JURNAL INFOTEL
VOL. 16, NO. 1, FEBRUARY 2024, PP. 58–70. DOI:10.20895/INFOTEL.V16I1.1070

RESEARCH ARTICLE

Understanding Customer Perception of Local Fashion Products on Online Marketplace through Content Analysis

Imam Adi Nata[1,*] and Muhammad Rifqi Maarif[2]

[1]Information Technology, Tidar University, Magelang 56116, Indonesia
[2]Industrial Engineering, Tidar University, Magelang 56116, Indonesia

*Corresponding email: imamadinata@untidar.ac.id

Received: November 11, 2023; Revised: November 24, 2023; Accepted: January 22, 2024.

Abstract: This research employs natural language processing (NLP) techniques to evaluate customer reviews from online marketplaces. It uses keyword extraction and clustering to identify thematic clusters in the data. The notable k-means algorithm was used in this research for keyword clustering purposes. These clusters reveal shared contextual significance and provide a higher-level perspective on customer perceptions of local fashion products. Sentiment analysis is also conducted within each theme to understand customer sentiment. This approach goes beyond binary sentiment classification and offers a more nuanced analysis. This research provides a thorough framework for comprehending customer perceptions in the digital marketplace by incorporating keyword extraction, clustering, and sentiment analysis. It contributes to the field of e-commerce by offering a robust methodology for decoding customer sentiments towards local fashion products. The findings have substantial implications for marketers, designers, and platform providers in online marketplaces, leading to a more consumer-centric e-commerce ecosystem.

Keywords: customer review, keyword extraction, keyword clustering, sentiment analysis

1 Introduction

In recent years, online shopping has become increasingly popular, and this trend is strongly evident in the fashion industry. As a result, online marketplaces have emerged as highly active centers of consumer engagement, facilitating access to a wide range of products. On the other hand, with its rich and diverse culture, Indonesia has seen a rise in local fashion

 ISSN: 2085-3688; e-ISSN: 2460-0997

brands making their mark in online marketplaces. These platforms offer a unique opportunity for local designers to showcase their products to a broader audience. Understanding how consumers perceive and interact with these products online is essential, yet this area has yet to be thoroughly explored, especially in the Indonesian context.

The intersection of e-commerce, consumer behavior, and sentiment analysis has become a topic of negligible interest in contemporary research within the academic community. Previous investigations have examined various aspects of customer perception and interaction within online marketplaces, thereby providing valuable insights into the evolving dynamics of digital commerce. A study by Kim [1] provided a sentiment classification in customer reviews, laying down a foundational framework for extracting valuable consumer insights from textual data. Other studies by Park *et al.* [2] and Lucini *et al.* [3] expanded the scope of research by implementing clustering algorithms to group keywords based on shared context. However, it is worth mentioning that their applications were primarily centered around customer feedback in the hospitality industry and did not extend into the realm of fashion products. Even though these studies have contributed to the existing body of knowledge, they have yet to tackle the nuances and complexities linked to customer sentiments in the fashion context. Another work by Kiakatswin *et al.* [4] employing latent dirichlet allocation (LDA) unveiled latent themes in consumer feedback, leading to a deeper comprehension of product sentiments.

Several studies mentioned in the previous paragraph have looked into online consumer behavior. Detailed exploration of keyword extraction techniques must still address a gap. While discussing topic modeling sentiment analysis and keyword identification, those studies did not extend their methodologies to include the clustering of these keywords for deeper thematic analysis. This gap is a crucial area for further research, where more specific analytical techniques to identify the keywords could reveal richer, more detailed insights into consumer behavior within the fashion industry.

This study aims to address these gaps in the literature by utilizing natural language processing (NLP) techniques to conduct a detailed exploration of customer reviews obtained from popular online marketplaces. We aim to disentangle the intricate tapestry of customer sentiments, preferences, and opinions by applying advanced keyword extraction methodologies. Based on the attributes and sentiments derived from the NLP analysis, a visual graph is formulated, effectively showcasing the attribute and sentiment data. Finally, a qualitative analysis is conducted to effectively interpret the visual graph derived from the entire research process.

