]. Latent Dirichlet Allocation (LDA) is a prevalent topic modeling technique. Massive-scale data understanding is vital for decision-making and innovation. LDA, a probabilistic model, is employed to analyze text data, deriving statistical relationships for classification, sentence implication, text similarity, and novelty detection, thus summarizing large text collections [61]. However, the data's magnitude requires advanced techniques and tools for managing and inferring insights. In response to these challenges, Grootendorst (2022) [59] introduces a tailored BERTopic architecture. This architecture is employed for clustering software description texts and automated refinement of application software tags. This refinement process hinges on the clusters derived from topic clustering, coupled with the extraction of salient subject words. Importantly, the model exhibits noteworthy performance. These results underscore the model's high Precision and effectiveness in categorizing software, minimizing instances of misclassification. Table 5 demonstrates BERTopic's superior performance compared to LDA.

Researchers have utilized BERTopic in several areas. For example, research by Aytaç and Khayet (2023) proposes the utilization of BERTopic; it was deployed for the comprehensive analysis of a substantial dataset comprising 3684 articles within the context of molecular dynamics (MD) literature [62]. The application of BERTopic yielded commendable results, discerning salient terminologies that encapsulated the essence of the dataset. Furthermore, it facilitated the identification of both pervasive overarching themes and intricately nuanced localized topics within the domain of MD research. These findings bear notable significance, furnishing MD researchers with valuable insights to inform their future research endeavors and offering a comprehensive overview of the present landscape within the field. In addition, research by Bu et al. (2023) This paper delves into the prevalent concern of inaccurate software recommendations within the application software market, primarily stemming from deficiencies in objectivity, hierarchical structure, and standardized classification tags [63].

Table 5. BERTopic comparison.

Case	Dataset	Performance
Detecting the user's interest topics in MOOCs by Zankadi et al. (2022) [64]	Tweet from Twitter with the keywords "Computer Science" and "Artificial Intelligence". The amount of data used is 10,000 tweets	BERTopic excelled with a coherence score of 0.61, while LDA_BOW and LDA_TFIDF were 0.50 and 0.59
Thompson & Mimno (2020) researched determining context-appropriate word representations using BERT and LDA [65]	Supreme Court of the United States (SCOTUS) and Amazon reviews	BERTopic achieves a word entropy score of 4, while LDA achieves 5. The coherence score of BERTopic is 0.6, LDA 0.5
de Groot et al. (2022) researched the generalization of short multi-domain text using BERT and LDA [66]	The dataset contains open-text comments with a total of 62,522 data. The data comes from students, and the 20 NG dataset contains 11,096 news articles	BERTopic achieves a word entropy score of 4, while LDA achieves 5. The coherence score of BERTopic is 0.6, LDA 0.5

3. Research Framework

Selecting an appropriate NLP model is crucial for optimal results in any task. RoBERTa is chosen for the multilabel classification task, given its suitability and superior performance compared to other models considered. The primary objective of this study is to enhance the quality of clothing products while also addressing sustainability issues through implementing a multilabel classification and topic modeling approach, employing state-of-the-art models like RoBERTa and BERTopic. By combining the power of multilabel classification and topic modeling, the research aims to identify and classify various quality dimensions associated with clothing products, enabling a comprehensive assessment of their overall quality.

In addition to improving product quality, the study seeks to integrate sustainability considerations into the classification and topic modeling process. The incorporation of sustainability issues is vital in today's environmentally conscious world, where the fashion industry faces increasing pressure to adopt eco-friendly practices.

Figure 1 represents the construction of the multilabel classification model using BERT and RoBERTa, spanning from data collection through model analysis. The model must be trained and evaluated to identify patterns in the training data for evaluation using testing data. A well-performing model is then chosen and subsequently fine-tuned for practical application in various conditions.

Figure 2 shows the processes of model construction for topic modeling. It includes data preparation (data collection and preprocessing) and a preliminary step for topic modeling.

The research workflow consists of two approaches, namely topic modeling and multilabel classification. These approaches are done in parallel; multilabel classification is utilized to classify the issue of clothing quality, whereas topic modeling can provide insight into the public voice. This comprehensive approach ensures a holistic understanding of clothing quality concerns and the multifaceted voices of consumers in this domain.

Figure 3 provides a comprehensive depiction of the research workflow, encompassing a series of tasks that collectively contribute to the study's methodology. These tasks include steps such as data labeling, a crucial process to assign appropriate labels to the dataset's components, ensuring accurate representation, and subsequent analysis. The preprocessing phase is undertaken to eliminate noise and standardize text, resulting in a cleaner and more coherent dataset for the model. Furthermore, the workflow involves the crucial step of splitting the dataset into distinct subsets, with 80% of the data designated for the training set and the remaining 20% allocated for model testing. As for topic modeling tasks, the first step is preprocessing, followed by visualization of the topic. A detailed explanation is given in the next section.

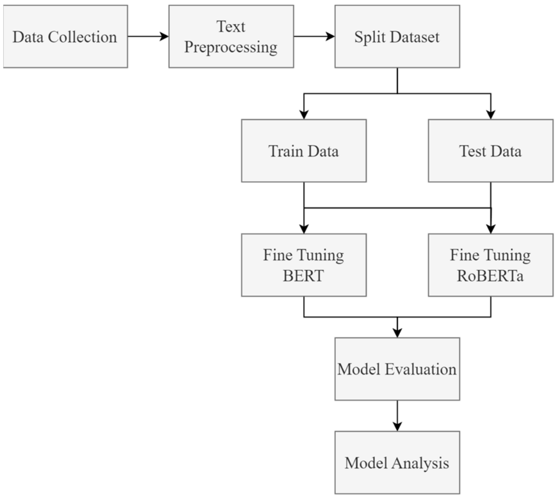

Figure 1. Model construction of multilabel classification.

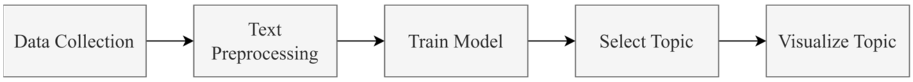

Figure 2. Model construction of topic modeling.

