Classification of E-Commerce Product Descriptions with The TF-IDF and SVM Methods

Dagobert Pakpahan[1)], **Veronika Siallagan**[2)], **Saut Dohot Siregar**[3)*]
[1,2,3)] Universitas Prima Indonesia, Indonesia
[1)]jeremianadeak02@gmail.com, [2)]veronikasiallagan48@gmail.com,[3)]saut.unpri@gmail.com

Submitted : Jul 25, 2023 | **Accepted** : Aug 5, 2023 | **Published** : Oct 1, 2023

Abstract: The rapidly growing e-commerce sector presents a significant challenge in navigating an abundance of products. Understanding and classifying product descriptions efficiently and accurately is crucial to improving user experience and business operations. This research employed the Term Frequency-Inverse Document Frequency (TF-IDF) algorithm and Support Vector Machine (SVM) for the classification of e-commerce product descriptions into four categories: Electronics, Household Items, Books, and Clothing. The initial phase involved pre-processing of text data which incorporated text cleaning, tokenization, part-of-speech tagging, entity recognition, and conversion into a vector representation. The resulting model was trained and tested using the SVM algorithm. Our model demonstrated a high degree of accuracy, achieving 99.2% during the training phase and 95.7% in the testing phase. This model provides a valuable tool for e-commerce businesses, as it allows for accurate classification of products based on their descriptions. This could lead to improved user navigation and overall user experience on e-commerce platforms.

Keywords: E-commerce Product Classification; Support Vector Machine; SVM; Natural Language Processing; TF-IDF

INTRODUCTION

In the recent study conducted by Nasution et al. (Nasution et al., 2020), it was demonstrated that the rapid development of e-commerce has created its market and has had an impact on the digital business sector, but it also brings a variety of challenges. The study found that people have difficulties choosing the right e-commerce product due to the multitude of available products. On the other hand, Alhamra in his research on designing a sneaker product search application explained that people find it hard to select the product with the best price and quality, emphasizing the need for better product search technology (Alhamra, 2018).

To analyze text data, a method capable of processing text data is required first. One method explored for text classification problems is the Term Frequency-Inverse Document Frequency (TF-IDF) (Naraloka, 2021). TF-IDF is used to extract important features from text data by calculating the frequency of word occurrences in a document and considering the importance of these words in the document and the entire document collection. By applying the TF-IDF method, text data can be represented by feature vectors reflecting the relative importance of the words in the description. The TF-IDF method has been used to solve various problems related to text data analysis, including text classification (Arifin et al., 2021), public sentiment analysis about the Covid-19 lockdown policy (Isnain et al., 2021), similarity search of thesis titles (Amrulloh & Adam, 2021), and public sentiment on the creative economy (Naraloka, 2021).

*saut.unpri@gmail.com

Furthermore, once the text has been successfully converted into a vector form understandable by the algorithm, an algorithm like the Support Vector Machine (SVM) (Isnain et al., 2021; Mutawalli et al., 2019; Ristiano, 2022) is needed to classify the text. SVM is an effective machine learning algorithm in handling classification problems by building a model that separates two classes with a maximum margin. Several previous studies have proven that SVM can produce high-accuracy text classification for classifying social media text about the news of Wiranto's stabbing (Mutawalli et al., 2019), SMS text of Spam (Ristiano, 2022), social media Twitter about Covid-19 lockdown (Isnain et al., 2021), and beauty product reviews (Pratiwi et al., 2021).

The dataset consists of big data in the form of text, which is product description text, categorized into four types, namely "Electronics," "Household," "Books," and "Clothing & Accessories," with a total of 50425 data. This dataset, in ".csv" format, consists of two columns, with the first column containing the class names and the second containing the related products and descriptions. This dataset is a representation of an e-commerce site, obtained through the scraping process, and has multivariate characteristics. The primary purpose of using this dataset is to carry out classification tasks with the help of the SVM and TF-IDF algorithms. The proposed solution is an algorithm designed with a Python 3-based Jupyter Notebook program to classify e-commerce product descriptions. The importance of assisting consumers in selecting the right e-commerce product motivates researchers to undertake research that can help solve this problem. Besides, the dataset used is a product description, and the algorithms used are TF-IDF and SVM

LITERATURE REVIEW

The importance of assisting consumers in choosing the right E-Commerce product motivates researchers to conduct studies that can help solve this problem. Moreover, the methods employed are inspired from the previous research which are TF-IDF and SVM, the dataset used is different from the previous research (Arifin et al., 2021) and consists of product descriptions and its product type.

