ECTI Transactions on Computer and Information Technology
Journal homepage: https://ph01.tci-thaijo.org/index.php/ecticit/
Published by the ECTI Association, Thailand, ISSN: 2286-9131

Optimizing Product Matching in E-Commerce with DOC2VEC: Leveraging Hierarchical Clustering Parameters Based on Product Titles

Yuliana Melita Pranoto[1], Anik Nur Handayani[2], Heru Wahyu Herwanto[3] and Yosi Kristian[4]

ABSTRACT

Information technology is pivotal in increasing efficiency and effectiveness in online retail, particularly in product matching. This research delves into the challenges associated with product matching in the e-commerce sector, addressing issues related to the diversity and ambiguity of product titles and the fast-paced introduction of new products to the market. As a solution, we implement a neural network-based approach. The main contribution of this research is the implementation and validation of the Doc2Vec method in the context of product matching for e-commerce products. Additionally, this study successfully identifies the optimal parameter combinations for Hierarchical Clustering, which has been tested and validated on 4,000 product title data points. The data for learning and evaluation comes from an online retail platform and includes 34,000 product names from various sectors. The research compares two Doc2Vec architectures for feature extraction from product titles and then integrates them with a Hierarchical Clustering approach to group similar products. The results indicate that the Doc2Vec model with the DBOW (Distributed Bag of Words) architecture yields a better average NMI (Normalized Mutual Information) Score than the DM (Distributed Memory) architecture.

Article information:

Keywords: Product Matching, Doc2Vec, Hierarchical Clustering, NMI Score, Neural Network

Article history:
Received: March 28, 2024
Revised: May 23, 2024
Accepted: July 18, 2024
Published: August 3, 2024
(Online)

DOI: 10.37936/ecti-cit.2024183.256164

1. INTRODUCTION

Information technology is crucial for the efficient operation of online businesses, especially in the e-commerce sector [1]. It facilitates the selling process and enhances the user experience for buyers. The importance of online shopping has been steadily increasing, offering a convenient and efficient method for consumers to make purchases. Despite these advantages, e-commerce also presents unique challenges that demand innovative solutions. Numerous studies have explored the ever-changing nature of online shopping, its impact on consumer behavior, and the emerging trends shaping the industry [2,3,4].

The growth of the online shopping industry is introducing new challenges. One of these challenges is that online buying has become more complicated for customers. They have so many online stores to choose from that making a choice can be challenging. Such a situation necessitates a product matching scheme to facilitate easy product searching [5] and to find the most suitable online stores. A method should be developed to match products based on customer descriptions, thereby improving the shopping experience. Product matching in online stores can be conducted more efficiently by adopting semantic markup language to interpret products [6].

In the online retail sector, product matching is a distinct form of record linkage that aims to identify and associate items from one situation to another [7]. Earlier studies on this topic mainly used rule-based and statistical methods. However, machine learning has become popular lately because it works well [8]–[11]. Furthermore, deep learning methods have been very successful in many areas [12] for solving the product matching problem.

The issue of matching products in online shopping

[1,2,3] The authors are with the Department of Electrical Engineering and Informatics, Universitas Negeri Malang, Indonesia, E-mail: yuliana.melita.2205349@students.um.ac.id, aniknur.ft@um.ac.id and heru_wh@um.ac.id

[1,4] The authors are with the Department of Informatics, Institut Sains dan Teknologi Terpadu Surabaya, Indonesia, E-mail: yuliana.melita.2205349@students.um.ac.id and yosi@stts.edu

[2] Corresponding author: aniknur.ft@um.ac.id

has been studied extensively, and researchers are interested in it. Earlier studies used specific algorithms that did not need supervision to match products just by their titles [13]. More recent methods focus on machine learning techniques; some even use advanced deep learning to match products based on their unstructured descriptions [14].

Some product matching techniques are designed to find similar products across different brands, while others aim to locate identical products from the same brand but sold in various online stores. Many algorithms have been proposed in studies that focus on finding text duplicates. Other research suggests using K-Combinations and Permutations to group different e-commerce product titles that refer to the same item. This method uses how close the words are and how often they appear in the brand, model, and product attributes as factors [15].

Neural network-based approaches are used to tackle the issue of matching products by their titles on e-commerce platforms [16]. This study compares the performance of these models with traditional methods like Support Vector Machines (SVM) and Random Forests. The results show that neural network models perform better in accuracy and can be trained very efficiently [15,16]. Meanwhile, [7] fine-tuned the network models to make them more effective at matching products. Using pre-trained networks has been proven to improve accuracy when matching products, compared to other traditional methods that do not use machine learning to adapt [19]. These networks can also handle challenges like varying product names, incomplete product information, and differences in how product descriptions are formatted across various e-commerce platforms.

New product recommendation systems for customers are developed based on text similarity with products they've bought before [20]. Machine learning models are trained using word2vec, a natural language processing technique, to get vector representations of product descriptions and reviews. These vector representations are then used to calculate how similar two products are; the higher the value, the more alike the products.

The Doc2Vec method, an extension of the word2vec model, offers an effective solution for generating numerical vectors from text documents [21]. Concurrently, clustering methods have become robust techniques for grouping similar text data. Hierarchical clustering is frequently employed due to its capacity to furnish a hierarchical structure for the data [22]. However, the combination of these two methods for analyzing product title text has not been extensively explored despite the critical role that product titles often play in purchase decisions.

