Automatic Categorization of Multi Marketplace FMCGs Products using TF-IDF and PCA Features

Sri Suci Indasari[1]*, Aris Tjahyanto[2]
Information Systems [1], [2]
Sepuluh Nopember Institute of Technology
Surabaya, Indonesia
sri.suciindasari@gmail.com [1], aristj@its.ac.id [2]

***Abstract*— The use of technology in line with the increasing number of internet users has caused a shift in the product sales ecosystem to the realm of electronic commerce (electronic commerce). A total of 73.23 customers made purchase transactions using e-commerce and the most purchased products were products classified as Fast Moving Consumer Goods (FMCGs). The increasingly varied FMCGs data coupled with the increasing number of marketplaces is felt to need to be broken down into specific groups. The process is carried out by analyzing e-commerce product information, especially product names, and descriptions. In this study, we propose an automatic categorization of multiple marketplaces using data from multiple marketplaces. Data text is converted into structured data with a series of preprocessing, and comprehensive experiments are carried out to see the extraction performance of variables including TF-IDF, BOW, and N-Gram. All three methods are used to validate text data sets with K-Means grouping results used with the help of PCA to reduce data dimensions. The results show that the performance of the TF-IDF algorithm with a dimension reduction value of 70 and the use of Python can provide optimal results for the percentage of grouping data.**

Keywords*—*E-commerce, Fast Moving Consumer Goods, K-Means, Text Clustering, TF-IDF

***Abstrak*— Pemanfaatan teknologi sejalan dengan jumlah pengguna internet yang semakin bertambah menyebabkan pergeseran ekosistem penjualan produk ke ranah perdagangan elektronik (*electronic commerce*). Sebanyak 73,23 pelanggan melakukan transaksi pembelian menggunakan e-commerce serta produk yang paling banyak dibeli merupakan produk yang tergolong dalam Fast Moving Consumer Goods (FMCGs). Data FMCGs yang semakin bervariatif ditambah dengan jumlah marketplace yang semakin banyak dirasa perlu diuraikan menjadi kelompok-kelompok spesifik. Proses tersebut dilakukan dengan cara menganalisis informasi produk e-commerce khususnya nama dan deskripsi produk. Pada penelitian ini, kami mengusulkan pengkategorian otomatis multi marketplace dengan menggunakan data dari beberapa marketplace. Data teks diubah menjadi data terstruktur dengan rangkaian preprocessing, serta dilakukan eksperimen komprehensif untuk melihat kinerja ekstraksi variabel diantaranya TF-IDF, BOW dan N-Gram. Ketiga metode tersebut digunakan untuk memvalidasi kumpulan data teks dengan hasil pengelompokan K-Means yang digunakan dengan bantuan PCA untuk mereduksi dimensi data. Hasilnya menunjukkan bahwa kinerja dari algoritma TF-IDF dengan nilai reduksi dimensi 70 serta pemanfaatan python dapat memberikan hasil yang optimal terhadap persentase pengelompokan data.**

Kata Kunci*—*E-commerce, Fast Moving Consumer Goods, K-Means, Text Clustering, TF-IDF

I. INTRODUCTION

At present, companies utilize information technology as an online buying and selling platform. This is in line with the number of internet users has reached 62.10 percent of the total population in Indonesia and this number has increased by 22.2 percent since 2018 which amounted to 39.90 percent [1]. The use of electronic commerce (e-commerce) is a company solution to market its products instantly and can make it easier for consumers to find a product. Thus, encouraging companies to carry out digital transformation and invest more in online purchasing platforms to be able to compete in meeting customer satisfaction and needs. Companies can optimize existing business processes by improving user experience [2]. According to the Indonesian Central Bureau of Statistics, as many as 71.23 percent of customers make purchase transactions using e-commerce services [3]. Based on these numbers, it means that user preferences in searching for needs tend to use e-commerce.

The products offered on the e-commerce platform vary. One of them is a product that has a relatively short useful life and a relatively large amount of consumption, that is Fast Moving Consumer Goods (FMCGs). FMCGs are terms for products that are often bought, consumed quickly, sold in bulk and have relatively low prices [4]. FMCGs is "non-durable" items required for daily use. The following are examples of products classified as FMCGs in Fig. 1.