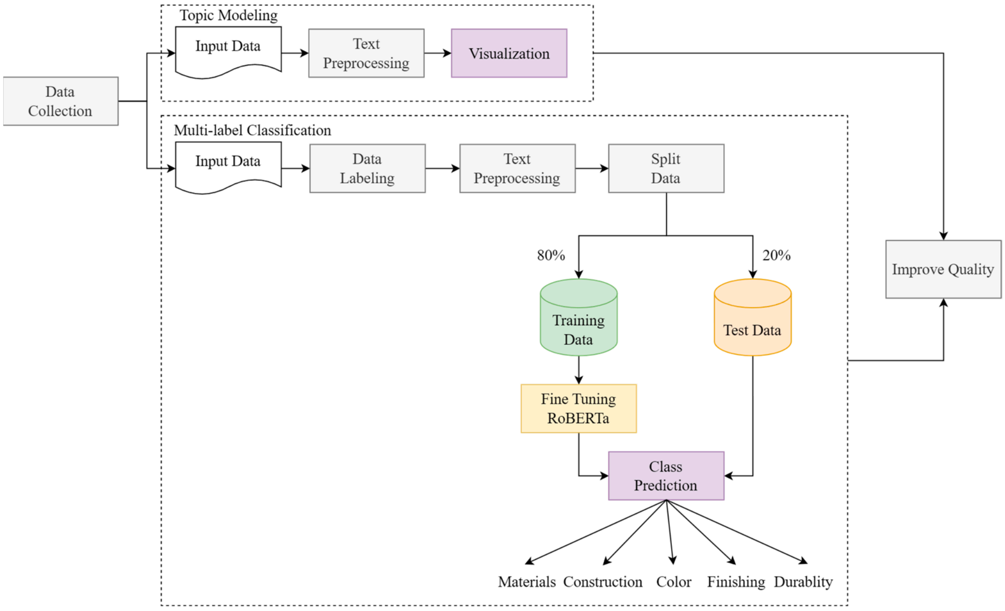

Figure 3. Research Workflow.

3.1. Data Source, Data Collection, and Dataset

There are two data sources. The first is the data collected using the AMZReviews tool to obtain Amazon reviews [67]. It is the primary data for this research [68]. The data we gather contain customer reviews for clothing products with 27,683 data points. Furthermore, to enrich our datasets, we added a dataset from Kaggle consisting of 23,486 clothing review data [69]. The combined dataset comprises 51,169 data points.

Table 6 comprises six columns, each with its own description. The sample for the User column is “R2S3IVS12Y6RC4”, the Rate column holds user ratings ranging from 1 to 5, with the sample rating as “5”, and the Format column describes product attributes, such as size and color. The “Title” column contains review headings like “Quality Baseball Pants”. The Content column presents detailed user reviews, in this case, about the fit of baseball pants for a child. The “Helpful” column indicates the number of users who found the review helpful. This structure is outlined in Table 6.

Table 6. Raw data structure.

Column Name	Description
User	Username or user ID. Data type varchar
Rate	The rating users give for a product—value from 1 to 5 (integer number)
Format	The type or variation of product the user buys. Value S, M, L, or XL
Title	The headline of the review. Data type text
Content	Detailed explanation of the user’s experience. Data type text
Helpful	Display the number of users who consider the review valuable. Data type integer

We concentrate on one primary feature in Table 6, Content, which offers insights into textual expression and sentiment. Furthermore, we include five Boolean features tied to the five quality dimensions listed in Table 1. Boolean feature is assigned a value of 0 if absent or 1 if present in the review. For instance, a review containing the "Materials" dimension receives a value of 1. There are, in total, nine features for the final dataset [67]. The final structure of the dataset is as follows: Title, Content, Rate, Class Name, Materials, Construction, Color, Finishing, and Durability. This final dataset serves as input for RoBERTa and BERTopic neural network models, employed in multilabel classification, RoBERTa fine-tuning, and BERT-based topic modeling.

3.2. Data Preprocessing

Preprocessing significantly impacts text classification success and topic modeling for enhanced insights [70]. BERT in multilabel classification task automatically preprocesses text data. The steps comprise tokenization with special tokens such as [CLS] for classification tasks and [SEP] to separate input segments. Input sequences are then padded or truncated to a fixed length, ensuring a uniform shape for efficient training and inference. Lastly, masking is employed, where BERT randomly replaces tokens with [MASK] during pretraining to learn bidirectional representations [13].

In contrast, BERTopic requires manual preprocessing, comprising tokenization, stopword removal, lowercase conversion, stemming, and lemmatization, eliminating extraneous information and reducing text complexity [71]. BERTopic was employed to create embeddings for each document in our dataset, utilizing a pre-trained BERT model.

3.3. Multilabel Classification Process

The first step of the process is labeling the data, which will be explained in Section 3.1. After the input data are labeled and preprocessed by the embedded tokenizer, the model is then fine-tuned to our specific dataset. We then conducted an extensive hyperparameter search to intricately optimize the model's performance. This comprehensive endeavor encompassed fine-tuning crucial parameters, such as learning rates, batch sizes, and the number of training epochs. The objective is to attain the most favorable outcomes tuned precisely to our specific task.

3.3.1. Data Labeling

Data labeling is vital for model construction, enabling accurate pattern learning. This process classifies raw data into correct classes and matches content with quality dimensions. With multilabel classification in mind, we added five columns to the dataset based on Table 1, namely Materials, Construction, Color, Finishing, and Durability, and filled them with numbers 0 or 1 based on data relevance. The labeled data comprised 3318 instances for clothing quality and 466 unrelated instances. To avoid bias, under-sampling balanced the dataset by selecting 1208 instances from the least-represented label—Durability.

Annotators assign one or more labels from a predefined set of labels to each data instance. This type of annotation allows for the simultaneous assignment of multiple labels to a single data point, reflecting the fact that the data instance may belong to multiple categories or classes. As an illustration, consider the following review: "The colors are gorgeous, but there was a tear in the seam of the dress I had not worn it yet." This review contextually indicates a concern related to durability, as the mention of tearing is commonly associated with the garment's resistance to wear and damage. Consequently, we fill the "Color" and "Durability" columns with a value of 1, signifying the presence of durability-related feedback, while the remaining columns are populated with 0 to indicate the absence of such feedback.

3.3.2. Fine-Tuning

The accuracy of model predictions depends on various factors, including the training data size and learning algorithms. A large enough training data size is crucial to achieving high accuracy values, but errors can still occur if there is insufficient data. In this study, we used the Random Split strategy to divide the data randomly [72]. The data have been split into a ratio of 80:20, where 80% of the data are allocated for training and 20% for testing. The total number of training data is 3029, while the number of test data is 758.