The beneficial contributions of this research are:

1. Creating vector representations of text in title features using the Doc2Vec method.

Table 1: *Examples of Product.*

Product Title		Product Image	Group
Original	Translated to English (for documentation purposes only)		
Sarung celana wadimor original 100% dewasa dan anak hitam dan putih polos	Wadimor Original 100% Adult and Children Plain Black and White Sarong Pants		1
SARUNG CELANA WADIMOR DEWASA HITAM POLOS SARCEL	WADIMOR ADULT PLAIN BLACK SARONG PANTS		1
WARNA RANDOM ACAK Sarung Celana Wadimor MURAH Celana Sarung WADIMOR	RANDOM COLOR MIX Cheap Wadimor Sarong Pants		1
GROSIR LVN COLLAGEN / COLAGEN STROBERI PEMUTIH KULIT 1 BOX ISI 10 SACHET	WHOLESALE LVN COLLAGEN / STRAWBERRY COLLAGEN SKIN WHITENER 1 BOX CONTAINS 10 SACHETS		2
DISTRIBUTOR LVN COLLAGEN STROBERI / COLAGEN 1 BOX (10 sachet)	DISTRIBUTOR OF LVN STRAWBERRY COLLAGEN / COLAGEN 1 BOX (10 sachets)		2
TERMURAH LVN COLLAGEN STROBERI 1 BOX 10 SACHET	CHEAPEST LVN STRAWBERRY COLLAGEN 1 BOX 10 SACHETS		2
LVN COLLAGEN - ORIGINAL TERMURAH - LVN STROBERI - GARANSI UANG KEMBALI	LVN COLLAGEN - CHEAPEST ORIGINAL - LVN STRAWBERRY - MONEY BACK GUARANTEE		2
I7s TWS Apple Airpods Android Nirkabel Bluetooth Headset Earphone Hitam Ready Stock	I7s TWS Apple Airpods Android Wireless Bluetooth Headset Earphone Black In Stock		3
I7 i7s TWS Earphone Mini Bluetooth 4.2 dengan Mic	I7 i7s TWS Mini Bluetooth 4.2 Earphone with Mic		3
i7s Wireless Bluetooth Earphone Stereo Earbuds With Charging Box	i7s Wireless Bluetooth Stereo Earbuds with Charging Box		3

2. Determining the most appropriate parameter criteria from a subset of product data to group those products into relevant clusters.
3. Employing Hierarchical Clustering techniques to group many products into existing clusters based on previously obtained parameter criteria.

2. MATERIALS AND METHODS

This section outlines the methods and approaches in this study that tackle the challenge of product matching in e-commerce. Given the complexity and vagueness often present in product titles, this research incorporates two critical methods for text processing and data evaluation: Doc2Vec and Hierarchical Clustering. Doc2Vec is responsible for extracting features from product titles, thereby transforming textual information into vector representations suitable for further analysis. On the other hand, Hierarchical Clustering groups similar products based on these vector representations. Additional tests are conducted to assess the efficacy of various metric and linkage parameters in Hierarchical Clustering. The dataset for this study originates from the Shopee e-commerce platform and comprises 34,000 product titles across diverse categories.

2.1 Dataset

In this study, we work with a dataset sourced from Kaggle. Kaggle serves as an online community designed to facilitate idea sharing, learning, and competition among its members in Data Science and Machine Learning. It holds a significant role in the fast-evolving domain of data science.

Shopee is one of the largest e-commerce platforms in Indonesia. According to a study by Ipsos concerning e-commerce competition in Indonesia at the end of 2021, Shopee was the most widely used platform in the country [23]. This dataset [24] is well-suited for tasks such as image classification, product matching, and product recommendations. Examples of products included in the dataset are displayed in Table 1.

In this research, we use 30,000 product titles to train the Doc2Vec model. This trained model is employed in several experiments involving an additional 4,000 product title texts. One such experiment is the application of Hierarchical Clustering to group product titles based on their similarity.

2.2 Proposed Methods

This study's dataset is divided into two parts: 88% is allocated for the product data (containing 30,000 product titles), and 12% is allocated for the test data (comprising 4,000 product titles). The product data serves as the foundation for training a model using Doc2Vec, while the test data is employed for making predictions based on the developed model. The results of these predictions, represented as text features, are then applied to the Hierarchical Clustering method to identify relevant product groups.

This model is constructed by leveraging the Doc2Vec algorithm and fine-tuning various parameters. Meanwhile, data clustering is executed through the Hierarchical Clustering technique, analyzing the thresholds of multiple parameters. This study aims to understand how the vector representations from Doc2Vec can impact the quality of the clusters generated by Hierarchical Clustering. The proposed method is illustrated in Fig. 1.

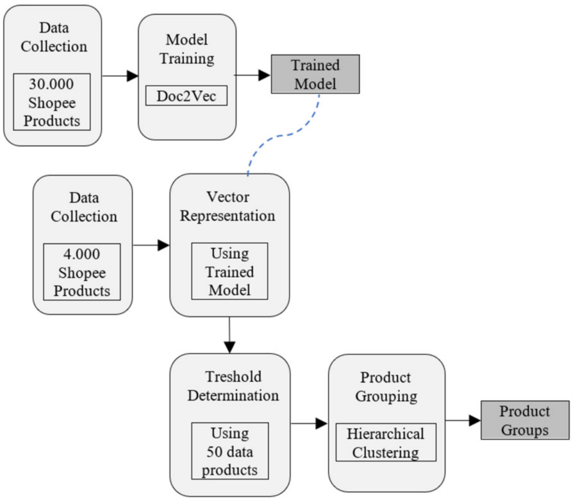


Fig.1: *The Proposed Method.*

2.3 Doc2Vec

Doc2Vec is a learning algorithm that converts text documents into numerical vector representations suitable for classification, clustering, and recommendation generation. It is an extension of the Word2Vec word embedding method, encoding the entire document as a vector rather than just individual words. Doc2Vec has produced superior features for automatic product classification compared to Word2Vec [21].

$$L = \sum\nolimits_{w\in D} \sum\nolimits_{c\in C(w)} \log p(c|w,D) \qquad (1)$$

Equation 1 represents the mathematical formulation for the Doc2Vec model, aiming to maximize the likelihood function based on the context C for each target word w in document D. Where L is the likelihood function to be maximized, w is the target word, D is the document id or tag, $C(w)$ are the context words of w.