Fig. 1. Examples of FMCGs Product Variety

The FMCGs product sector is the largest industry in the world consisting of various products such as food, beverages,

electronic devices, household appliances, medicines, and others [5]. Consumers usually buy products in this category at least once a month. Based on data from the Central Statistics Agency, the most sold products via the Internet in 2020 were the food, beverage, and grocery categories by 40.86 percent, clothing by 20.71 percent, household needs by 20.30 percent, cosmetics by 8.05 percent and 38.40 percent for other product categories [3].

FMCGs products continue to emerge, resulting in more diverse data consumed by customers. Coupled with the emergence of various e-commerce with various conveniences offered. So multi-marketplace product grouping is needed by customers in product search for a price comparison between marketplaces. Each product has detailed information in the form of product name, price, rating, description, number of products sold, store name, store address, and others. This information can be used in identifying product similarities in multiple marketplaces.

The process of grouping products based on multi-marketplace categories generally uses soft computing methods. Several studies related to categorizing a product have been carried out with the use of machine learning. For example, research on the categorization of e-commerce products is carried out by proposing a machine translation paradigm using classification techniques by looking at nodes in the taxonomic tree [6]. Meanwhile, Chavaltada et al compared the performance of various machine learning techniques on product categorization in the proposed framework [7].

This study aims to group categories based on product information (e.g. product names and descriptions) available by emphasizing the use of tokenization processes and word weighting so that automatic product grouping is obtained. Several algorithms can be used in performing a grouping, especially with the use of unsupervised learning. A series of studies have applied unsupervised learning such as in analyzing mass spectrometry imaging [8], health [9][10][11], agriculture [12][13] and various other fields. Some use K-means clustering, hierarchical clustering, principal component analysis, and factor analysis algorithms. This research will use the concept of text clustering because it will identify, and group products based on product information (for example product names and descriptions) which are classified as unstructured data. This is done because there is research that states that text clustering based on similarities between texts is an efficient technique used to partition several documents into groups [14]. So that it will be tested in the process of grouping FMCGs product categories. In addition, the grouping will focus on the application of the K-means algorithm which is a non-hierarchical grouping method used to identify similarities between objects based on distance vectors and has efficiency, conciseness, and speed in its implementation [15] And this method is considered capable of carrying out the machine learning process quickly and providing optimal grouping results.

However, product information (product name and description) is unstructured text data, so the data will be processed first into structured data before being processed in unsupervised learning. So that experiments were conducted and analyzed to determine the performance of the process of converting unstructured data into structured with TF-IDF, BOW, and N-Gram in case text grouping based on FMCGs products can be more accurate. This process is done by converting text data to numeric by tokenizing each record and extracting it into a structured value. As well as using the Principal Component Analysis (PCA) method to reduce the dimensions of the data tokens formed and find out the relationship between variables.

Based on the explanation above, this study is expected to be able to provide a grouping of product name and description data based on the categorization of FMCGs products Automatically by applying machine learning technology that can divide data into clusters so that the same data group is formed.

II. THEORETICAL BASIS

Source data be present in deep shapes Virtual Form yard web, media, articles, blogs, and more. Partial big information gets taken deep a yard Web with access URL address websites and configuring such shape. This research retrieval data with the use of the technique of Crawling on an e-commerce website. This approach is already used in various research deep gather data picture [16][17][18], and text [19][20][21][22]. Crawling techniques be an Approach that gets gather data text and pictures from the website this research will Focuses on data name product and description of product that are classified as deep product FMCGs.

Machine Learning (ML) is a series of processes that can perform automated analysis of large amounts of data by finding patterns in data. In ML, developers strive to make the processes implemented in machines resemble human thought processes. ML is becoming an excellent tool for innovation due to its low computational costs, short development cycles, robust data analysis, and predictive capabilities [23]. Today, ML has been used in sensor utilization [24], digital transformation [25], manufacturing [26], healthcare [27] and has even entered the financial field [28][29].

In ML, there is a clustering method which is an unattended method that has no labeled inputs and problem-solving is based on the experience that algorithms gain from solving similar problems [30]. Because the data in this study uses text data, this study will focus on converting text data to vectors derived from the tokenization process. This clustering will group the vector value data based on the appropriate similarity.

Research conducted by Sen Xu et al [31] using text data to do clustering, that the algorithm applied provides optimal results using small data, but in large-scale data is still not able to provide good results. In addition, Diallo et al [32] in conducting text clustering compared similarities to Cosine and Euclidean Distances. However, this study will focus on comparing text weighting methods to obtain optimal clustering results, including TF-IDF, BOW, and N-Gram.