The BERT and RoBERTa models are pre-trained, which allows researchers to focus on fine-tuning them for specific datasets. The initial step entails establishing a scheduler to ascertain the optimal learning rate, ultimately contributing to improved model performance. Determining the model's hyperparameters is crucial. The "Dropout" parameter, set at 0.1, suggests the use of dropout regularization with a rate of 10%, a common practice to prevent overfitting. The "Batch Size" is configured as 12, indicating that the model processes data in batches of 12 during training. The "Learning Rate (AdamW)" is specified as 0.0001, which controls the step size in the optimization process, with lower values often indicating a cautious learning approach. The "Epoch" parameter is set to 5, representing the number of complete passes through the training dataset. "Hidden Size" is noted as 758, potentially denoting the dimensionality of the hidden layers in the model architecture. Lastly, "Max Position Embeddings" is listed as 512, indicating the maximum positional information the model can consider, which is essential for tasks involving sequential data.

Dropout is a technique used in deep-learning model training to prevent overfitting by randomly disabling nodes. Batch size, the number of samples used in each epoch, affects performance; improper sizes may yield suboptimal results or extended epoch durations with minimal accuracy gains [73]. The learning rate impacts model accuracy and convergence [74,75]. An epoch, a single pass through the training data, indicates how often a dataset is trained on a model. Hidden size refers to the number of embedding representations, while maximum position embedding denotes the maximum number of words in a sentence eligible for embedding [11].

3.4. Topic Modeling Process

BERTopic, a method for topic modeling, was employed to harness the capabilities of BERT (Bidirectional Encoder Representations from Transformers) in extracting meaningful topics from textual data. The preprocessed data can be input data for BERTopic. The next step is the optimization of BERTopic's performance for our specific research objective, which entailed the fine-tuning of hyperparameters, encompassing critical adjustments such as determining the optimal number of topics and creating a meaningful topic.

The value of each parameter is based on the recommendation stated in BERTopic. Parameters specific to the BERTopic model, a method primarily used for topic modeling. "stop_words" are defined as "english", specifying the removal of common English stop words during the analysis, which can enhance the focus on more meaningful content. "top_n_words" is set at 5, signifying that the model focuses on the top five most important words within each document, based on the optimal number of terms is below 30 and TF-IDF score, contributing to the overall representation. "min_topic_size" is established as 20, designating the minimum number of documents required for a topic to be considered valid, with smaller topics potentially excluded. "nr_topics" is specified as 10, indicating the intended number of topics to be extracted from the text data, influencing the granularity of the topic modeling [59].

3.5. Model Evaluation

The model necessitates evaluation to assess its performance following fine-tuning with training data to assess the performance of multilabel classification and topic modeling tasks.

We utilize the metrics presented in Table 2 for multilabel classification. These metrics serve as an essential indicator of the model's effectiveness in handling multiple labels and provide a thorough evaluation of its overall performance [76,77].

BERTopic can be evaluated using coherence score, perplexity, and topic diversity. Coherence scores assess how semantically related the words within each topic are, essentially quantifying how well the words co-occur and form coherent topics. Higher coherence scores indicate more coherent and interpretable topics, implying that the words within each topic are closely related and represent a cohesive theme [76]. These scores are designed to distinguish between topics that are readily understood. On the other hand, perplexity scores measure how well a topic model can predict a set of documents or text data [78,79]. As for topic diversity, it captures the variety and distinctiveness of topics in a corpus. The goal of this study is to help clothing companies improve the product's quality by understanding the consumer's feedback thoroughly [59].

3.6. Multilabel Classification Task Model Selection

To substantiate the rationale for selecting RoBERTa, a comparative analysis of BERT and RoBERTa's performance is conducted utilizing our dataset. As depicted in Figure 1, the methodology for comparing these two models is outlined. Macro and micro metrics are employed for evaluation. The resulting macro and micro F1 scores reveal that BERT yields a score of 0.86, while RoBERTa achieves a score of 0.87. Based on these findings, the study adopts the RoBERTa model for the multilabel classification.

4. Results and Discussion

4.1. Multilabel Classification Results

This section describes the results of multilabel classification by fine-tuning RoBERTa. We use five epochs to fine-tune the model. Evaluation metrics are used on each label and the model's overall training process over five epochs. It records the training loss, validation loss, and the time taken during each epoch. In the initial epoch, the training loss begins at 0.39 and gradually decreases to 0.06 in the final epoch. This diminishing training loss indicates a reduction in the prediction errors made by the model as it is being trained on the dataset. Similarly, the validation loss commences at 0.34, experiences fluctuations, and stabilizes at 0.24 in the fourth and fifth epochs. The time taken for each epoch remains relatively consistent, ranging between 4 min and 33 s to 4 min and 35 s. This consistency suggests efficient model training without significant variations in the time required for each epoch.

Table 7 provides a comprehensive overview of the model's performance metrics, showcasing results for both the training and test datasets. These metrics are instrumental in evaluating the model's efficiency in making predictions or classifications. In the training phase, "MicroP" (Micro Precision) indicates a Precision rate of 0.86, highlighting that, on average, the model's predictions are accurate around 86% of the time. This Precision remains consistent in the test phase, affirming the model's reliability. "MacroP" (Macro Precision) emphasizes class-specific Precision scores, with 0.88 in the training phase, demonstrating the model's consistent Precision across individual classes. Although it slightly decreases to 0.86 in the test phase, the Precision level remains high. "MicroR" (Micro Recall) attains a score of 0.87 in both training and test phases, signifying that, on average, the model successfully recalls about 87% of relevant instances. "MacroR" (Macro Recall) showcases robust Recall across individual classes, reaching 0.86 in the training phase and increasing to 0.88 in the test phase. "MicroF1" (Micro F1 Score) achieves a high balance between Precision and Recall with a score of 0.89 in the training phase and maintains strength at 0.87 in the test phase. "MacroF1" (Macro F1 Score) affirms a consistent trade-off between Precision and Recall across various classes in both training and test phases, with a score of 0.87. "Accuracy" signifies the model's overall correctness, achieving 0.90 in the training phase and improving to 0.91 in the test phase, demonstrating high accuracy in predicting instances within the test dataset.

The noteworthy observation in this evaluation's context is that all the performance metrics considered have consistently surpassed the critical threshold of 0.8. Established as a benchmark, this threshold carries profound implications as it signifies a significant level of model performance. The model's ability to consistently achieve metrics exceeding this threshold implies a remarkable proficiency in sentence prediction, with an error rate well below 20%. This achievement highlights the model's robustness and reliability in natural language processing tasks, particularly in its role as an indispensable tool for effectively classifying issues related to the quality of clothing products. These exemplary results affirm the model's efficacy and underscore its invaluable utility, positioning it as a highly dependable resource in addressing and categorizing concerns about clothing quality.