2 Research Methods

In this study, we utilize online customer review data about local shoe products available in various marketplaces within Indonesia. We present Figure 1 to illustrate the step-by-step experiments conducted in this work. The research process commences with retrieving customer data from the marketplace platform, which is accomplished through the employment of the web scrapping technique for data collection purposes. Subsequently, the collected data undergoes a cleaning process before being analyzed using the esteemed NLP techniques.

The second phase of this research includes utilizing NLP techniques in the data processing phase, which ultimately results in extracting keywords that are considered repre-

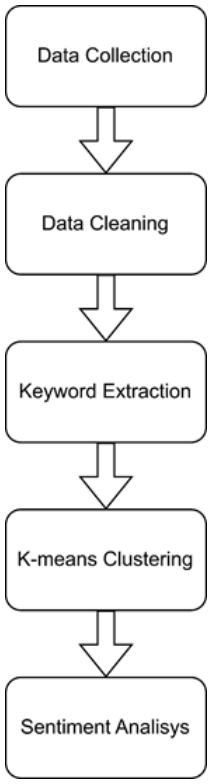

Figure 1: Flow of research.

sentations of product attributes. After the keywords were successfully extracted, the study implements the k-means clustering algorithm, which enables the formation of coherent thematic clusters that encapsulate shared contextual significance. This transformative process distills the raw data into abstract themes, providing a higher-level perspective on customer perceptions. By clustering the extracted keywords, we can uncover hidden patterns and connections that might not be immediately obvious, enriching our understanding of the underlying themes within the customer reviews.

Once thematic clusters are formed, sentiment analysis is conducted on each identified theme. This allows for a quantitative measure of customer sentiment within specific thematic domains, providing a more comprehensive picture of customer perceptions. By analyzing sentiment within different thematic clusters, we can gain valuable insights into the customer's sentiment towards critical aspects of local fashion offerings, such as quality, price, and customer service. The results of this study can inform decision-making processes by offering data-driven insights that cater to the specific needs and preferences of customers in the domain of local fashion products.

2.1 Data Acquisition and Preparation

Figure 1 shows an overview of the system in the monitoring system. The sensor system is inserted in the hatchery area In this study, the customer review data was obtained from the official stores of local shoe brands in the marketplace. Specifically, the chosen marketplace for this research was Shopee Indonesia, which was selected based on its ranking as the second most visited marketplace in Indonesia according to data from similar websites. We used the scrapping technique on the website www.shopee.co.id to gather the necessary data. This technique allowed us to extract online customer reviews for each product that was the focus of our research. We collected a substantial amount of data, specifically 243,487 review data points, encompassing 534 different products from the six selected local shoe brands. The average review length was 9.34 words, resulting in a total of 2,263,984 words across all the reviews. However, it is essential to note that only a fraction of the collected data, less than 50 % or 110,107 data points, contained actual reviews.

Before the collected data could be utilized for further analysis, it was necessary to clean it using predefined stages. These data-cleaning stages, as outlined in the study by Anandarajan *et al*. [5], include the following:

1. Case folding: This process involves converting all letters in the document to a consistent form, either by converting capital letters to lowercase or vice versa. This step was crucial due to the case sensitivity of the Python programming language, which treats uppercase and lowercase letters as distinct characters.
2. Removal of punctuation marks, numbers, and non-letter characters: To focus solely on The textual content of the reviews and all extraneous characters were removed from the dataset.
3. Eliminating stop words: Stop words, such as "*yang*" and "in," which do not have significant meaning, were removed from the dataset. This step was done using the stopword list provided by the Sastrawi programming library, a widely used Python library for processing Indonesian language texts.
4. Stemming/normalization: Words were transformed into their base form through stemming or normalization. This step was also performed using the Sastrawi library.
5. Merge Similar Words: This stage involved combining words that had the same meaning but were expressed differently, such as merging the words "Seller" and "Seller" into a single form. This step aimed to ensure consistency and eliminate redundancy within the dataset.