Doc2Vec can address challenges like data limitations and language variations commonly found in product descriptions. It can generate vectors that effectively and efficiently represent the meaning and context of the documents [23,24]. [27] demonstrated that Doc2Vec performs well when using a model trained on a large external corpus and can be further enhanced using pre-trained word embeddings. Doc2Vec can produce high-quality product embeddings that align well with existing product clusters [28].

There are two types of architectures in Doc2Vec: Distributed Memory (DM) and Distributed Bag of Words (DBOW). DM utilizes both document vectors and word vectors as inputs to predict the subsequent word in the context of a document. DM captures the sequence of words within the document and integrates information from the document and the words. On the other hand, DBOW employs only the document vector as input to predict words appearing in the document. DBOW disregards the sequence of words within the document and focuses exclusively on the information derived from it [29]. In this research, we aim to compare these two Doc2Vec architectures: DM and DBOW.

The DM architecture leverages document and word vectors as inputs to project the next word in a given document context. DM excels in retaining the word sequence within the document and combining contextual information from the words and the document itself. Conversely, DBOW requires only the document vector as input to predict the presence of words in a document. Unlike DM, DBOW does not consider the word sequence, and its primary focus is on the global information extracted from the document [30].

Table 2: *Exploration of Initial Parameters in Doc2Vec Model Configuration.*

Parameters	Value
Vector size	1000
Alpha / Learning rate	0.025
Epoch	100
Architecture	DM and DBOW

In this research, Doc2Vec handles feature extraction by engaging the words within the documents as part of the learning process. Two distinct Doc2Vec models are constructed, one using the DBOW architecture and the other using the DM architecture. While DBOW concentrates on learning document vectors, DM simultaneously learns document and word vectors. Notably, model construction with DM takes more time than DBOW. Subsequently, we conduct a comparative analysis between the two architectures (DBOW and DM) to determine which is more effective for the objectives of this study.

2.4 Hierarchical Clustering

Hierarchical clustering is a machine-learning method that helps cluster data with similar characteristics. This method represents the data in a tree-like structure known as a dendrogram, where each node signifies a combination of multiple leaf nodes. The dendrogram displays the levels of closeness among the nodes within the network [22]. There are two types of Hierarchical Clustering: Agglomerative Clustering and Divisive Clustering. In this research, we employ the Agglomerative Clustering approach, where each data point is considered a separate cluster and progressively merged into larger clusters until a primary cluster is formed [29,30].

One of the advantages of Hierarchical Clustering is the dendrogram, which provides visual information about the relationships among data and the flexibility in choosing the desired number of clusters. However, the limitations of this method depend on the metrics and linkage methods employed [33],32]. We explore all combinations of distance metrics and linkage methods, including Euclidean Distance, Manhattan Distance, Cosine Similarity, Ward Linkage, Single Linkage, Complete Linkage, and Average Linkage [35]. Specifically, Ward Linkage can only be paired with Euclidean Distance; thus, the number of parameter combinations is 10 pairs.

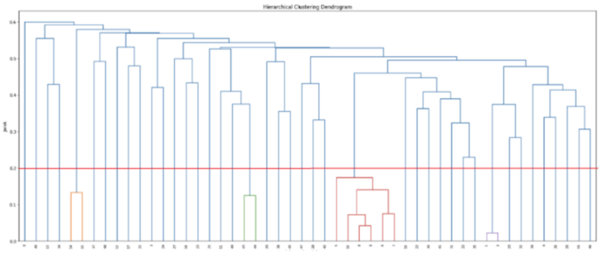


Fig.2: *Dendogram with DBOW Architecture, Metric Cosine Similarity, Linkage Average, Threshold = 0.2.*

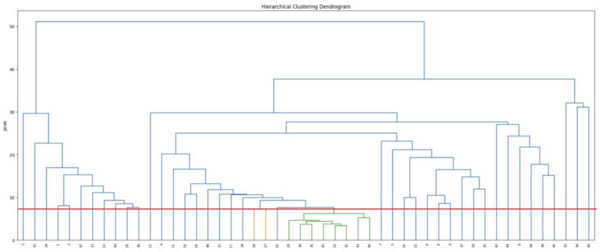


Fig.3: *Dendogram with DM Architecture, Metric Manhattan Distance, Linkage Complete, Threshold = 7.5.*

This study uses a subset of data to identify the most appropriate parameter settings for clustering products into suitable groups. We randomly select 50 product titles from a dataset of 4,000 product titles. Using the previously developed Doc2Vec model, we extract textual features from these 50 titles and visualize them in a dendrogram through agglomerative clustering. We establish a threshold based on the actual number of clusters, which corresponds to the absolute number of groups determined by observing these randomly selected product titles with the aid of a dendrogram. The threshold results for each parameter combination in hierarchical clustering are presented in Table 3.

3. RESULTS AND DISCUSSION

Testing was conducted on 4,000 product titles using previously constructed Doc2Vec models with DM and DBOW architectures. The products were then grouped into the nearest clusters using hierarchical clustering based on the parameters specified in Table 3. The testing results were evaluated using the NMI (Normalized Mutual Information) score and are presented in Table 4.

Table 3: *Parameter Criteria in Hierarchical Clustering.*

Doc2Vec Architecture	Metric	Linkage	Best Threshold
Distributed Bag of Words (DBOW)	Euclidean Distance	Ward	0.557
		Single	0.45
		Complete	0.555
		Average	0.54
	Manhattan Distance	Single	12.21
		Complete	14.2
		Average	13.5
	Cosine Similarity	Single	0.2
		Complete	0.227
		Average	0.2
Distributed Memory (DM)	Euclidean Distance	Ward	0.29
		Single	0.22
		Complete	0.29
		Average	0.27
	Manhattan Distance	Single	6
		Complete	7.5
		Average	6.8
	Cosine Similarity	Single	0.18
		Complete	0.227
		Average	0.215

The NMI score is employed to determine the similarity between two clusters, ranging from 0 to 1. A score of 0 indicates that the two clusters are entirely distinct, while a score of 1 signifies that the two clusters are identical. The NMI score is subsequently used to label clusters considered to be equivalent.