Table 7. Multilabel classification evaluation metrics results.

Metrics	Training Score	Test Score
MicroP	0.86	0.86
MacroP	0.88	0.86
MicroR	0.87	0.87
MacroR	0.86	0.88
MicroF1	0.89	0.87
MacroF1	0.87	0.87
Accuracy	0.90	0.91

Table 8 presents each label's Precision, Recall, and F1-score values. Notably, despite having an equal number of labels, each label exhibits distinct evaluation metric values. The Precision values for the labels are as follows: Materials at 0.85, Construction at 0.84, Color at 0.87, Finishing with the lowest value of 0.83, and Durability with the highest score of 0.92. The Recall scores are 0.81 for Materials (the lowest Recall value), 0.85 for Construction, 0.95 for Color, 0.82 for Finishing, and 0.95 for Durability, with Color and Durability achieving the highest Recall scores. The F1 scores for the labels are 0.81 for Materials, 0.84 for Construction, 0.91 for Color, 0.83 for Finishing, and 0.94 for Durability, where Materials labels receive the lowest F1 score, and Durability obtains the highest F1 score. Although the differences are slight, they may be attributed to factors such as variations in label combinations within the data and the number of sentences. The difference in the Precision, Recall, and F1 score values necessitate the utilization of specialized multilabel classification metrics to evaluate the model's comprehensive performance.

Table 8. Evaluation metrics per label.

Label	Precision	Recall	F1-Score
Materials	0.85	0.81	0.83
Construction	0.84	0.85	0.84
Color	0.87	0.95	0.91
Finishing	0.83	0.82	0.83
Durability	0.92	0.95	0.94

The findings of the RoBERTa model demonstrate a commendable score using a relatively small dataset consisting of merely 3784 customer reviews. It is crucial to acknowledge that this dataset is considered modest in size, yet the RoBERTa model demonstrates its effectiveness even when the training dataset is constrained. This assertion is substantiated by the evaluation results, showing that the model's score surpasses 0.5 for individual quality dimensions and overall performance. This finding strongly suggests that the RoBERTa model holds the potential to reliably function as a valuable instrument for pinpointing product quality weaknesses in clothing companies. Additionally, the RoBERTa model's successful performance with a modest-sized dataset ensures promising implications for other NLP tasks. If the RoBERTa model demonstrates such impressive outcomes with a smaller dataset, it assures that performance can be further enhanced with extensive training data. This assurance could prove particularly valuable for industries or domains where acquiring extensive quantities of labeled data is essential.

The integration of RoBERTa can facilitate a proactive approach to the feedback loop. Manufacturers can utilize the model to predict potential material handling issues before they escalate, enabling preemptive measures that enhance the quality of products and their handling attributes. This predictive capability is especially valuable in optimizing logistics, as manufacturers can swiftly identify potential bottlenecks or vulnerabilities in the material handling process and take corrective actions.

4.2. Topic Modeling Results

This segment presents the top ten topics derived from topic modeling, determined by their highest frequency. To assess the performance of BERTopic in this study, we employ the coherence score metric. The model achieved a coherence score of 0.67 in 20 topics. We experimented with min_topic_size ranging from 10 to 22. We experiment with this set of numbers based on the minimum optimal value recommended by the author of BERTopic [59].

We reduce the topic to 10 to remove redundant or overlapping topics to represent most consumer reviews. To visualize the topics, we use five Topic Word Scores based on TF-IDF to find keywords in a particular topic and a Similarity Matrix to find the correlation between topics. It helps multiple issues to be addressed with one solution. BERTopic manages to give a decent coherence score, and it is above 0.5. The highest coherence score means BERTopic can produce a cluster of words that form topics.

Figure 4 presents ten topics, each featuring the top five words based on their highest TF-IDF scores, which indicate word relevance and provide insight into each topic's general theme. Topic 0 addresses attributes like softness, warmth, and color consumers value in sweaters. Topic 1 focuses on the feel of white T-shirts. Topic 2 examines thin fabrics associated with inexpensive clothing. Topic 3 reviews work pants with high durability needs in the sensitive groin area. Topic 4 covers winter jackets and coats desired for warmth. Topic 5 reveals consumer disappointment with dresses featuring fragile back zippers. Topic 6 discusses color preferences for attractive dresses. Topic 7 highlights the common purchase of medium-sized clothing, though consumers often find it too small and lightweight. Topic 8 explores tank top attributes, such as color and detail. Topic 9 considers the appearance of consumers' buttocks when wearing dresses and skirts. The topic encapsulates consumer desires and requirements in product reviews, assisting companies in achieving sustainability objectives. By integrating consumer feedback, businesses can develop products better tailored to meet needs, minimizing disposal and related environmental impacts. This strategy fosters sustainability and curbs the fashion industry's waste generation. Concentrating on fulfilling consumers' needs allows companies to decrease waste and promote eco-friendly production.

Figure 5 shows a similarity between topics. It is important to qualitatively analyze the words that appear on the topic based on domain experts. The highest similarity is between Topic 0 and Topic 8, with a similarity score of 0.66. As for the topic with the lowest similarity score between Topic 3 and Topic 4, these generally discuss the characteristics required by clothing, according to its function, for companies to be included in the top and sweater attributes. Topic 3 discusses the durability of work pants. In contrast, Topic 4 discusses whether winter jackets should provide warmth for users; these two topics each discuss clothing in different contexts, and Topic 3 discusses clothes worn in winter. In contrast, Topic 4 discusses clothing used while working. Another correlation worth mentioning is Topic 0 and 2, as it can be interpreted that a sweater that is made from fabric is often soft and thin with a lot of color variation while being cheap.

A higher similarity score's importance lies in its ability to pinpoint sustainable solutions across various topics. Companies can create versatile solutions when there is a high degree of similarity between problems in different areas, reducing research and development costs. This method preserves resources and fosters sustainability by lessening the creation of potentially harmful products or processes. Employing a similarity-focused approach helps companies lower their environmental impact and foster a sustainable future. Thus, a high similarity score is valuable for its practical uses and for encouraging eco-friendly business practices. For example, topics 0 and 8 strongly correlate due to the word "color". Companies can address related issues in both clothing categories using uniform color codes and natural dyes, ensuring color consistency without procuring additional raw materials. Additionally, Topics 0 and 2 are essential for incorporating improved material handling considerations into material selection, which is crucial for clothing manufacturers, particularly when dealing with diverse textiles, such as thin fabrics that cannot be made into sweaters. The materials require careful attention to ensure that the handling processes, from manufacturing to distribution, do not compromise their integrity or quality over time.