2.2 Keywords Extraction

The NLP method is employed for data processing techniques for extracting the keywords. These are essentially keywords found in user reviews pertaining to one of the marketing mix attributes. The process of extracting these keywords involves the selection of words in the review that function as nouns. To filter out words that are nouns, the part-of-speech (POS) technique is utilized within the realm of NLP. Part of speech tagging (POS-tagging) is a precious process that automatically assigns a word class label to each word in a sentence [6].

The results obtained from POS tagging significantly impact the output of the parsing process. The challenge that arises pertains to getting the correct word class labeling within the context of each sentence. The POS tagging algorithm diligently annotates words based

Figure 2: Wordcloud of extracted keyword candidates.

on their type and position within the sentence, whether they are nouns (NN), noun phrases (NNP), and so forth.

In this research, the hidden Markov model (HMM) algorithm, readily available within the Python library known as the natural language processing toolkit (NLTKhas), has been utilized to create the POS Tagging model. The creation of this POS tagging model involves the usage of publicly available Indonesian POS tagging training data [7]. The Hidden Markov Model consists of two distinct parts, namely the observed state and the hidden state. The experimental state represents the portion that can be directly observed, while the hidden state represents the portion that cannot be observed [8]. In the case of Part of Speech Tagging, the order of word classes cannot be directly observed; it is treated as the hidden state, whereas the order of words serves as the observed state. The main objective is to determine the most appropriate order of word classes based on the order of words [9].

2.3 Keyword Clustering

To cluster the keywords, the initial step involves constructing an attribute similarity metric that comprehends the inherent grouping of keywords. Therefore, a method is required to evaluate the similarity between these keywords based on the context in which they manifest within user reviews. This similarity calculation necessitates the conversion of the keyword into a numerical vector. The technique commonly employed for this conversion process is known as word2vec [10]. Word2vec is a type of word embedding that takes the form of a weight matrix acquired from the results of unsupervised learning neural network training [11]. By representing words as vectors, word2vec effectively captures the semantic essence of each word. This word embedding model is an example of an unsupervised learning application employing a neural network consisting of hidden and fully connected layers [12].

Wod2vec relies on local information derived from the language itself. The surrounding words influence the learned semantics of a specific word. This model aptly demonstrates its ability to discern linguistic patterns by establishing linear vector relationships [13]. Within the confines of this research, creating a word2vec model from keywords extracted from all review data is accomplished by utilizing the word2vec algorithm library. This library has been implemented in Python via the Gensim library, serving as a valuable resource.

Utilizing the word2vec algorithm, similarity metrics were constructed to aid keyword clustering. This critical step allows for the discovery of patterns within the dataset. The k-means algorithm was then applied to group the keywords that have shared closely related contexts. The algorithm divides the data into distinct clusters based on their proximity to the mean [14]. In the context of keyword clustering, this approach allowed for grouping semantically related terms, providing a more detailed understanding of the concepts and themes present in the dataset [15]. By applying k-means to the precomputed similarity metrics, we could identify specific clusters of keywords, which represented their semantic associations and improved the overall interpretability of the data [16]. This multi-step process, which involves constructing similarity metrics using word2vec and clustering the keywords using k-means, is a robust framework for uncovering valuable insights from text data.

2.4 Sentiment Analysis

Sentiment analysis is NLP, a technique to ascertain whether data contains positive or negative sentiment. In this research, sentiment analysis is utilized to evaluate the emotional context of the user related to the particular keywords present within a review [17]. Currently, two standard methods are employed in sentiment analysis: term-based sentiment analysis and machine learning [18]. As part of this research, the term-based sentiment analysis approach is selected for analysis.

In a term-based approach, sentiment terms or adjectives that possess an inherent sentiment, such as "good" or "bad," are identified and matched with the keywords that they are closely associated with [19]. For instance, if one were to encounter a review that states, "Thank God the color is good, but the delivery took a long time," this review would be assigned a positive sentiment for the keywords "color" and a negative sentiment for the keywords "delivery".