This research calculates the NMI score from 0 to 1 to evaluate the clustering outcomes obtained from various parameter settings. A higher NMI score indicates better clustering quality produced by the algorithm with a specific parameter configuration. The NMI score measures how effectively the clustering algorithm has grouped similar data points. A higher NMI value suggests that the algorithm has successfully identified the underlying patterns in the data and that the resulting clusters are more accurate [36][37]. Compared to other methods, NMI provides superior results in the issues of cluster comparison and labeling [38].

We observed that the time required to construct a model using the DM architecture is longer than that needed for the DBOW architecture. While the longer training duration may be a factor, the insights gained from the DM model remain valuable for product clustering.

From various combinations of metric and linkage parameters depicted in the dendrogram for agglomerative clustering, we identified the best test results for both the DBOW and DM models.

Based on the experimental results in Table 4, the average NMI Score for the Doc2Vec model using the DBOW architecture outperforms that of the DM architecture. The average NMI Score stands at 0.91 for DBOW and 0.89 for DM. This performance difference can be attributed to the nature of product titles, which are generally short and information-rich. DBOW is more effective at processing concise sentences and prioritizes the presence of individual words over the sequence in which they appear.

Table 4: *Test results using NMI score.*

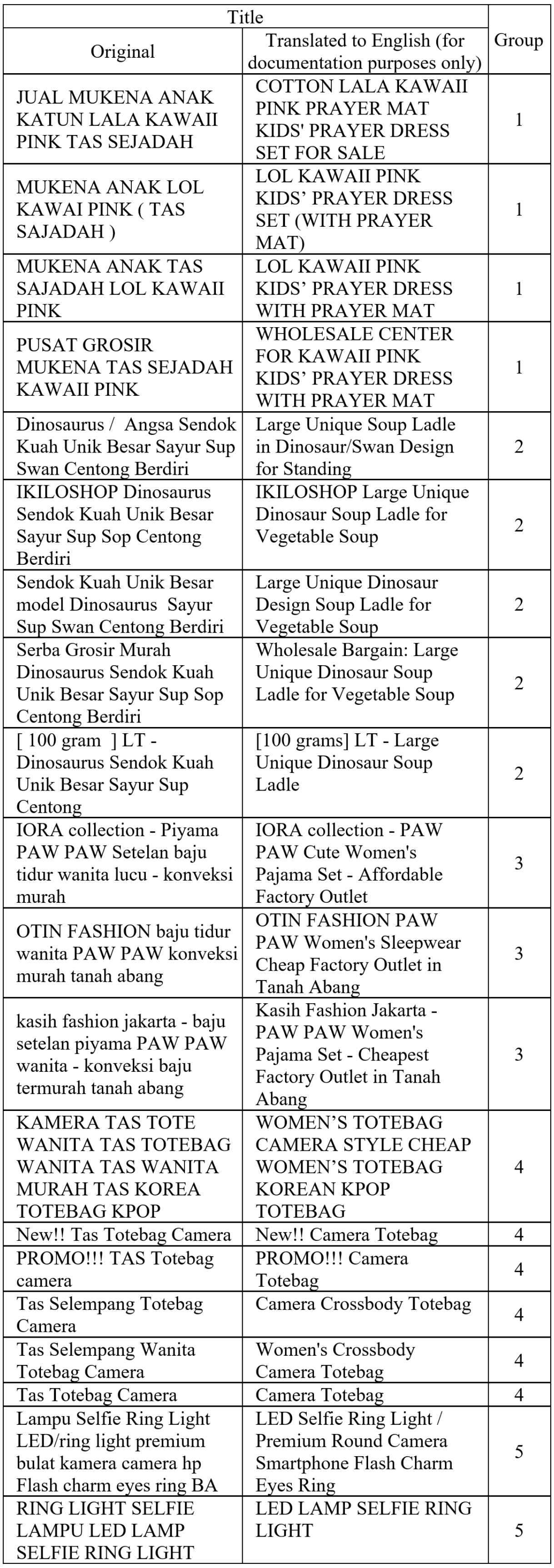
Doc2Vec Architecture	Metric	Linkage	NMI Score
Distributed Bag of Words (DBOW)	Euclidean Distance	Ward	0.96
		Single	0.86
		Complete	0.95
		Average	0.93
	Manhattan Distance	Single	0.81
		Complete	0.95
		Average	0.94
	Cosine Similarity	Single	0.81
		Complete	0.96
		Average	0.96
Distributed Memory (DM)	Euclidean Distance	Ward	0.93
		Single	0.86
		Complete	0.92
		Average	0.88
	Manhattan Distance	Single	0.84
		Complete	0.91
		Average	0.88
	Cosine Similarity	Single	0.84
		Complete	0.93
		Average	0.89

Table 5: *Combination of Vector Size and Epochs in the Doc2Vec Model.*

Parameters in the Doc2Vec Model		Parameters in the Hierarchical Clustering		Threshold	NMI score
Vector Size	Epoch	Metric	Linkage		
100	50	Euclidean Distance	Ward	1.1	0.965
500	50			1.1	0.966
1000	100			0.557	0.960
1500	150			0.252	0,924
100	50	Cosine Similarity	Complete	0.3	0.958
500	50			0.3	0.960
1000	100			0.227	0.959
1500	150			0.0077	0,926

The DBOW model achieved the highest NMI score of 0.96 using the parameters of Euclidean Distance and Ward linkage, Cosine Similarity, Complete linkage, and Cosine Similarity and Average linkage. Meanwhile, the DM model reached its best NMI score of 0.93 with Euclidean Distance and Ward linkage, Cosine Similarity, and Complete linkage. These findings indicate that the optimal parameters for metric and linkage are Euclidean Distance and Ward, along