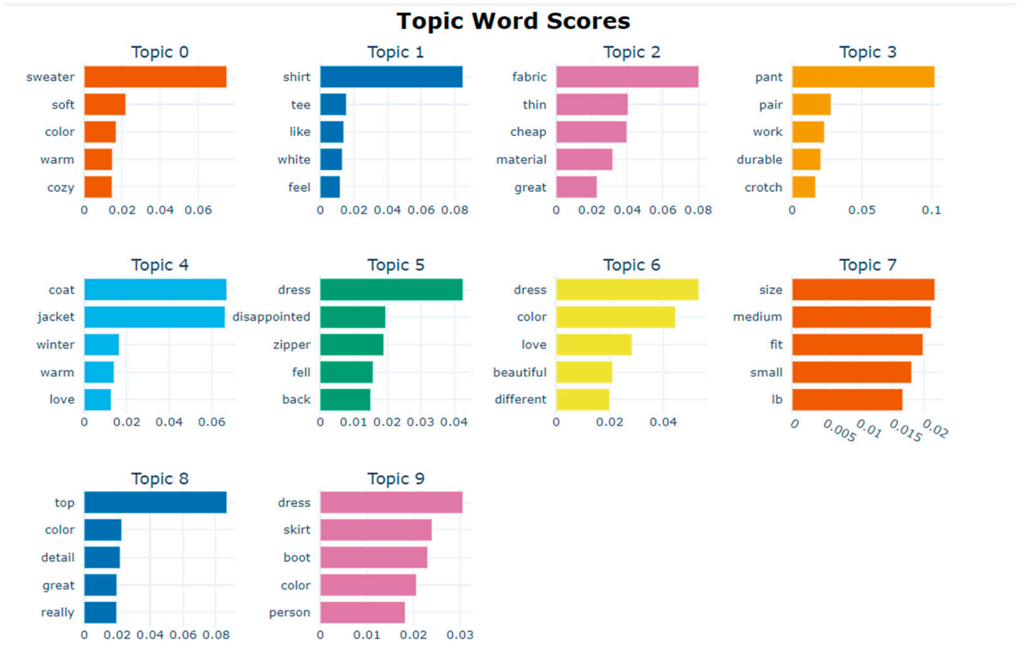

Figure 4. Top Ten Topics.

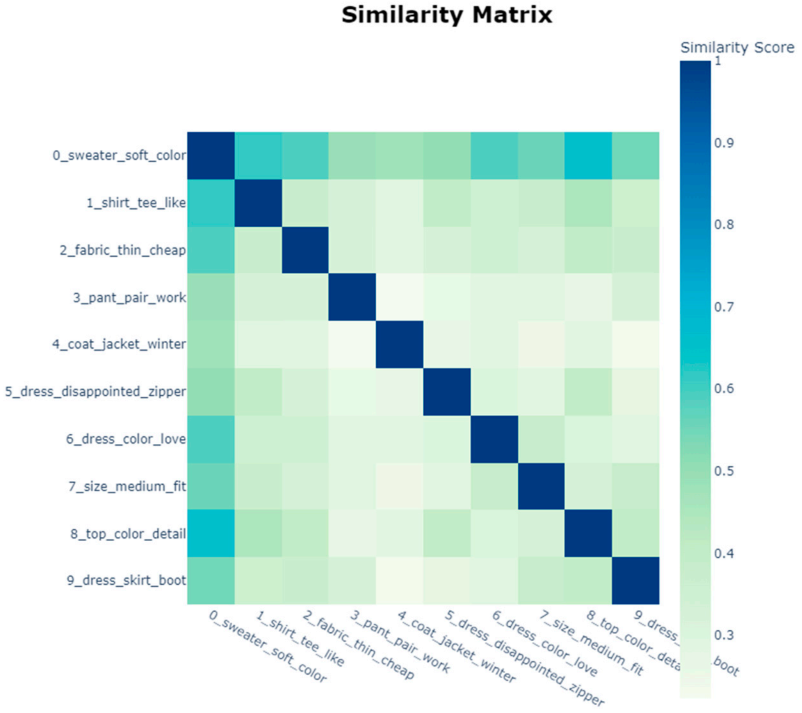

Figure 5. Similarity Matrix.

5. Conclusions

Companies should examine ever-growing and diverse consumer reviews for product development ideas to enhance quality and minimize waste. This study suggests a novel approach by merging multilabel classification using a modified RoBERTa model and topic modeling with BERTopic. The combination enables companies to comprehend customer feedback better, pinpoint improvement areas, and reduce waste by developing products in line with consumer preferences.

The research demonstrates RoBERTa's powerful performance in multilabel classification tasks, separating consumer reviews based on clothing quality dimensions. Metrics such as micro and macro-averaged Precision, Recall, and F1 scores support these findings. Additionally, BERTopic's topic modeling offers valuable insights, as each topic contains quality-related keywords frequently reviewed on Amazon's e-commerce platform. We discovered that distinct clothing types have different topics and quality-related issues.

Integrating RoBERTa opens opportunities for a proactive approach to addressing product quality. Manufacturers can use the model to predict material handling issues before they escalate, allowing them to take swift corrective actions, which is particularly valuable in optimizing logistics. BERTopic for topic modeling offers a practical framework for extracting meaningful topics from consumer feedback. Its coherence score of 0.67 at 20 topics effectively captures vital aspects of clothing products valued or criticized by consumers. These insights empower businesses to tailor product development strategies, reducing waste and promoting eco-friendly production practices that align with sustainability goals. Analyzing topic similarities helps companies create versatile and sustainable solutions. High similarity scores between different problems guide the development of unified solutions, lowering research and development costs and reducing environmental impact. For instance, the correlation between topics discussing color consistency encourages the adoption of uniform color codes and natural dyes. Understanding topic relationships, such as those related to material handling, informs material choices to maintain quality during handling processes.

This study acknowledges certain limitations. First, the data quantity used for pretraining models and topic modeling may not fully encompass all clothing quality aspects, warranting more extensive data in future research. Second, this study did not compare other BERT variations or alternative language models. Despite these limitations, the research contributes to the understanding that the RoBERTa model effectively interprets clothing quality-related sentences and excels in multilabel classification tasks. Additionally, BERTopic can extract meaningful topics. Overall, clothing companies can utilize this model to improve product development by leveraging consumer reviews and identifying quality areas needing enhancement. The model boasts high accuracy and rapid analysis of vast data volumes.