In addition, this study will perform dimension reduction using Principal Component Analysis (PCA). PCA can be used to analyze relationships between variables. In clustering, PCA is useful for identifying the main variable and its effect on the target variable [33]. It is used to reduce the main variables in the clustering process.

III. RESEARCH METHODOLOGY

This research involves the process of preprocessing, variable selection using text weighting methods based on words

contained in the data, dimensional reduction, and K-Means clustering techniques for grouping data based on the similarity of each record, as in Fig. 2.

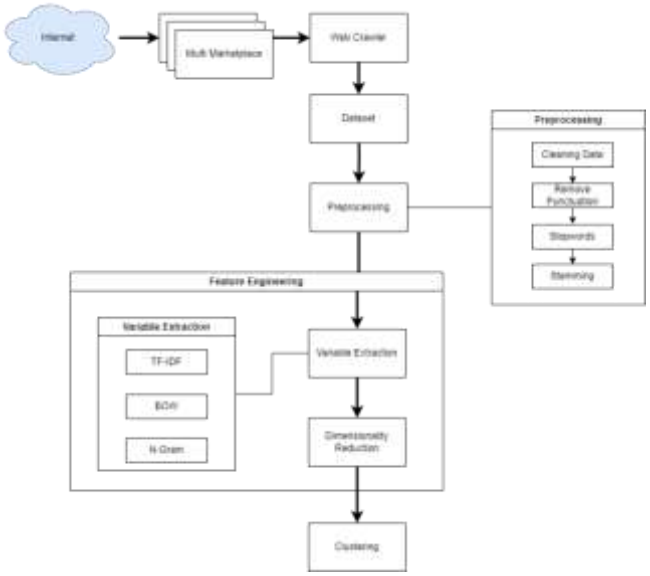

Fig. 2. Research Methodology

A. Internet, Multimarketplace, Web Crawler & Files

Various information is available on the internet, ranging from the latest news, stock prices, buying and selling products, to future predictions. Therefore, data from this study was obtained from several marketplaces, namely Shopee, Bukalapak, and Lazada which can be accessed via the internet using web crawlers. The web crawler is a data collection technique by capturing information on e-commerce websites so that the information taken is in the form of product name and description data.

B. Preprocessing

Preprocessing is a series of stages that are carried out before data is processed. This stage aims to generalize the form of data such as removing symbols or numbers that are not important, eliminating words that often appear, and returning data to its original form. In this study, a series of preprocessing processes were carried out by cleaning data, removing punctuation, symbols, and numbers (remove punctuation), removing unused words (remove stopwords), and changing the form of data to basic words (stemming).

C. Variable Extraction

Variable extraction is a process for converting text data into numeric form. This study will test and compare the results of variable extraction between TF-IDF, BOW, and N-Gram with the resulting K-Means grouping.

1) TF-IDF is applied to measure term frequency, then filter out words that appear with very low frequency [32]. This algorithm performs TF and IDF calculations of documents. Here's the TF and IDF formula in equations (1) and (2).

$$TF = \frac{Number\ of\ times\ the\ term\ appears\ in\ the\ document}{Total\ number\ of\ term\ in\ the\ document} \quad (1)$$

$$IDF = Log(\frac{Number\ of\ documnet\ in\ the\ corpus}{Number\ of\ document\ in\ the\ corpus\ contain\ the\ term}) \quad (2)$$

2) BOW (Bag Of Words): one way of extracting variables from text into numbers by representing textual documents as sparse vectors of word counts [34].
3) N-Gram: a text preprocessing model that has a method to improve character transformations. N-Gram is a frequently used method in which n indicates a continuous number of terms or words as well as consists of a collection of document-size character sets (bi-gram, tri-gram, quad-gram) [35].

D. Dimensionality Reduction

Dimension reduction is a stage used to minimize the number of variable inputs before clustering. The dimension reduction used in this study is *Principal Component Analysis* (PCA). PCA is used to transform high-dimensional data to lower dimensions, this is done by using several main components so that the transformed dimensions are reduced [36].