There are various methodologies through which sentiment analysis can be conducted, and employing a dictionary is the most straightforward approach. Sentiment analysis dictionaries provide valuable information regarding the emotions or polarities that are expressed by various words, phrases, or concepts [20]. In this research, a sentiment dictionary was compiled, comprising words with either a negative or positive sentiment, which would then facilitate the execution of term-based sentiment analysis.

The sentiment terms included in this dictionary were meticulously extracted and manually selected from the words featured in the customer review dataset utilized for this research. These sentiment terms were labeled 1 for words that exhibited positive sentiment and -1 for words that displayed negative sentiment. To conduct sentiment analysis based on terms, it is initially necessary to generate a dictionary of sentiment terms that frequently occur. Following this, the review itself is deconstructed into word fragments or phrases, considering the keywords present within the sentence of the review. The sentiment terms

that subsequently appear with the keywords in a given phrase are then utilized as a foundation for providing a sentiment value.

3 Result

The data that was processed as part of this research was obtained through web scraping, specifically by extracting information from the comments or review columns found within six local shoe product stores that have achieved the highest sales levels on the Shopee e-marketplace. It is important to note that the information gathered relates to user comments or evaluations concerning the products they bought between September 1, 2021, and September 1, 2022. Collecting this data occurred over the period above and resulted in 243,487 individual data points being obtained for analysis. These data points focus specifically on six local shoe brands, encompassing a variety of 534 different products. Furthermore, a compelling finding can be made regarding the typical number of words detected within a single review, totaling roughly 9.34. This information allows us to ascertain that the total number of words present within the collected reviews is a staggering 2,263,984 words. However, it is worth mentioning that out of the total 243,487 data points collected, a mere fraction of less than 50 %, specifically 110,107 data points, consisted of user reviews. Conversely, the remaining majority of users chose to rate the products solely without providing any written reviews, resulting in blank entries within the dataset.

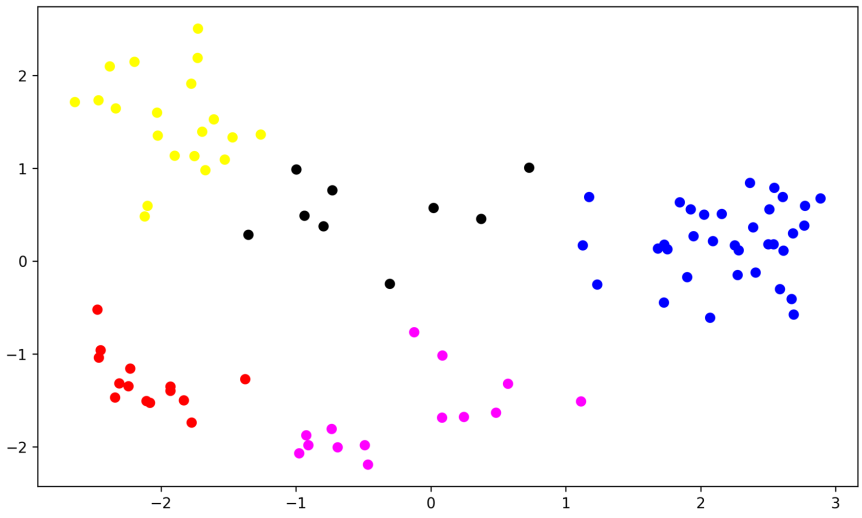

Figure 3: Cluster of keywords generated by k-means algorithm.

3.1 Extracted Keywords

After completing the procedures of data cleaning and POS tagging, the initial phase of keyword extraction commences, wherein a set of potential keywords are recognized as candidates. It is noteworthy to mention that, at this particular phase, approximately 477

keywords have been successfully identified as candidates. To provide a comprehensive overview of these extracted keywords, Figure 2 showcases a word cloud, visually encapsulating the essence of these linguistic entities.

However, it is essential to acknowledge that the abundance of keywords obtained during this stage remains quite substantial. Regrettably, some of them may not possess the desired level of relevance. This regrettable circumstance can primarily be attributed to the limited availability of POS Tagging models specifically developed for the Indonesian language, thereby impeding optimizing the keyword filtering process. Consequently, a rigorous filtering mechanism is meticulously implemented to acquire keywords that are genuinely pertinent to the keywords.