Author Contributions: Conceptualization, A.A.; Methodology N.D.G.; Validation A.A.; Writing—Original Draft Preparation, A.A. and N.D.G.; Writing—Review & Editing, A.A. and N.D.G.; Visualization, N.D.G.; Supervision, A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are available to the public.

Acknowledgments: The staff at Telkom University research support services were instrumental in helping us access valuable resources and literature related to our research topic.

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

References

1. Amed, I.; Berg, A.; Balchandani, A.; Hedrich, S.; Jensen, J.E.; Straub, M.; Rölkens, F.; Young, R.; Brown, P.; Merle, L.L.; et al. *The State of Fashion 2022*; McKinsey & Company: New York, NY, USA, 2022; pp. 1–144.
2. Piippo, R.; Niinimäki, K.; Aakko, M. Fit for the Future: Garment Quality and Product Lifetimes in a CE Context. *Sustainability* **2022**, *14*, 726. [CrossRef]
3. Weber, S.; Lynes, J.; Young, S.B. Fashion interest as a driver for consumer textile waste management: Reuse, recycle or disposal. *Int. J. Consum. Stud.* **2017**, *41*, 207–215. [CrossRef]
4. Bhardwaj, V.; Fairhurst, A. Fast fashion: Response to changes in the fashion industry. *Int. Rev. Retail. Distrib. Consum. Res.* **2010**, *20*, 165–173. [CrossRef]

5. Macchion, L.; Moretto, A.M.; Caniato, F.; Caridi, M.; Danese, P.; Vinelli, A. International e-commerce for fashion products: What is the relationship with performance? *Int. J. Retail Distrib. Manag.* **2017**, *45*, 1011–1031. [CrossRef]
6. Statista. eCommerce Report 2021-Fashion Statista Digital Market Outlook-Segment Report Bilder Immer Einfärben in: Blue, Accent Color 1. No. June, 2021. Available online: https://www.statista.com/study/38340/ecommerce-report-fashion/ (accessed on 1 October 2022).
7. Ho-Dac, N.N. The value of online user generated content in product development. *J. Bus. Res.* **2020**, *112*, 136–146. [CrossRef]
8. Hong, Y.; Shao, X. Emotional Analysis of Clothing Product Reviews Based on Machine Learning. In Proceedings of the 2021 3rd International Conference on Applied Machine Learning (ICAML), Changsha, China, 23–25 July 2021; pp. 398–401. [CrossRef]
9. Satinet, C.; Fouss, F. A Supervised Machine Learning Classification Framework for Clothing Products' Sustainability. *Sustainability* **2022**, *14*, 1334. [CrossRef]
10. Cai, L.; Song, Y.; Liu, T.; Zhang, K. A Hybrid BERT Model That Incorporates Label Semantics via Adjustive Attention for Multi-Label Text Classification. *IEEE Access* **2020**, *8*, 152183–152192. [CrossRef]
11. Rahmawati, A.; Alamsyah, A.; Romadhony, A. Hoax News Detection Analysis using IndoBERT Deep Learning Methodology. In Proceedings of the 2022 10th International Conference on Information and Communication Technology (ICoICT), Bandung, Indonesia, 2–3 August 2022; pp. 368–373. [CrossRef]
12. Dudija, N.; Natalia, L.; Alamsyah, A.; Romadhony, A. Identification of Extraversion and Neuroticism Personality Dimensions Using IndoBERT's Deep Learning Model. In Proceedings of the 2022 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT), Bali, Indonesia, 28–30 July 2022; pp. 155–159. [CrossRef]
13. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the NAACL-HLT 2019, Minneapolis, MN, USA, 2–7 June 2019; Volume 1, pp. 4171–4186.
14. Elleuch, M.; Mezghani, A.; Khemakhem, M.; Kherallah, M. Clothing classification using deep cnn architecture based on transfer learning. *Adv. Intell. Syst. Comput.* **2021**, *1179*, 240–248. [CrossRef]
15. Dirting, B.D.; Chukwudebe, G.A.; Nwokorie, E.C.; Ayogu, I.I. Multi-Label Classification of Hate Speech Severity on Social Media using BERT Model. In Proceedings of the 2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development, NIGERCON 2022, Lagos, Nigeria, 5–7 April 2022. [CrossRef]
16. Binotto, C.; Payne, A. The Poetics of Waste: Contemporary Fashion Practice in the Context of Wastefulness. *Fash. Pr.* **2016**, *9*, 5–29. [CrossRef]
17. Shirvanimoghaddam, K.; Motamed, B.; Ramakrishna, S.; Naebe, M. Death by waste: Fashion and textile circular economy case. *Sci. Total Environ.* **2020**, *718*, 137317. [CrossRef]
18. Martin, J.; Elg, M.; Gremyr, I. The Many Meanings of Quality: Towards a Definition in Support of Sustainable Operations. *Total Qual. Manag. Bus. Excell.* **2020**, 1–14. [CrossRef]
19. Haule, L.V.; Nambela, L. Sustainable application of nanomaterial for finishing of textile material. In *Green Nanomaterials for Industrial Applications*; Elsevier: Amsterdam, The Netherlands, 2022; pp. 177–206. [CrossRef]
20. Goworek, H.; Oxborrow, L.; Claxton, S.; McLaren, A.; Cooper, T.; Hill, H. Managing sustainability in the fashion business: Challenges in product development for clothing longevity in the UK. *J. Bus. Res.* **2018**, *117*, 629–641. [CrossRef]
21. AOlad, A.; Valilai, O.F. Using of social media data analytics for applying digital twins in product development. In Proceedings of the IEEE International Conference on Industrial Engineering and Engineering Management, Singapore, 14–17 December 2020; pp. 319–323. [CrossRef]
22. Kumar, P.; Ramanan, T.R.; Keelath, M.S. The mediating role of quality management capability on the dynamic capability—New product development performance relationship: An empirical study among new product development units in the electronics sector. *Qual. Manag. J.* **2020**, *27*, 80–94. [CrossRef]
23. Manz, S. Chapter 14—Alignment. In *Medical Device Quality Management Systems: Strategy and Techniques for Improving Efficiency and Effectiveness*; Academic Press: Havre De Grace, MD, USA, 2019; pp. 197–204. [CrossRef]
24. Shen, B.; Chen, C. Quality management in outsourced global fashion supply chains: An exploratory case study. *Prod. Plan. Control.* **2019**, *31*, 757–769. [CrossRef]
25. Liu, J.; Zhuang, D.; Shen, W. The impact of quality management practices on manufacturing performance: An empirical study based on system theory. *Soft Comput.* **2022**, *27*, 4077–4092. [CrossRef]
26. Bartholomew, D.J. *Encyclopedia of Operations Research and Management Science*; Springer Science and Business Media LLC: Dordrecht, The Netherlands, 2013. [CrossRef]
27. Hunter, L.; Fan, J. *Adding Functionality to Garments*; Elsevier Ltd.: Amsterdam, The Netherlands, 2015. [CrossRef]
28. Aakko, M.; Niinimäki, K. Quality matters: Reviewing the connections between perceived quality and clothing use time. *J. Fash. Mark. Manag. Int. J.* **2021**, *26*, 107–125. [CrossRef]
29. Motlogelwa, S. *Comfort and Durability in High-Performance Clothing*; Elsevier Ltd.: Amsterdam, The Netherlands, 2017. [CrossRef]
30. Xie, J.; Burstein, F. Using machine learning to support resource quality assessment: An adaptive attribute-based approach for health information portals. In *Database Systems for Advanced Applications, Proceedings of the 16th International Conference, DASFAA 2011, International Workshops: GDB, SIM3, FlashDB, SNSMW, DaMEN, DQIS, Hong Kong, China, 22–25 April 2011*; Springer: Berlin/Heidelberg, Germany, 2011; Volume 6637, pp. 526–537. [CrossRef]