E. Clustering

Clustering is a type of unsupervised machine learning, where this study will focus on K-Means that can identify similarities between variables based on distance vectors and have efficiency, conciseness, and speed in its implementation. In general, K-Means Clustering in grouping variables starts from the following stages:

1) Specify the K value as the number of initial clusters (centroids) you want to form.
2) Calculate the distance of data with a centroid using the Euclidean Distance formula to find the closest distance of data with a centroid. Here's the Euclidean Distance equation (3):

$$d(x_i, \mu_j) = \sqrt{\sum(x_i - \mu_j)^2} \quad (3)$$

Where, x_i and μ_j is the number of attribute values of the variable.
3) Classify each data based on its proximity to the centroid.
4) Update the new centroid value obtained from the cluster average, using the formula:

$$\mu j\ (t+1) = \frac{1}{N_{sj}} \sum j \in sj^{xj} \quad (3)$$

Where $\mu j\ (t+1)$ is the new centroid in the (t+1) iteration and Nsj is a lot of data on the Sj cluster.
5) Steps 2 through 4 are repeated until the value of the centroid point is constant.

However, in its implementation, Python already has a library to run the K-Means algorithm, namely with the sklearn.cluster library and import K-Means.

IV. Results and Discussion

A total of 300 e-commerce data were collected in this research consisting of three e-commerce, namely Lazada, Shopee, and Bukalapak. Each e-commerce was composed of

categories of clothing, beauty equipment, electronics, and health. The category data is used as *ground truth* to analyze and compare the results of groupings carried out in this research design.

The following are the preprocess stages carried out in this study:

1) Data Cleaning: steps to remove unnecessary columns in the crawled dataset. This aims to streamline data and only focuses on the data you want to process. Here's a comparison of before and after cleaning on Table I and II.

TABLE I. BEFORE CLEANING DATA

Web-scrap-order	Web-scraper-start-URL	Linked-href	Name	Description
1670898723-453	https://www.bukalapak.com/products?search%5Bkeywords%5D=baju%20pria	https://www.bukalapak.com/p/fashion-pria/kaos-165/3310ht3-jual-termurah-baju-kaos-polos-pendek-distro-katun-combed-premium-misty-unisex-cocok-untuk-pria-dan-wanita?from=list-product&pos=9	TERMURAH - Baju Kaos Polos Pendek Distro Katun Combed Premium Misty - Unisex Cocok Untuk Pria Dan Wanita	CATATAN ORDER : JIKA VARIAN WARNA YANG DI INGINKAN TIDAK TERSEDIA DI PILIHAN SILAHKAN TULIS WARNA YANG DI INGINKAN DI CATATAN AGAR MENGHINDARI TERJADINYA PENGIRIMAN RANDOM ……..

TABLE II. AFTER CLEANING DATA

Name	Description
TERMURAH - Baju Kaos Polos Pendek Distro Katun Combed Premium Misty - Unisex Cocok Untuk Pria Dan Wanita	CATATAN ORDER: JIKA VARIAN WARNA YANG DI INGINKAN TIDAK TERSEDIA DI PILIHAN SILAHKAN TULIS WARNA YANG DI INGINKAN DI CATATAN AGAR MENGHINDARI TERJADINYA PENGIRIMAN RANDOM…

2) Remove Punctuation: steps to remove punctuation, symbols, numbers, and uniformity of letters into lowercase. This aims to homogenize the form of data to be more effective in the subsequent data processing. Here's a comparison before and after the remove of the punctuation process in Table III.

TABLE III. COMPARISON OF DATA BEFORE AND AFTER REMOVE PUNCTUATION

Before Remove Punctuation	After Remove Punctuation
TERMURAH - Baju Kaos Polos Pendek Distro Katun Combed Premium Misty - Unisex Cocok Untuk Pria Dan Wanita	termurah baju kaos polos pendek distro katun combed premium misty unisex cocok untuk pria dan wanita

3) Stopwords: steps to remove unimportant words, such as connecting words, subjects that indicate people, or words that are most encountered. The set of words can be found in the Indonesian version of the Python nltk library and can be added manually by creating a list of words that you want to remove. The set of words you want to remove is called a stoplist. Here's a comparison before and after the stopword in Table IV.

TABLE IV. COMPARISON OF DATA BEFORE AND AFTER STOPWORDS

Before Stopword	After Stopword
termurah baju kaos polos pendek distro katun combed premium misty unisex cocok untuk pria dan wanita	baju kaos distro katun combed misty unisex

It is caused because the words "termurah", "pendek", "premium", "pria", "dan", "wanita" are included in the stoplist, so the word is omitted.

4) Stemming: the step to change the form of a word into its basic structure by removing the affixes present in the word. The porter stemming algorithm is applied in this stage. Here's a comparison before and after stemming from Table V.