Table 6: *Example of Product Titles Successfully Identified in the Same Group.*



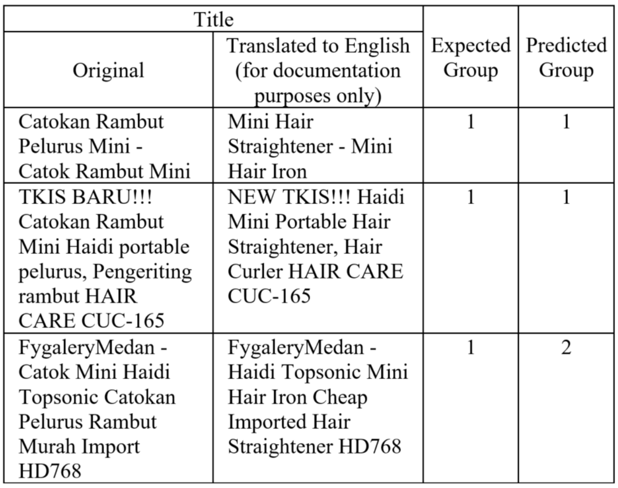
Title		Group
Original	Translated to English (for documentation purposes only)	
JUAL MUKENA ANAK KATUN LALA KAWAII PINK TAS SEJADAH	COTTON LALA KAWAII PINK PRAYER MAT KIDS' PRAYER DRESS SET FOR SALE	1
MUKENA ANAK LOL KAWAI PINK (TAS SAJADAH)	LOL KAWAII PINK KIDS' PRAYER DRESS SET (WITH PRAYER MAT)	1
MUKENA ANAK TAS SAJADAH LOL KAWAII PINK	LOL KAWAII PINK KIDS' PRAYER DRESS WITH PRAYER MAT	1
PUSAT GROSIR MUKENA TAS SEJADAH KAWAII PINK	WHOLESALE CENTER FOR KAWAII PINK KIDS' PRAYER DRESS WITH PRAYER MAT	1
Dinosaurus / Angsa Sendok Kuah Unik Besar Sayur Sup Swan Centong Berdiri	Large Unique Soup Ladle in Dinosaur/Swan Design for Standing	2
IKILOSHOP Dinosaurus Sendok Kuah Unik Besar Sayur Sup Sop Centong Berdiri	IKILOSHOP Large Unique Dinosaur Soup Ladle for Vegetable Soup	2
Sendok Kuah Unik Besar model Dinosaurus Sayur Sup Swan Centong Berdiri	Large Unique Dinosaur Design Soup Ladle for Vegetable Soup	2
Serba Grosir Murah Dinosaurus Sendok Kuah Unik Besar Sayur Sup Sop Centong Berdiri	Wholesale Bargain: Large Unique Dinosaur Soup Ladle for Vegetable Soup	2
[100 gram] LT - Dinosaurus Sendok Kuah Unik Besar Sayur Sup Centong	[100 grams] LT - Large Unique Dinosaur Soup Ladle	2
IORA collection - Piyama PAW PAW Setelan baju tidur wanita lucu - konveksi murah	IORA collection - PAW PAW Cute Women's Pajama Set - Affordable Factory Outlet	3
OTIN FASHION baju tidur wanita PAW PAW konveksi murah tanah abang	OTIN FASHION PAW PAW Women's Sleepwear Cheap Factory Outlet in Tanah Abang	3
kasih fashion jakarta - baju setelan piyama PAW PAW wanita - konveksi baju termurah tanah abang	Kasih Fashion Jakarta - PAW PAW Women's Pajama Set - Cheapest Factory Outlet in Tanah Abang	3
KAMERA TAS TOTE WANITA TAS TOTEBAG WANITA TAS WANITA MURAH TAS KOREA TOTEBAG KPOP	WOMEN'S TOTEBAG CAMERA STYLE CHEAP WOMEN'S TOTEBAG KOREAN KPOP TOTEBAG	4
New!! Tas Totebag Camera	New!! Camera Totebag	4
PROMO!!! TAS Totebag camera	PROMO!!! Camera Totebag	4
Tas Selempang Totebag Camera	Camera Crossbody Totebag	4
Tas Selempang Wanita Totebag Camera	Women's Crossbody Camera Totebag	4
Tas Totebag Camera	Camera Totebag	4
Lampu Selfie Ring Light LED/ring light premium bulat kamera camera hp Flash charm eyes ring BA	LED Selfie Ring Light / Premium Round Camera Smartphone Flash Charm Eyes Ring	5
RING LIGHT SELFIE LAMPU LED LAMP SELFIE RING LIGHT	LED LAMP SELFIE RING LIGHT	5

with Cosine Similarity and Complete Linkage.

Subsequently, we experimented to adjust the vector size and the number of epochs in creating a Doc2Vec model utilizing the DBOW architecture. The vector sizes used were 100, 500, 1000, and 1500, with epochs set at 50, 100, and 150. After successfully establishing the Doc2Vec model, we conducted further testing on 4,000 product titles. The threshold values were recalculated using a small data subset aided by the dendrogram representation of agglomerative clustering. We selected the metric and linkage parameters from Table 4, which had the best values. The NMI scores are presented in Table 5.

Table 5 indicates that a vector size of 500 and 50 epochs constitute the best combination, yielding the highest NMI score. Interestingly, increasing the vector size and the number of epochs does not always correlate with improving the NMI score; instead, it results in a longer computational time.