From the initial keyword extraction of approximately 477 candidate keywords, we performed a manual examination process to ascertain the suitability of each keyword for the subsequent clustering analysis. The manual examination aimed to evaluate the keywords' relevance in consumer reviews, ensuring they were representative of consumer sentiments and product attributes, and we analyzed the frequency of these keywords. Keywords that needed to be more sparse or relevant to consumer preferences were excluded to maintain the integrity of the topic clusters. Table 1 outlines the manually selected keywords from the candidate keywords extracted from our algorithm and their corresponding frequency of appearance throughout the consumer review corpus.

3.2 Keyword Clustering

The selected keywords from consumer reviews shown in Table 1 exhibit several distinct consumer preferences. Those keywords present a complex interplay of consumer focus areas. While each keyword carries a different meaning, their contextual use in consumer reviews often blurs the lines between categories, creating a challenge in determining their relationships and thematic connections.

Table 1: The manually selected keywords from extracted keywords candidate

Keyword	**Freq.**	**Keyword**	**Freq.**
pengiriman	10,123	*empuk*	1,828
warna	9,279	*ringan*	1,598
penjual	9,190	*jahitan*	1,474
harga	8,072	*tebal*	787
packaging	7,438	bubble	733
tampilan	7,175	*diskon*	716
nyaman	4,313	*promo*	581
kardus	4,294	*pelayanan*	579
respon	2,910	*benang*	237
murah	2,691	*kokoh*	213
paket	2,200	*lembut*	198
kurir	2,062	finishing	173
lem	1,948		

We then grouped those keywords into clusters to reveal the natural categories associated with those keywords for further analysis. This study employs the k-means clustering method to generate multiple data clusters. K-means is one of the clustering algorithms used to group data into clusters with similar characteristics. This process can help extract

topics or patterns that emerge in the data. After several attempts at data clustering using the k-means method, five data clusters were obtained, represented by the colors yellow, blue, black, magenta, and red, as seen in Figure 2. For further explanation, Table 2 outlines the keywords assigned to each cluster and the estimated representative topics of those keywords.

Table 2: The membership of each keyword clusters

Cluster	Representative Keywords	Estimated Topic
Yellow	*tampilan, warna, finishing, lem, benang, jahitan*	Product appearance
Blue	*lembut, empuk, ringan, kokoh, tebal, nyaman*	Product convenience
Black	*promo, murah, diskon, harga*	Price and promotion
Magenta	*paket, pengiriman, kurir, penjual, pelayanan, respon*	Service and delivery
Red	*bubble, kardus, packaging*	Packaging

The yellow cluster contains data related to product appearance, covering aspects such as product appearance, color, finishing, glue, thread, stitching, and *so on*. The blue cluster includes data related to product convenience, focusing on attributes like softness, thickness, and more. The black cluster pertains to price and promotions, encompassing promotions, discounts, and all matters concerning product pricing. The magenta cluster addresses service and product delivery data, covering topics like delivery packages, couriers, delivery times, and *so on*. The final cluster, marked in red, contains data related to packaging, including product packaging, packaging safety, and more.

3.3 Sentiment Analysis

Based on the clustering process in the previous section, sentiment analysis will be conducted to evaluate or assess the feelings, opinions, or sentiments contained in product reviews provided by customers. Determining a positive sentiment label requires a sentiment mapping score greater than 0.2, while for a negative sentiment label, the sentiment mapping score should be less than -0.2. Otherwise, a neutral label will be assigned. Sentiment analysis will be performed by categorizing the data into groups based on the earlier segmentation. Within the product appearance group, there are 56,262 with a positive sentiment, 5,965 with a negative sentiment, and 3,673 with a neutral sentiment. In the product convenience group, there are 89,911 with a positive sentiment, 536 with a negative sentiment, and 9,554 with a neutral sentiment. In the price promotion group, there are 57,418 with a positive sentiment, 5,602 with a negative sentiment, and 36,650 with a neutral sentiment. In the service and delivery group, there are 56,464 with a positive sentiment, 5,920 with a negative sentiment, and 3,821 with a neutral sentiment. In the last group, packaging, there are 55,584 with a positive sentiment, 6,051 with a negative sentiment, and 1,010 with a neutral sentiment. Figure 3 summarizes the sentiment analysis results and shows the positive, neutral, and negative sentiment distribution from each of the discovered clusters.