TABLE V. COMPARISON OF DATA BEFORE AND AFTER STEMMING

Before Stemming	After Stemming
suplemen makanan kaya kasium melindungi tulang kepadatannya diperkaya vitamin c dan b6	suplemen makan kaya kasium lindung tulang padat kaya vitamin c b6

The next stage is variable extraction, stemming data is processed by testing three methods, namely TF-IDF, BOW, and N-Gram. In terms of experimentation, this study also verified the complexity of PCA and its effect on clustering, results in FMCGs product datasets. And will compare the results of clustering against Weka and Python tools.

TABLE VI. VARIABLE EXTRACTION BASED ON PCA VALUE USING WEKA

Variable Extraction	Cluster	Principal Component Analysis			
		PCA40	*PCA50*	*PCA60*	*PCA70*
TF-IDF	C1	13%	4%	22%	20%
	C2	60%	18%	18%	31%
	C3	14%	42%	41%	34%
	C4	14%	36%	19%	15%
BOW	C1	96%	96%	96%	96%
	C2	0%	0%	0%	0%
	C3	0%	0%	0%	0%
	C4	3%	3%	3%	3%
N-Gram	C1	96%	96%	96%	9.5%
	C2	0%	0%	1%	1%
	C3	0%	0%	0%	0%
	C4	3%	4%	3%	4%

*C is the cluster formed along with the percentage of each cluster

Based on Table VI, if the processing uses Weka tools, the application of TF-IDF with PCA 60 shows the highest results so the results obtained are also optimal. Meanwhile, when using Python in development, the results obtained are as in the following Table VII.

TABLE VII. VARIABLE EXTRACTION BASED ON PCA VALUE USING PYTHON

Variable Extraction	Cluster	Principal Component Analysis			
		PCA40	*PCA50*	*PCA60*	*PCA70*
TF-IDF	C1	53%	31%	38%	26%
	C2	4%	46%	4%	12%
	C3	6%	18%	42%	34%
	C4	36%	5%	16%	27%
BOW	C1	5%	49%	51%	16%
	C2	84%	38%	6%	11%
	C3	6%	7%	34%	12%
	C4	4%	7%	9%	60%
N-Gram	C1	91%	4%	76%	78%
	C2	2%	5%	5%	11%
	C3	4%	88%	10%	7%
	C4	3%	2%	9%	4%

The application of Table VII above can be concluded that the application of TF-IDF with a PCA value of 70 is closer to the value of balanced data division of 25%. This is different from the use of WEKA tools, with Python implementation on datasets able to provide better clustering of variable extraction using BOW and N-Gram. These results corroborate the results of research conducted by Huong et al regarding the comparison of several word vector representation techniques applied to Vietnamese sentiment analysis [37]. It is also in agreement with the study using Arabic Twitter data on distance education attitude analysis that the application of TF-IDF variable extraction works better than the AraVec model.

Furthermore, evaluation was carried out in this study, including the Elbow technique to determine the optimal cluster formed. In addition, Elbow was able to see the percentage comparison of the number of clusters with the number of clusters added [39] and the evaluation of the Silhouette Coefficient and Davies Bouldin Index. The Davies Bouldin Index is one of the methods used to measure cluster validity by summing the proximity of the cluster center point of the cluster followed [40], while silhouette coefficients are used to see how well a particular cluster is separated from others [41]. The following results of the Elbow method can be seen in Fig. 3.

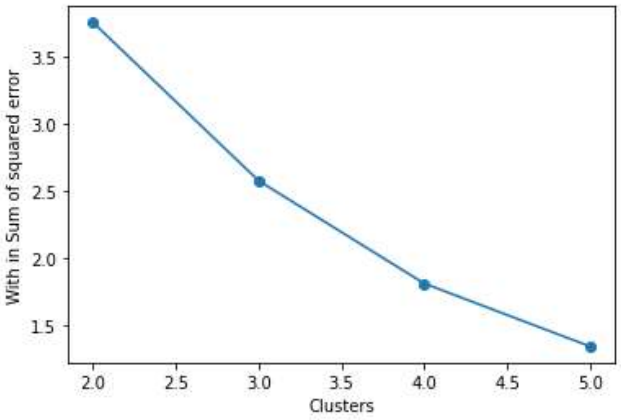

Fig. 3. Elbow Method Matrix

Meanwhile, to find out the results of the Davies Bouldin Index and Silhouette Coefficients can be seen in Table VIII.

TABLE VIII. EVALUATION OF CLUSTERING RESULTS

K value	*Silhouette Coefficient*	*Davies Bouldin Index*
2	0.517	0.667
3	0.531	0.838
4	0.562	0.669
5	0.424	0.726

The highest Silhouette Coefficient evaluation results show good results, while the Davies Bouldin Index evaluation shows good results indicated by minimum values. Based on Table VIII, the value that has optimal results is found in the value K=4 or with the division of 4 clusters.