Table 7: *Example of Product Titles Failed to Cluster into a Single Group (1).*



Title		Expected Group	Predicted Group
Original	Translated to English (for documentation purposes only)		
Catokan Rambut Pelurus Mini - Catok Rambut Mini	Mini Hair Straightener - Mini Hair Iron	1	1
TKIS BARU!!! Catokan Rambut Mini Haidi portable pelurus, Pengeriting rambut HAIR CARE CUC-165	NEW TKIS!!! Haidi Mini Portable Hair Straightener, Hair Curler HAIR CARE CUC-165	1	1
FygaleryMedan - Catok Mini Haidi Topsonic Catokan Pelurus Rambut Murah Import HD768	FygaleryMedan - Haidi Topsonic Mini Hair Iron Cheap Imported Hair Straightener HD768	1	2

The results of clustering the 4,000 product titles into their nearest groups using hierarchical clustering are presented in Table 6. In that table, we showcase examples from 8 groups successfully identified in their respective clusters. From this, it is evident that Doc2Vec can recognize text with a reasonably high level of accuracy, thereby simplifying the task of categorizing product titles into relevant groups.

Further, in our examination of 4,000 product titles, we observed some title texts that could not be grouped into a single cluster. This observation was carried out on one of the highest NMI score results, specifically the DBOW architecture with the Cosine Similarity metric and Complete Linkage. The findings are presented in Tables 7 and 8.

Table 7 displays three title texts that should ideally be in the same group; however, the test results indicate that the third title is separated into a different group. Meanwhile, Table 8 shows four title texts that should ideally be divided into two groups, but

Table 8: *Example of Product Titles Failed to Cluster into a Single Group (2).*

Title		Expected Group	Predicted Group
Original	Translated to English (for documentation purposes only)		
(BISA COD) TERMURAH CELANA JEANS PRIA	(CASH ON DELIVERY AVAILABLE) CHEAPEST MEN'S JEANS	1	1
BISA COD BOXSER PRIA & WANITA	CASH ON DELIVERY AVAILABLE FOR MEN'S & WOMEN'S BOXERS	1	2
Celana JEANS LEVIS pria skinny STRETCH PENSIL / Celana Jeans Panjang Pria	Men's LEVIS Skinny Stretch Pencil Jeans / Men's Long Jeans	2	3
Celana Jeans Panjang Pria Skinny / CELANA JEANS LEVIS PENSIL PRIA STRECTH/MELAR	Men's Long Skinny Jeans / MEN'S LEVIS PENCIL JEANS STRETCH/EXPANDABLE	2	1

the test results yield three distinct groups.

We conducted experiments using several small datasets from a test dataset comprising 4,000 product titles to determine the most appropriate parameter criteria for clustering products into suitable groups. Each small dataset, sampled thrice, contained 50 product titles randomly selected from the test dataset. The threshold obtained for each subset varied depending on the random sampling results. The NMI scores are displayed in Table 9. From these experiments, we calculated the average threshold for the three subsets to be 0.51 (for the Euclidean Distance and Ward Linkage metrics) and 0.20 (for the Cosine Similarity and Complete Linkage metrics).

In addition, we added several datasets from other e-commerce platforms to our experiments (beyond the original 4,000 product titles test set). Two hundred-one product titles from two e-commerce platforms, Tokopedia and Blibli, were analyzed for their NMI scores. We then compared the actual number of clusters with the predicted number. The parameters used here were the Cosine Similarity metric and Complete Linkage, utilizing an average threshold of 0.20 from Table 9. The details of the experimental results are displayed in Table 10, with an average NMI score for the 201 product titles being 0.95.

Table 9: *Experiments with Multiple Small Data Subsets.*

Small Subsets of Data	Metric	Linkage	Threshold	NMI Score
First Data Subset	Euclidean Distance	Ward	0.557	0.96
	Cosine Similarity	Complete	0.227	0.96
Second Data Subset	Euclidean Distance	Ward	0.48	0.97
	Cosine Similarity	Complete	0.25	0.95
Third Data Subset	Euclidean Distance	Ward	0.5	0.97
	Cosine Similarity	Complete	0.115	0.98

Table 10: *Experiment with Adding Data from Other E-Commerce Platforms.*

E-Commerce Platform	Number of Products	NMI Score	Number of Clusters	
			Real	Predict
Tokopedia	113	0.954	17	24
Blibli	88	0.955	14	19

4. CONCLUSION

After a thorough series of experiments and analyses, we have discerned several essential points that can be considered the conclusions of this study. Our test results indicate that the DBOW model yields a higher average NMI score than the DM model. The NMI score reached 0.91 for the DBOW model and 0.89 for the DM model. These results suggest that the vector representations obtained from the DBOW model are better suited for clustering products based on their titles.

Additionally, we found that utilizing a small subset of randomly selected test data, even if it only includes 50 product titles, can offer invaluable guidance for determining the optimal parameters for clustering a more extensive dataset of up to 4,000 product titles. The results confirm that these new products can be automatically classified into existing clusters or allocated to new ones if they do not fit well. While the study provides valuable insights, it is not without its limitations. Future work should focus on expanding the product dataset and introducing more product diversity.

Moreover, enriching the feature set to include elements like product images could add depth to the research. These refinements are crucial for strengthening the study's relevance and accuracy.

AUTHOR CONTRIBUTIONS

Conceptualization, Yuliana Melita Pranoto, Anik Nur Handayani, Heru Wahyu Herwanto, and Yosi Kristian; methodology, Yuliana Melita Pranoto; software, Yuliana Melita Pranoto; validation, Yuliana Melita Pranoto, Anik Nur Handayani, Heru Wahyu Herwanto, and Yosi Kristian; formal analysis, Yuliana Melita Pranoto; investigation, Yuliana Melita Pranoto; data curation, Yuliana Melita Pranoto; writing—original draft preparation, Yuliana Melita Pranoto; writing—review and editing, Yuliana Melita Pranoto, Anik Nur Handayani, Heru Wahyu Herwanto, and Yosi Kristian; visualization, Yuliana Melita Pranoto; supervision, Yuliana Melita Pranoto, Anik Nur Handayani, Heru Wahyu Herwanto, and Yosi Kristian; funding acquisition, Yuliana Melita Pranoto. All authors have read and agreed to the published version of the manuscript.

References

[1] N. Binsaif, "Application of information technology to e-commerce," *Int. J. Comput. Appl. Technol.*, vol. 68, no. 3, pp. 305–311, 2022.