4 Discussion

Applying POS labeling and ensuing utilization of k-means clustering and word2vec similarity calculations yielded five separate clusters in our exploration. These clusters, specif-

ically named product appearance, product convenience, price and promotion, service and delivery, and packaging, offer valuable insights into the main topics of customer conversations about local fashion products in Indonesia. The identified clusters illuminate the significant aspects that customers highlight in their discussions.

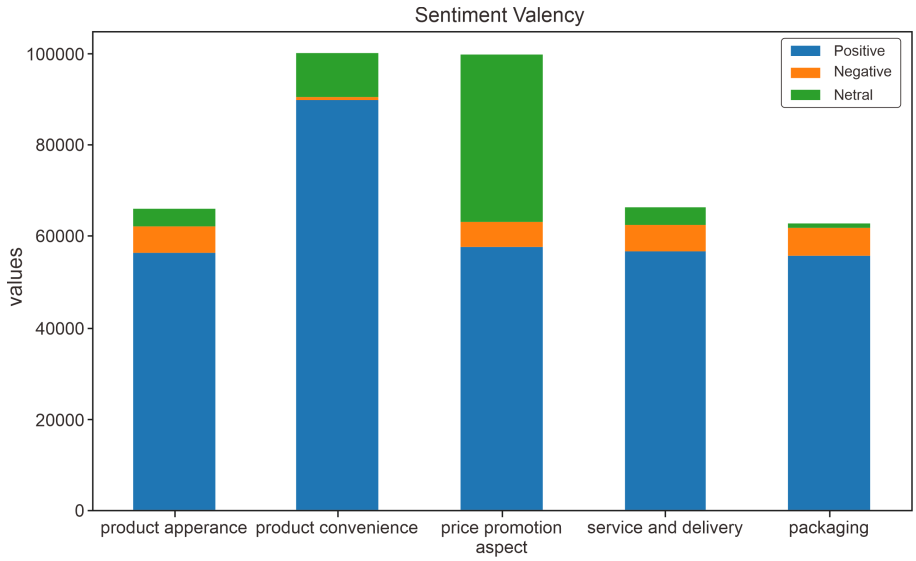

Figure 4: Sentiment analysis of each keyword cluster.

The first cluster, named 'product appearance', underscores the significance of visual aesthetics in customer perceptions. Moreover, the 'product convenience' cluster highlights the relevance of usability and practicality in the eyes of the consumer. The cluster centered around 'price and promotion' suggests that pricing strategies and promotional activities are pivotal in shaping customer sentiment. This discovery underlined the importance of companies carefully considering their pricing models and promotional campaigns to align with customer expectations and preferences.

The cluster centered around 'price and promotion' suggests that pricing strategies and promotional activities are pivotal in shaping customer sentiment. This finding emphasizes the need for Businesses should consider pricing models and promotional campaigns to align with customer expectations and preferences. The 'packaging' cluster highlights the influence of visual presentation and packaging on customer opinions. This finding underscores the need for businesses to invest in appealing and functional packaging designs to enhance the overall customer experience.

The positive sentiment predominantly observed across all clusters indicates the satisfaction expressed by customers in their discussions about local fashion products. This sentiment likely arises from an alignment between the product offerings and the expectations of the discerning consumer base. This synchronicity signifies a foundation for fostering customer loyalty and cultivating a positive perception of the local fashion product.

When comparing our study with previous works, such as those by Park *et al.* [2] and Lucini *et al.* [3], we notice a significant difference in application and context. Park and Lucini utilized clustering algorithms to group keywords in customer feedback, primarily

within the hospitality industry. Their studies laid a foundational framework for extracting consumer insights from textual data but have yet to extend to the fashion product context. The specificity and complexity of customer sentiments in the fashion sector, particularly in an Indonesian setting, should have been addressed in their research.