V. CONCLUSION

Based on the description above, this research proposes to automatically FMCGs products as categorized multi-marketplace. Because it uses relatively much data, the deciphering of words that are not used needs to be considered, especially at the preprocessing stage. This will affect the results of variable extraction and the groupings formed. Experiments

were conducted on three variable extraction algorithms namely TF-IDF, BOW, and N-Gram. In addition, this study also applies PCA to summarize variable data tables from a large scale into smaller sets of variables (summary index).

The experimental analysis above shows that the extraction of TF-IDF variables results in a multi-marketplace product categorization that is close to the optimal value of the cluster if the grouping formed is four clusters. This is to research conducted [22] that variable extraction is very influential in clustering models, and TF-IDF is a variable extraction method that can provide significant results [42]. Meanwhile, the application of PCA also has a significant influence on cluster results, namely on TF-IDF with a PCA value of 70, and the use of Python can optimize the data so that the results provided can be better. In addition, this research will continue to be developed both in testing the results formed when the number of data product is expanded and by conducting multilevel grouping to determine multi-automatic product subcategories marketplace by specifying the results of the category grouping produced in this study.

REFERENCES

[1] B.-S. Indonesia, "Statistik Telekomunikasi Indonesia," 2021.

[2] H. Al Mashalah, E. Hassini, A. Gunasekaran, and D. Bhatt (Mishra), "The impact of digital transformation on supply chains through e-commerce: Literature review and a conceptual framework," *Transp. Res. Part E Logist. Transp. Rev.*, vol. 165, no. August, p. 102837, 2022, doi: 10.1016/j.tre.2022.102837.

[3] B.-S. Indonesia, *Statistik E-Commerce 2021*. 2021.

[4] A. Etuk, J. A. Anyadighibe, E. E. James, and P. M. Egemba, "Trade sales promotion and distributors' performance of fast-moving consumer goods (FMCGS)," *Int. Res. J. Manag. IT Soc. Sci.*, vol. 9, no. 2, pp. 254–263, 2022, doi: 10.21744/irjmis.v9n2.2011.

[5] A. O. Binuyo, H. Ekpe, and B. O. Binuyo, "Innovative strategies and firm growth: Evidence from selected fast moving consumer goods firms in Lagos state, Nigeria," *Probl. Perspect. Manag.*, vol. 17, no. 2, pp. 313–322, 2019, doi: 10.21511/ppm.17(2).2019.24.

[6] L. Tan, M. Y. Li, and S. Kok, "E-Commerce Product Categorization via Machine Translation," *ACM Trans. Manag. Inf. Syst.*, vol. 11, no. 3, 2020, doi: 10.1145/3382189.

[7] C. Chavaltada, K. Pasupa, and D. R. Hardoon, "A comparative study of machine learning techniques for Automatic Product Categorisation," *Adv. Intell. Syst. Comput.*, vol. 906, pp. 459–464, 2019, doi: 10.1007/978-981-13-6001-5_37.

[8] N. Verbeeck, R. M. Caprioli, and R. Van de Plas, "Unsupervised machine learning for exploratory data analysis in imaging mass spectrometry," *Mass Spectrom. Rev.*, vol. 39, no. 3, pp. 245–291, 2020, doi: 10.1002/mas.21602.

[9] C. M. Eckhardt *et al.*, "Unsupervised machine learning methods and emerging applications in healthcare," *Knee Surgery, Sport. Traumatol. Arthrosc.*, no. 0123456789, 2022, doi: 10.1007/s00167-022-07233-7.

[10] S. M. D. A. C. Jayatilake and G. U. Ganegoda, "Involvement of Machine Learning Tools in Healthcare Decision Making," *J. Healthc. Eng.*, vol. 2021, 2021, doi: 10.1155/2021/6679512.

[11] J. Pereira and M. Silveira, "Learning Representations from Healthcare Time Series Data for Unsupervised Anomaly Detection," *2019 IEEE Int. Conf. Big Data Smart Comput. BigComp 2019 - Proc.*, 2019, doi: 10.1109/BIGCOMP.2019.8679157.

[12] S. T. Jagtap, K. Phasinam, T. Kassanuk, S. S. Jha, T. Ghosh, and C. M. Thakar, "Towards application of various machine learning techniques in agriculture," *Mater. Today Proc.*, vol. 51, pp. 793–797, 2021, doi: 10.1016/j.matpr.2021.06.236.