Support Vector Machine (SVM) is a powerful and widely used machine learning algorithm for both classification and regression tasks. It is a supervised learning algorithm that can be applied to solve binary and multiclass classification problems. The main idea behind SVM is to find the optimal hyperplane in a high-dimensional feature space that best separates different classes of data points. The hyperplane is chosen in such a way that it maximizes the margin, which is the distance between the hyperplane and the closest data points of each class. These closest data points are called support vectors, hence the name Support Vector Machine.The kernel trick allows SVM to implicitly map the data points into a higher-dimensional space, where it becomes possible to find a hyperplane that can separate the data. After feature extraction, the data is ready to be classified. In this research, the SVM method is used to classify product descriptions into four categories. SVM works by building a model that can separate two classes with a maximum margin. SVM is considered effective for text classification because it can handle high-dimensional data and produce high accuracy (Suryati et al., 2023).

METHOD

The type of research is a trial analysis aimed at automatically classifying product descriptions based on categories using the Term Frequency-Inverse Document Frequency (TF-IDF) method and Support Vector Machine (SVM). The following are the stages of research with the SVM method:

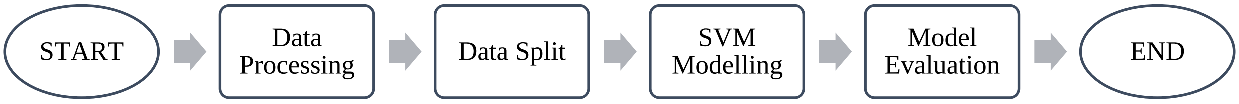


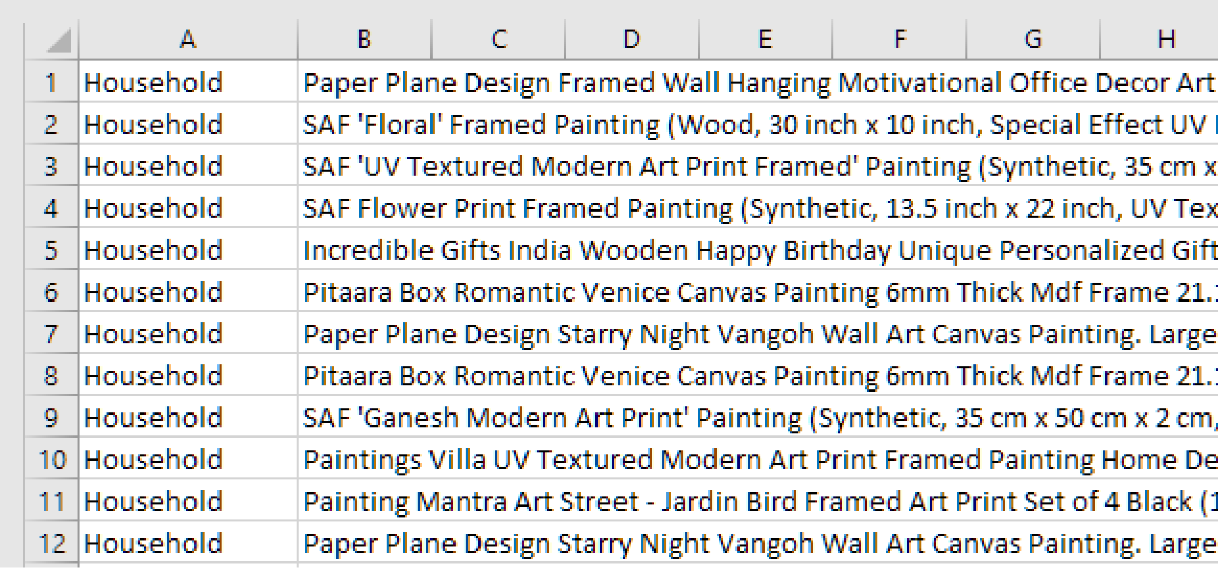
Fig 1. Research Process

As shown in the image above, the stages of this research are as follows:

*saut.unpri@gmail.com

1. Data Preprocessing: Cleaning and normalizing text, as well as removing non-informative words and extracting features with TF-IDF which converts text into vectors.
2. Data Split: Dividing the product description data into 80% Training Data and 20% Testing Data.
3. SVM Modelling: Creating a classification model with the Support Vector Machine algorithm to group product descriptions.
4. Model Evaluation: Measuring the performance of the model with the Accuracy metric.

Product descriptions are informative texts that explain the details and features of a product. They are commonly used in e-commerce platforms to provide information to consumers before making a purchase. By conducting research on product descriptions, we can optimize the way information is presented effectively, automatically classify products into the appropriate categories, and ultimately enhance customer satisfaction and the operational efficiency of sellers or e-commerce platforms (Rahmawati & others, 2019).



	A	B	C	D	E	F	G	H
1	Household	Paper Plane Design Framed Wall Hanging Motivational Office Decor Art						
2	Household	SAF 'Floral' Framed Painting (Wood, 30 inch x 10 inch, Special Effect UV I						
3	Household	SAF 'UV Textured Modern Art Print Framed' Painting (Synthetic, 35 cm x						
4	Household	SAF Flower Print Framed Painting (Synthetic, 13.5 inch x 22 inch, UV Tex						
5	Household	Incredible Gifts India Wooden Happy Birthday Unique Personalized Gift						
6	Household	Pitaara Box Romantic Venice Canvas Painting 6mm Thick Mdf Frame 21.:						
7	Household	Paper Plane Design Starry Night Vangoh Wall Art Canvas Painting. Large						
8	Household	Pitaara Box Romantic Venice Canvas Painting 6mm Thick Mdf Frame 21.:						
9	Household	SAF 'Ganesh Modern Art Print' Painting (Synthetic, 35 cm x 50 cm x 2 cm,						
10	Household	Paintings Villa UV Textured Modern Art Print Framed Painting Home De						
11	Household	Painting Mantra Art Street - Jardin Bird Framed Art Print Set of 4 Black (1						
12	Household	Paper Plane Design Starry Night Vangoh Wall Art Canvas Painting. Large						

Fig 2. Dataset View in .CSV Format

The data used in this research is sourced from https://www.kaggle.com/datasets/saurabhshahane/ecommerce-text-classification. The dataset is in text format and contains 50,425 samples of product descriptions with two columns. The first column contains the product categories (electronics, household items, books, clothing & accessories), and the second column contains the product descriptions. The data is collected in .csv format and was obtained through a scraping process from e-commerce websites.

Before the data is processed, several preprocessing steps cam be performed to improve the quality of data (Kurniawan et al., 2020), including:

1. Data Cleaning: Removing irrelevant data, such as symbols and numbers unrelated to the description.
2. Normalization: Converting all text into lowercase.
3. Stopword Removal: Removing common words that do not have significant meaning, such as "and", "in", "which".

After preprocessing, feature extraction is done using the TF-IDF method. TF-IDF is used to convert text into a vector form and identify the most important words in each document. This is done by calculating the frequency of word occurrences in a document and the importance of these words in the entire document collection (Prastyo et al., 2020).

After feature extraction, the data is ready to be classified. In this research, the SVM method is used to classify product descriptions into four categories. SVM works by building a model that can separate two classes with a maximum margin. SVM is considered effective for text classification because it can handle high-dimensional data and produce high accuracy (Suryati et al., 2023).