[2] A. Bhatti, H. Akram, H. M. Basit, A. U. Khan, S. M. Raza, and M. B. Naqvi, "E-commerce trends during COVID-19 Pandemic," *Int. J. Futur. Gener. Commun. Netw.*, vol. 13, no. 2, pp. 1449–1452, 2020.

[3] M. K. Susmitha, "Impact of COVID-19 on E-Commerce," *J. Interdiscipl. Cycle Res.*, vol. 12, no. 9, pp. 1161–1165, 2021.

[4] A.-L. Scutariu, Ştefăniţă Şuşu, C.-E. Huidumac-Petrescu, and R.-M. Gogonea, "A cluster analysis concerning the behavior of enterprises with e-commerce activity in the context of the COVID-19 pandemic," *J. Theor. Appl. Electron. Commer. Res.*, vol. 17, no. 1, pp. 47–68, 2021.

[5] M. Takayanagi, O. Fukuda, N. Yamaguchi, H. Okumura and A. N. Handayani, "Vision-based Scene Recognition for Product Search," *2021 7th International Conference on Electrical, Electronics and Information Engineering (ICEEIE)*, Malang, Indonesia, pp. 1-5, 2021.

[6] P. Ristoski, P. Petrovski, P. Mika, and H. Paulheim, "A machine learning approach for product matching and categorization," *Semant. Web*, vol. 9, no. 5, pp. 707–728, 2018.

[7] N. Kertkeidkachorn and R. Ichise, "PMap: Ensemble Pre-training Models for Product Matching.," in *MWPD@ ISWC*, 2020.

[8] H. W. Herwanto, A. N. Handayani, A. P. Wibawa, K. L. Chandrika and K. Arai, "Comparison of Min-Max, Z-Score and Decimal Scaling Normalization for Zoning Feature Extraction on Javanese Character Recognition," *2021 7th International Conference on Electrical, Electronics and Information Engineering (ICEEIE)*, Malang, Indonesia, pp. 1-3, 2021.

[9] K. Dedes, A. B. P. Utama, A. P. Wibawa, A. N. Afandi, A. N. Handayani, and L. Hernandez, "Neural Machine Translation of Spanish-English Food Recipes Using LSTM," *JOIV Int. J. Informatics Vis.*, vol. 6, no. 2, pp. 290–297, 2022.

[10] E. I. Setiawan, A. Ferdianto, J. Santoso, Y. Kristian, S. Sumpeno, and M. H. Purnomo, "Analisis Pendapat Masyarakat terhadap Berita Kesehatan Indonesia menggunakan Pemodelan Kalimat berbasis LSTM (Indonesian Stance Analysis of Healthcare News using Sentence Embedding Based on LSTM)," *J. Nas. Tek. Elektro dan Teknol. Inf*, vol. 9, no. 1, pp. 8–17, 2020.

[11] A. Rachmadany, Y. M. Pranoto, and G. Gunawan, "Classification of Words of Wisdom in Indonesian on Twitter Using Naïve Bayes and Multinomial Naive Bayes," *Acad. Open*, vol. 3, pp. 10–21070, 2020.

[12] S. Mudgal *et al.*, "Deep learning for entity matching: A design space exploration," in *Proceedings of the 2018 International Conference on Management of Data*, pp. 19–34, 2018.

[13] L. Akritidis, A. Fevgas, P. Bozanis, and C. Makris, "A self-verifying clustering approach to unsupervised matching of product titles," *Artif. Intell. Rev.*, vol. 53, no. 7, pp. 4777–4820, Oct. 2020.

[14] A. Alabdullatif and M. Aloud, "AraProdMatch: A Machine Learning Approach for Product Matching in E-Commerce," *Int. J. Comput. Sci. Netw. Secur.*, vol. 21, no. 4, pp. 214–222, 2021.

[15] L. Akritidis and P. Bozanis, "Effective Unsupervised Matching of Product Titles with k-Combinations and Permutations," *2018 Innovations in Intelligent Systems and Applications (INISTA)*, Thessaloniki, Greece, pp. 1-10, 2018.

[16] K. Shah, S. Kopru, and J. D. Ruvini, "Neural network based extreme classification and similarity models for product matching," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language* Technologies, vol. 3 (Industry Papers), pp. 8–15, , 2018.

[17] D. Anggreani, I. A. E. Zaeni, A. N. Handayani, H. Azis and A. R. Manga', "Multivariate Data Model Prediction Analysis Using Backpropagation Neural Network Method," *2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT)*, Surabaya, Indonesia, pp. 239-243, 2021.

[18] A. N. Handayani, M. I. Akbar, H. Ar-Rosyid, M. Ilham, R. A. Asmara and O. Fukuda, "Design of SIBI Sign Language Recognition Using Artificial Neural Network Backpropagation," *2022 2nd International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA)*, Bandung, Indonesia, pp. 192-197, 2022.

[19] Y. M. Pranoto, A. N. Handayani, and Y. Kris-

tian, "Marketplace Product Image Grouping Using Transfer Learning of Deep Convolutional Neural Network in COVID-19 Post-Pandemic Situation," in *The Spirit of Recovery, CRC Press*, pp. 55–63, 2023.

[20] R. Shrivastava and D. S. Sisodia, "Product Recommendations Using Textual Similarity Based Learning Models," *2019 International Conference on Computer Communication and Informatics (ICCCI)*, Coimbatore, India, pp. 1-7, 2019.

[21] Q. Chen and M. Sokolova, "Specialists, scientists, and sentiments: Word2Vec and Doc2Vec in the analysis of scientific and medical texts," *SN Comput. Sci.*, vol. 2, pp. 1–11, 2021.

[22] A. Habib, M. Akram, and C. Kahraman, "Minimum spanning tree hierarchical clustering algorithm: A new Pythagorean fuzzy similarity measure for analyzing functional brain networks," *Expert Syst. Appl.*, vol. 201, p. 117016, 2022.