[13] M. Xu, S. Yoon, J. Lee, and D. S. Park, "Unsupervised Transfer Learning for Plant Anomaly Recognition," vol. 11, no. 4, pp. 30–37, 2022.

[14] L. Abualigah *et al.*, "Nature-inspired optimization algorithms for text document clustering—a comprehensive analysis," *Algorithms*, vol. 13, no. 12, pp. 1–32, 2020, doi: 10.3390/a13120345.

[15] D. Abdullah, S. Susilo, A. S. Ahmar, R. Rusli, and R. Hidayat, "The application of K-means clustering for province clustering in Indonesia of the risk of the COVID-19 pandemic based on COVID-19 data," *Qual. Quant.*, vol. 56, no. 3, pp. 1283–1291, 2022, doi: 10.1007/s11135-021-01176-w.

[16] J. Hwang, J. Kim, S. Chi, and J. O. Seo, "Development of training image database using web crawling for vision-based site monitoring," *Autom. Constr.*, vol. 135, p. 104141, 2022, doi: 10.1016/j.autcon.2022.104141.

[17] X. Zhouyi, H. Weijun, and H. Yanrong, "INTELLIGENT ACQUISITION METHOD OF HERBACEOUS FLOWERS IMAGE BASED ON THEME CRAWLER, DEEP LEARNING AND GAME THEORY," vol. 3, no. 65, pp. 44–52, 2022.

[18] J. Yang *et al.*, "Unified Contrastive Learning in Image-Text-Label Space," pp. 19141–19151, 2022, doi: 10.1109/cvpr52688.2022.01857.

[19] N. Kumar, M. Gupta, D. Sharma, and I. Ofori, "Technical Job Recommendation System Using APIs and Web Crawling," *Comput. Intell. Neurosci.*, vol. 2022, p. 7797548, 2022, doi: 10.1155/2022/7797548.

[20] S. Neelakandan, A. Arun, R. R. Bhukya, B. M. Hardas, T. C. Anil Kumar, and M. Ashok, "An Automated Word Embedding with Parameter Tuned Model for Web Crawling," *Intell. Autom. Soft Comput.*, vol. 32, no. 3, pp. 1617–1632, 2022, doi: 10.32604/IASC.2022.022209.

[21] S. Gupta and K. K. Bhatia, "Design of a Parallel and Scalable Crawler for the Hidden Web," *Int. J. Inf. Retr. Res.*, vol. 12, no. 1, pp. 1–23, 2022, doi: 10.4018/ijirr.289612.

[22] P. Weninggalih and Y. Sibaroni, "Identify User Behavior based on Tweet Type on Twitter Platform using Agglomerative Hierarchical Clustering," *J. Media Inform. Budidarma*, vol. 6, no. 3, p. 1404, 2022, doi: 10.30865/mib.v6i3.4342.

[23] C. Gao *et al.*, "Innovative Materials Science via Machine Learning," *Advantaced Funct. Mater.*, vol. 32, no. 1, 2022.

[24] A. Venketeswaran *et al.*, "Recent Advances in Machine Learning for Fiber Optic Sensor Applications," *Adv. Intell. Syst.*, vol. 4, no. 1, p. 2100067, 2022, doi: 10.1002/aisy.202100067.

[25] A. Sarirete, Z. Balfagih, T. Brahimi, M. D. Lytras, and A. Visvizi, "Artificial intelligence and machine learning research: towards digital transformation at a global scale," *J. Ambient Intell. Humaniz. Comput.*, vol. 13, no. 7, pp. 3319–3321, 2022, doi: 10.1007/s12652-021-03168-y.

[26] J. Qin *et al.*, "Research and application of machine learning for additive manufacturing," *Addit. Manuf.*, vol. 52, no. October 2021, 2022, doi: 10.1016/j.addma.2022.102691.

[27] M. D. McCradden *et al.*, "A Research Ethics Framework for the Clinical Translation of Healthcare Machine Learning," *Am. J. Bioeth.*, vol. 22, no. 5, pp. 8–22, 2022, doi: 10.1080/15265161.2021.2013977.

[28] M. M. M. Megdad, B. S. Abu-Nasser, and S. S. Abu-Naser, "Fraudulent Financial Transactions Detection Using Machine Learning," *Int. J. Acad. Inf. Syst. Res.*, vol. 6, no. 3, pp. 30–39, 2022, [Online]. Available: www.ijeais.org/ijaisr.