31. Alamsyah, A.; Friscintia, P.B.A. Artificial neural network for Indonesian tourism demand forecasting. In Proceedings of the 7th International Conference on Information and Communication Technology, ICoICT 2019, Kuala Lumpur, Malaysia, 24–26 July 2019; pp. 1–7. [CrossRef]
32. Liu, X.X.; Chen, Z.Y. Service quality evaluation and service improvement using online reviews: A framework combining deep learning with a hierarchical service quality model. *Electron. Commer. Res. Appl.* **2022**, *54*, 101174. [CrossRef]
33. Guo, L.; Jin, B.; Yu, R.; Yao, C.; Sun, C.; Huang, D. Multi-label classification methods for green computing and application for mobile medical recommendations. *IEEE Access* **2016**, *4*, 3201–3209. [CrossRef]
34. Deniz, E.; Erbay, H.; Coşar, M. Multi-Label Classification of E-Commerce Customer Reviews via Machine Learning. *Axioms* **2022**, *11*, 436. [CrossRef]
35. Pereira, R.B.; Plastino, A.; Zadrozny, B.; Merschmann, L.H.C. Correlation analysis of performance measures for multi-label classification. *Inf. Process. Manag.* **2018**, *54*, 359–369. [CrossRef]
36. Wei, X.; Huang, J.; Zhao, R.; Yu, H.; Xu, Z. Multi-Label Text Classification Model Based on Multi-Level Constraint Augmentation and Label Association Attention. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **2023**. [CrossRef]
37. Lin, N.; Fu, S.; Lin, X.; Wang, L. Multi-label emotion classification based on adversarial multi-task learning. *Inf. Process. Manag.* **2022**, *59*, 103097. [CrossRef]
38. Read, J.; Pfahringer, B.; Holmes, G.; Frank, E. Classifier chains for multi-label classification. *Mach. Learn.* **2011**, *85*, 333–359. [CrossRef]
39. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 5999–6009.
40. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, 16–20 November 2020; pp. 38–45. [CrossRef]
41. Alerskans, E.; Nyborg, J.; Birk, M.; Kaas, E. A transformer neural network for predicting near-surface temperature. *Meteorol. Appl.* **2022**, *29*, e2098. [CrossRef]
42. Nechikkat, M.I.; Pattilikattil, B.V.V.; Varma, S.; James, A. Video Captioning Using Transformer Network. *AIP Conf. Proc.* **2022**, *2494*, 050003. [CrossRef]
43. Li, Z.; Jiao, Z.; He, A.; Xu, N. A denoising-classification neural network for power transformer protection. *Prot. Control. Mod. Power Syst.* **2022**, *7*, 52. [CrossRef]
44. Atiea, M.A.; Adel, M. Transformer-based Neural Network for Electrocardiogram Classification. *Int. J. Adv. Comput. Sci. Appl.* **2022**, *13*, 357–363. [CrossRef]
45. Pitz, E.; Pochiraju, K. A Neural Network Transformer Model for Composite Microstructure Homogenization. *arXiv* **2023**, arXiv:2304.07877.
46. Arroyo, R.; Jiménez-Cabello, D.; Martínez-Cebrián, J. Multi-Label Classification of Promotions in Digital Leaflets Using Textual and Visual Information. 2020. Available online: http://arxiv.org/abs/2010.03331 (accessed on 28 October 2022).
47. Lee, J.S.; Hsiang, J. Patent classification by fine-tuning BERT language model. *World Pat. Inf.* **2020**, *61*, 101965. [CrossRef]
48. Biswas, J.; Rahman, M.M.; Biswas, A.A.; Khan, M.A.Z.; Rajbongshi, A.; Niloy, H.A. Sentiment Analysis on User Reaction for Online Food Delivery Services using BERT Model. In Proceedings of the 2021 7th International Conference on Advanced Computing and Communication Systems, ICACCS 2021, Coimbatore, India, 19–20 March 2021; pp. 1019–1023. [CrossRef]
49. Heidari, M.; Rafatirad, S. Semantic Convolutional Neural Network model for Safe Business Investment by Using BERT. In Proceedings of the 2020 7th International Conference on Social Network Analysis, Management and Security, SNAMS 2020, Paris, France, 14–16 December 2020. [CrossRef]
50. Bilal, M.; Almazroi, A.A. Effectiveness of Fine-tuned BERT Model in Classification of Helpful and Unhelpful Online Customer Reviews. *Electron. Commer. Res.* **2022**, 1–21. [CrossRef]
51. Cao, Y.; Sun, Z.; Li, L.; Mo, W. A Study of Sentiment Analysis Algorithms for Agricultural Product Reviews Based on Improved BERT Model. *Symmetry* **2022**, *14*, 1604. [CrossRef]
52. Malik, M.S.I.; Nazarova, A.; Jamjoom, M.M.; Ignatov, D.I. Multilingual hope speech detection: A Robust framework using transfer learning of fine-tuning RoBERTa model. *J. King Saud Univ.—Comput. Inf. Sci.* **2023**, *35*, 101736. [CrossRef]
53. You, L.; Han, F.; Peng, J.; Jin, H.; Claramunt, C. ASK-RoBERTa: A pretraining model for aspect-based sentiment classification via sentiment knowledge mining. *Knowl.-Based Syst.* **2022**, *253*, 109511. [CrossRef]
54. Pavlov, T.; Mirceva, G. COVID-19 Fake News Detection by Using BERT and RoBERTa models. In Proceedings of the 2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO), Opatija, Croatia, 23–27 May 2022; pp. 312–316. [CrossRef]
55. Cortiz, D. Exploring Transformers models for Emotion Recognition: A comparision of BERT, DistilBERT, RoBERTa, XLNET and ELECTRA. In Proceedings of the 2022 3rd International Conference on Control, Robotics and Intelligent System, Virtual, 26–28 August 2022; pp. 230–234. [CrossRef]
56. Gupta, P.; Gandhi, S.; Chakravarthi, B.R. Leveraging Transfer learning techniques- BERT, RoBERTa, ALBERT and DistilBERT for Fake Review Detection. In Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation, Virtual, 13–17 December 2021; pp. 75–82. [CrossRef]