[23] E. S. Darmawan, "Ipsos Research Results: Shopee Named Most Used E-Commerce Platform in 2021." Kompas. Com. https://money.kompas.Com/read/2022/01/31/204500426, 2022.

[24] Suresh, "Shopee Train Images WithLabels Dataset. Retrieved June 24, 2022, from https://www.kaggle.com/datasets/dharmiksv/shopee-train-images-withlabels.," 2021.

[25] H. Lee and Y. Yoon, "Engineering doc2vec for automatic classification of product descriptions on O2O applications," *Electron. Commer. Res.*, vol. 18, pp. 433–456, 2018.

[26] H. B. Dogru, S. Tilki, A. Jamil and A. Ali Hameed, "Deep Learning-Based Classification of News Texts Using Doc2Vec Model," *2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA)*, Riyadh, Saudi Arabia, pp. 91-96, 2021.

[27] T. H. J. Hidayat, Y. Ruldeviyani, A. R. Aditama, G. R. Madya, A. W. Nugraha, and M. W. Adisaputra, "Sentiment analysis of Twitter data related to Rinca Island development using Doc2Vec and SVM and logistic regression as classifier," *Procedia Comput. Sci.*, vol. 197, pp. 660–667, 2022.

[28] M. Hanifi, H. Chibane, R. Houssin, and D. Cavallucci, ""Problem formulation in inventive design using Doc2vec and Cosine Similarity as Artificial Intelligence methods and Scientific Papers," *Eng. Appl. Artif. Intell.*, vol. 109, p. 104661, 2022.

[29] M. S El-Rahmany, E. Hussein Mohamed, and M. H Haggag, "Semantic detection of targeted attacks using DOC2VEC embedding," *J. Commun. Softw. Syst.*, vol. 17, no. 4, pp. 334–341, 2021.

[30] G. Wang and S. W. H. Kwok, "Using K-Means Clustering Method with Doc2Vec to Understand the Twitter Users' Opinions on COVID-19 Vaccination," *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, Athens, Greece, pp. 1-4, 2021.

[31] P. Shetty and S. Singh, "Hierarchical clustering: a survey," *Int. J. Appl. Res.*, vol. 7, no. 4, pp. 178–181, 2021.

[32] T. Li, A. Rezaeipanah, and E. M. T. El Din, "An ensemble agglomerative hierarchical clustering algorithm based on clusters clustering technique and the novel similarity measurement," *J. King Saud Univ. Inf. Sci.*, vol. 34, no. 6, pp. 3828–3842, 2022.

[33] A. Dogan and D. Birant, "K-centroid link: a novel hierarchical clustering linkage method," *Appl. Intell.*, pp. 1–24, 2022.

[34] L. L. Gao, J. Bien, and D. Witten, "Selective inference for hierarchical clustering," *J. Am. Stat. Assoc.*, pp. 1–11, 2022.

[35] Vijaya, S. Sharma and N. Batra, "Comparative Study of Single Linkage, Complete Linkage, and Ward Method of Agglomerative Clustering," *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, Faridabad, India, pp. 568-573, 2019.

[36] S. Abbasi, S. Nejatian, H. Parvin, V. Rezaie, and K. Bagherifard, "Clustering ensemble selection considering quality and diversity," *Artif. Intell. Rev.*, vol. 52, no. 2, pp. 1311–1340, 2019.

[37] X. Yang, J. Yan, Y. Cheng and Y. Zhang, "Learning Deep Generative Clustering via Mutual Information Maximization," in *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 9, pp. 6263-6275, Sept. 2023.

[38] M. Rahmanian and E. G. Mansoori, "An unsupervised gene selection method based on multivariate normalized mutual information of genes," *Chemom. Intell. Lab. Syst.*, vol. 222, p. 104512, 2022.

Yuliana Melita Pranoto born in Madiun, East Java, Indonesia, in 1982, has been a lecturer in the Department of Informatics at Institut STTS since 2005. She earned her bachelor's degree in Computer Science in 2005 and her master's degree in Information Technology in 2008 from Sekolah Tinggi Teknik Surabaya, Indonesia (now Institut Sains dan Teknologi Terpadu Surabaya or Institut STTS). Since 2022, she has been pursuing her doctoral studies at Universitas Negeri Malang, Indonesia. Her research interests are in Computer Vision and Deep Learning.

Anik Nur Handayani received the B.E. degree in Electronics Engineering from Brawijaya University, Malang, Indonesia, and the M.S. degree in Electronics Engineering, from Institute of Technology Sepuluh Nopember, Surabaya, Indonesia, in 2004 and 2008, and graduated from Saga University, Japan from Computer Science Department in 2014. Currently a lecture in Electrical and Informatic Engineering Department at Universitas Negeri Malang, Indonesia, started from 2005. Over 17 years on the institutions, have principally taught courses dealing with Artificial Intelligence and computer vision for Asisstive Device and Technology.

Heru Wahyu Herwanto is an academic at Universitas Negeri Malang. His expertise is in Image Processing, Artificial Intelligence, and their applications in education. He has published 20 scientific papers in reputable international journals and presented at international conferences. He is also actively involved in research and development in the fields of Image Processing, Artificial Intelligence, and their educational applications.

Yosi Kristian was born in Tuban, East Java, Indonesia, in 1981. He obtained his bachelor's degree in Computer Science and master's degree in Information Technology from Sekolah Tinggi Teknik Surabaya (now known as the Institut Sains dan Teknologi Terpadu Surabaya or Institut STTS) in 2004 and 2008, respectively. In 2018, he earned his Ph.D. from the Institut Teknologi Sepuluh Nopember (ITS) in Surabaya, Indonesia. In 2015, he was a research student at Osaka City University. Since 2004, Yosi has been a faculty member at Institut STTS, currently serving as an Associate Professor and Head of the Department of Informatics. His research interests include Machine Learning, Deep Learning, Artificial Intelligence, and Computer Vision.