[29] R. Rawat, Y. N. Rimal, P. William, S. Dahima, S. Gupta, and K. S. Sankaran, "Malware Threat Affecting Financial Organization Analysis Using Machine Learning Approach," *Int. J. Inf. Technol. Web Eng.*, vol. 17, no. 1, pp. 1–20, 2022, doi: 10.4018/ijitwe.304051.

[30] K. K. Bhardwaj, S. Banyal, and D. K. Sharma, "Chapter 7 - Artificial Intelligence Based Diagnostics, Therapeutics and Applications in Biomedical Engineering and Bioinformatics," *Internet Things Biomed. Eng.*, pp. 161–187, 2019.

[31] S. Xu *et al.*, "An integrated K-means – Laplacian cluster ensemble approach for document datasets," *Neurocomputing*, vol. 214, pp. 495–507, 2016, doi: 10.1016/j.neucom.2016.06.034.

[32] B. Diallo, J. Hu, T. Li, G. A. Khan, and A. S. Hussein, "Multi-view document clustering based on geometrical similarity measurement," *Int. J. Mach. Learn. Cybern.*, vol. 13, no. 3, pp. 663–675, 2022, doi: 10.1007/s13042-021-01295-8.

[33] M. kyu Kim, J. W. Chang, K. Park, and D. R. Yang, "Comprehensive assessment of the effects of operating conditions on membrane

intrinsic parameters of forward osmosis (FO) based on principal component analysis (PCA)," *J. Memb. Sci.*, vol. 641, no. September 2021, p. 119909, 2022, doi: 10.1016/j.memsci.2021.119909.

[34] S. Wattanakriengkrai *et al.*, "Automatic Classifying Self-Admitted Technical Debt Using N-Gram IDF," *Proc. - Asia-Pacific Softw. Eng. Conf. APSEC*, vol. 2019-December, pp. 316–322, 2019, doi: 10.1109/APSEC48747.2019.00050.

[35] T. Hasan and A. Matin, *Extract Sentiment from Customer Reviews: A Better Approach of TF-IDF and BOW-Based Text Classification Using N-Gram Technique*. Springer Singapore, 2021.

[36] M. R. Mahmoudi, M. H. Heydari, S. N. Qasem, A. Mosavi, and S. S. Band, "Principal component analysis to study the relations between the spread rates of COVID-19 in high risks countries," *Alexandria Eng. J.*, vol. 60, no. 1, pp. 457–464, 2021, doi: 10.1016/j.aej.2020.09.013.

[37] T. H. Huong, K. Tran-Trung, D. T. C. Lai, and V. T. Hoang, "Sentiment Analysis based on word vector representation for short comments in Vietnamese language," *Proc. - 2022 9th NAFOSTED Conf. Inf. Comput. Sci. NICS 2022*, pp. 165–169, 2022, doi: 10.1109/NICS56915.2022.10013426.

[38] T. Alqurashi, "Stance Analysis of Distance Education in the Kingdom of Saudi Arabia during the COVID-19 Pandemic Using Arabic Twitter Data," *Sensors*, vol. 22, no. 3, 2022, doi: 10.3390/s22031006.

[39] R. Nainggolan, R. Perangin-Angin, E. Simarmata, and A. F. Tarigan, "Improved the Performance of the K-Means Cluster Using the Sum of Squared Error (SSE) optimized by using the Elbow Method," *J. Phys. Conf. Ser.*, vol. 1361, no. 1, 2019, doi: 10.1088/1742-6596/1361/1/012015.

[40] M. Mughnyanti, S. Efendi, and M. Zarlis, "Analysis of determining centroid clustering x-means algorithm with davies-bouldin index evaluation," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 725, no. 1, 2020, doi: 10.1088/1757-899X/725/1/012128.

[41] A. M. Bagirov, R. M. Aliguliyev, and N. Sultanova, "Finding compact and well-separated clusters: Clustering using silhouette coefficients," *Pattern Recognit.*, vol. 135, 2023, doi: 10.1016/j.patcog.2022.109144.

[42] J. ZHU, S. HUANG, Y. SHI, K. WU, and Y. WANG, "A Method of K-Means Clustering Based on TF-IDF for Software Requirements Documents Written in Chinese Language," *IEICE Trans. Inf. Syst.*, vol. 105, no. 4, pp. 736–754, 2022, doi: 10.1587/transinf.2021EDP7144.