57. Rajapaksha, P.; Farahbakhsh, R.; Crespi, N. BERT, XLNet or RoBERTa: The Best Transfer Learning Model to Detect Clickbaits. *IEEE Access* **2021**, *9*, 154704–154716. [CrossRef]
58. Naseer, M.; Asvial, M.; Sari, R.F. An Empirical Comparison of BERT, RoBERTa, and Electra for Fact Verification. In Proceedings of the 3rd International Conference on Artificial Intelligence in Information and Communication, Marrakesh, Morocco, 5–7 May 2021; pp. 241–246. [CrossRef]
59. Grootendorst, M. BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure. 2022. Available online: http://arxiv.org/abs/2203.05794 (accessed on 10 August 2023).
60. Anwar, A.; Ilyas, H.; Yaqub, U.; Zaman, S. Analyzing QAnon on Twitter in Context of US Elections 2020: Analysis of User Messages and Profiles Using VADER and BERT Topic modeling. In Proceedings of the 22nd Annual International Conference on Digital Government Research, Omaha, NE, USA, 9–11 June 2021; pp. 82–88. [CrossRef]
61. Ozdemirci, S.M.; Turan, M. Case Study on well-known Topic Modeling Methods for Document Classification. In Proceedings of the 6th International Conference on Inventive Computation Technologies, ICICT 2021, Coimbatore, India, 20–22 January 2021; pp. 1304–1309. [CrossRef]
62. Aytaç, E.; Khayet, M. A Topic Modeling Approach to Discover the Global and Local Subjects in Membrane Distillation Separation Process. *Separations* **2023**, *10*, 482. [CrossRef]
63. Bu, W.; Shu, H.; Kang, F.; Hu, Q.; Zhao, Y. Software Subclassification Based on BERTopic-BERT-BiLSTM Model. *Electronics* **2023**, *12*, 3798. [CrossRef]
64. Zankadi, H.; Idrissi, A.; Daoudi, N.; Hilal, I. Identifying learners' topical interests from social media content to enrich their course preferences in MOOCs using topic modeling and NLP techniques. *Educ. Inf. Technol.* **2022**, *28*, 5567–5584. [CrossRef]
65. Thompson, L.; Mimno, D. Topic Modeling with Contextualized Word Representation Clusters. 2020. Available online: http://arxiv.org/abs/2010.12626 (accessed on 5 November 2022).
66. de Groot, M.; Aliannejadi, M.; Haas, M.R. Experiments on Generalizability of BERTopic on Multi-Domain Short Text. 2022, pp. 1–3. Available online: http://arxiv.org/abs/2212.08459 (accessed on 31 July 2023).
67. Getwebooster. AMZReviews—Amazon Review Scraper. 2023. Available online: https://chrome.google.com/webstore/detail/amzreviews-amazon-review/epnapacjpnonncagggmmhppncbmnpecl (accessed on 15 October 2023).
68. Girawan, N.; Alamsyah, A. Consumer Review of Clothing Product. *Mendeley Data* **2023**. [CrossRef]
69. NICAPOTATO. Women's E-Commerce Clothing Reviews. 2023. Available online: https://www.kaggle.com/datasets/nicapotato/womens-ecommerce-clothing-reviews/data (accessed on 15 October 2023).
70. Scarpino, I.; Zucco, C.; Vallelunga, R.; Luzza, F.; Cannataro, M. Investigating Topic Modeling Techniques to Extract Meaningful Insights in Italian Long COVID Narration. *BioTech* **2022**, *11*, 41. [CrossRef]
71. Lamba, M.; Madhusudhan, M. Text Pre-Processing. In *Text Mining for Information Professionals*; Springer International Publishing: Cham, Switzerland, 2022; pp. 79–103. [CrossRef]
72. Meng, Z.; McCreadie, R.; MacDonald, C.; Ounis, I. Exploring Data Splitting Strategies for the Evaluation of Recommendation Models. In Proceedings of the RecSys 2020—14th ACM Conference on Recommender Systems, Virtual Event, Brazil, 22–26 September 2020; pp. 681–686. [CrossRef]
73. Kandel, I.; Castelli, M. The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset. *ICT Express* **2020**, *6*, 312–315. [CrossRef]
74. Yu, C.; Qi, X.; Ma, H.; He, X.; Wang, C.; Zhao, Y. LLR: Learning learning rates by LSTM for training neural networks. *Neurocomputing* **2020**, *394*, 41–50. [CrossRef]
75. Zhao, H.; Liu, F.; Zhang, H.; Liang, Z. Research on a learning rate with energy index in deep learning. *Neural Netw.* **2019**, *110*, 225–231. [CrossRef] [PubMed]
76. He, J.; Weerkamp, W.; Larson, M.; de Rijke, M. An effective coherence measure to determine topical consistency in user-generated content. *Int. J. Doc. Anal. Recognit.* **2009**, *12*, 185–203. [CrossRef]
77. Heydarian, M.; Doyle, T.E. MLCM: Multi-Label Confusion Matrix. *IEEE Access* **2022**, *10*, 19083–19095. [CrossRef]
78. Lau, J.H.; Baldwin, T. The Sensitivity of Topic Coherence Evaluation to Topic Cardinality. Available online: https://github.com/jhlau/ (accessed on 31 July 2023).
79. Colla, D.; Delsanto, M.; Agosto, M.; Vitiello, B.; Radicioni, D.P. Semantic coherence markers: The contribution of perplexity metrics. *Artif. Intell. Med.* **2022**, *134*, 102393. [CrossRef]

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.