Similarly, Kiakatswin *et al.*'s [4] study, which employed LDA to uncover latent themes in consumer feedback, offered a more profound comprehension of product sentiments. However, this research did not explore the potential of clustering these extracted keywords for a more abstract and thematic analysis. In contrast, our study extracts and clusters keywords, allowing us to identify and analyze overarching themes within the customer reviews. This approach provides a more nuanced understanding of the various aspects influencing customer perceptions, such as design, quality, and overall experience with fashion products in online marketplaces.

5 Conclusion

Our study provides valuable insights into customer perceptions of local fashion products online. The identified clusters and their associated sentiments offer actionable information for businesses seeking to enhance their offerings and customer interactions. Businesses can strengthen their competitive edge and foster positive customer relationships by addressing the focal points highlighted in this study. This paper discusses product evaluation and customer perceptions in the marketplace. We have collected data from the marketplace, totaling 110,109 review entries. We performed a set of text processing approaches, including keyword extraction, keyword clustering, and sentiment analysis. From the extracted keywords, we employed k-means to cluster those keywords into certain groups, which represent more abstract concepts of user preference; we then employed term-based sentiment analysis to reveal the sentiment of each group.

From these findings, Indonesian society, especially marketplace users, has various opinions about products, ranging from appearance to pricing. On average, various product-related topics have a positive sentiment. This is evident in the analysis results graph, which shows that sentiment is positive for more than 70 % of the topics. However, businesses should still pay attention to the negative sentiments that exist. There is a need for product quality and service improvement to meet customer demands, which, in turn, will increase business profitability.

It is essential to acknowledge the limitations of this study. The analysis focused on a specific sample of local fashion products in Indonesia, which may represent different products from other regions or product categories. Furthermore, our study predominantly relied on textual data extracted from online reviews. While this method is effective for understanding customer sentiments expressed online, it might not capture the full spectrum of consumer behavior. Factors such as actual purchasing decisions, long-term brand loyalty, and post-purchase satisfaction, which are often better understood through direct customer interviews or longitudinal studies, were outside the scope of this research.

Future research could expand the scope to encompass a broader range of products and markets to provide a more comprehensive understanding of customer perceptions in online marketplaces. In advance, future research should consider incorporating a more diverse range of fashion products from various geographical regions to enhance the universality of the findings. Additionally, integrating different research methodologies, such as qualitative

data from consumer interviews or quantitative data from sales figures, could provide a more holistic view of consumer behavior in the fashion industry.

Acknowledgments

This research and publication were fully funded by The Ministry of Education, Culture, Research, and Technology through a research grant for capacity improvement (Penelitian Dosen Pemula/PDP) 2023.

References

[1] R. Y. Kim, "Using online reviews for customer sentiment analysis," *IEEE Engineering Management Review*, vol. 49, no. 4, pp. 162–168, 2021, doi: 10.1109/EMR.2021.3103835.

[2] E. O. Park, B. K. Chae, J. Kwon, and W. H. Kim, "The effects of green restaurant attributes on customer satisfaction using the structural topic model on online customer reviews," *Sustainability*, vol. 12, no. 7, p. 2843, 2020, doi: 10.3390/SU12072843.

[3] F. R. Lucini, L. M. Tonetto, F. S. Fogliatto, and M. J. Anzanello, "Text mining approach to explore dimensions of airline customer satisfaction using online customer reviews," *J Air Transp Manag*, vol. 83, p. 101760, 2020, doi: 10.1016/J.JAIRTRAMAN.2019.101760.

[4] K. Kiatkawsin, I. Sutherland, and J. Y. Kim, "A comparative automated text analysis of Airbnb reviews in Hong Kong and Singapore using latent dirichlet allocation," *Sustainability*, vol. 12, no. 16, p. 6673, 2020, doi: 10.3390/SU12166673.

[5] R. Shanmugam, "Practical text analytics: Maximizing the value of text data," *J Stat Comput Simul*, vol. 90, no. 7, pp. 1346–1346, 2020, doi: 10.1080/00949655.2019.1628899.

[6] A. Chiche and B. Yitagesu, "Part of speech tagging: a systematic review of deep learning and machine learning approaches," *J Big Data*, vol. 9, no. 1, pp. 1–25, 2022, doi: 10.1186/S40537-022-00561-Y/FIGURES/5.

[7] K. Widhiyantil and A. Harjoko, "POS tagging Bahasa Indonesia dengan HMM dan rule based," *Jurnal Informatika*, vol. 8, no. 2, pp. 151-167, 2012.

[8] T. Almutiri and F. Nadeem, "Markov models applications in natural language processing: A survey," *I.J. Information Technology and Computer Science*, vol. 2, pp. 1–16, 2022, doi: 10.5815/ijitcs.2022.02.01.

[9] T. Xia and X. Chen, "A weighted feature enhanced hidden Markov model for spam SMS filtering," *Neurocomputing*, vol. 444, pp. 48–58, 2021, doi: 10.1016/J.NEUCOM.2021.02.075.

[10] K. W. Church, "Word2Vec," *Nat Lang Eng*, vol. 23, no. 1, pp. 155–162, 2017, doi: 10.1017/S1351324916000334.

[11] A. Sharma and S. Kumar, "Ontology-based semantic retrieval of documents using Word2vec model," *Data Knowl Eng*, vol. 144, p. 102110, 2023, doi: 10.1016/J.DATAK.2022.102110.

[12] S. Ruder, I. Vulić, and A. Søgaard, "A survey of cross-lingual word embedding models," *Journal of Artificial Intelligence Research*, vol. 65, pp. 569–631, 2019, doi 10.1613/JAIR.1.11640.

[13] S. J. Johnson, M. R. Murty, and I. Navakanth, "A detailed review on word embedding techniques with emphasis on word2vec," *Multimed Tools Appl*, pp. 1–29, 2023, doi: 10.1007/S11042-023-17007-Z/METRICS.

[14] M. Ahmed, R. Seraj, and S. M. S. Islam, "The k-means algorithm: A comprehensive survey and performance evaluation," *Electronics*, vol. 9, no. 8, p. 1295, 2020, doi: 10.3390/ELECTRONICS9081295.

[15] A. K. Abasi, A. T. Khader, M. A. Al-Betar, S. Naim, Z. A. A. Alyasseri, and S. N. Makhadmeh, "A novel hybrid multi-verse optimizer with k-means for text documents clustering," *Neural Comput Appl*, vol. 32, no. 23, pp. 17703–17729, 2020, doi: 10.1007/S00521-020-04945-0/METRICS.

[16] Y. Rong and Y. Liu, "Staged text clustering algorithm based on k-means and hierarchical agglomeration clustering," in *Proceedings of 2020 IEEE International Conference on Artificial Intelligence and Computer Applications, ICAICA 2020*, 27-29 June 2020, Dalian, China, pp. 124–127, doi: 10.1109/ICAICA50127.2020.9182394.

[17] P. Nandwani and R. Verma, "A review on sentiment analysis and emotion detection from the text," *Soc Netw Anal Min*, vol. 11, no. 1, pp. 1–19, 2021, doi: 10.1007/S13278-021-00776-6/TABLES/1.

[18] D. Zimbra, A. Abbasi, D. Zeng, and H. Chen, "The state-of-the-art in Twitter sentiment analysis: A review and benchmark evaluation," *ACM Trans Manag Inf Syst*, vol. 9, no. 2, Apr. 2018, doi: 10.1145/3185045.

[19] N. Kumaresh, J. Naulegari, V. Bonta, and N. Janardhan, "A comprehensive study on lexicon based approaches for sentiment analysis," *Asian Journal of Computer Science and Technology*, vol. 8, no. 2, pp. 1–6, 2019, doi 10.51983/adjust-2019.8.S2.2037.

[20] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Computational Linguistics*, vol. 37, no. 2, pp. 267–307, 2011, doi 10.1162/COLI_A_00049.