After the classification model is built, model evaluation needs to be conducted to determine the model's performance (Suryati et al., 2023). Evaluation uses the following accuracy formula:

$$Accuracy = \frac{The\ number\ of\ correctly\ classified\ products}{Total\ number\ of\ Classifications} \quad (1)$$

RESULT

In this section, the researcher will explain the results of the research obtained. Researchers can also use images, tables, and curves to explain the results of the study. These results should present the raw data or the results after applying the techniques outlined in the methods section. The results are simply results; they do not conclude.

The data collected in this research is a .csv format dataset containing 50,425 product description samples obtained through a scraping process from an e-commerce site by Gautam on Zenodo.org. This dataset consists of two columns, with the first column containing the class name and the second column containing the associated product and description. Moreover, the sample of the dataset is as follows:

Table 1. Collected Dataset Sample

Product	Product Description
Household	Paper Plane Design Framed Wall Hanging Motivational Office Decor Art Prints (8.7 X 8.7 inch) - Set of 4 Painting made up in synthetic frame with uv textured print which ...
Household	SAF 'Floral' Framed Painting (Wood, 30 inch x 10 inch, Special Effect UV Print Textured, SAO297) Painting made up in synthetic frame with UV textured print which gives...
Books	"Finding Ultra, Revised and Updated Edition: Rejecting Middle Age, Becoming One of the World's Fittest Men, and Discovering Myself Review ""Finding Ultra blends ...
...	...
Clothing & Accessories	Dupatta Bazaar Women's Dupatta Give a touch of charismatic splendour to your ethnic silhouette with this gorgeous dupatta in vibrant red hue. Adorned with golden...
Electronics	HP 245 G5 AMD A6 14-inch Laptop (4GB/500GB HDD/DOS/Black/2.76 kg) HP world leader in PCs helps equip you with a fully functional notebook ready to connect...
Electronics	Samsung Guru FM Plus (SM-B110E/D, Black) Colour:Black Compact Design If you are looking for a phone that is both simple and sturdy, checkout the Samsung Guru FM Plus...

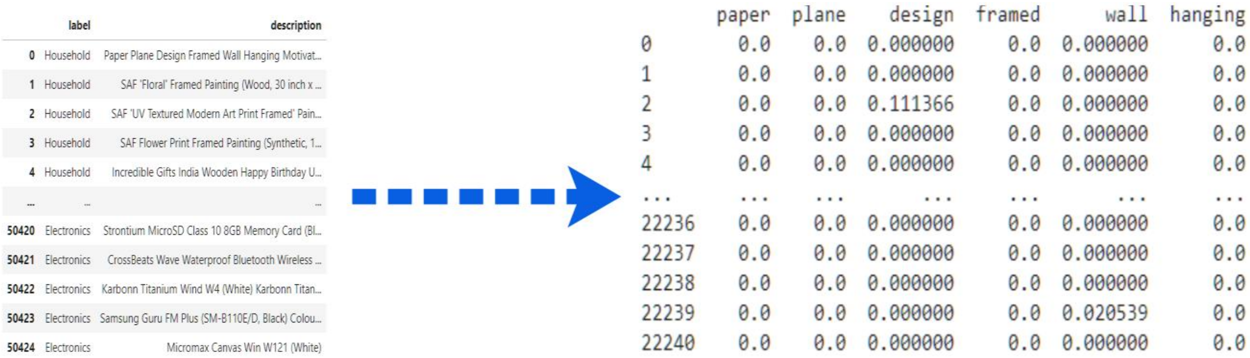
This stage involves loading the data used to train and test the model. The data can be loaded using libraries such as pandas or numpy. After conducting various trials, the researchers successfully produced a Data Preprocessing stages that could generate optimal accuracy:

Table 2. Preprocessing Result

No	Steps	Explanation

1.	Text Cleaning (Stopword Removal)	Removing common and less meaningful words (such as "the", "is", "and", etc.) from the text.
2.	Tokenization	Breaking down the text into smaller units or tokens or words.
3.	Part-of-Speech Tagging	Tagging each word in the text with its grammatical role (e.g., noun, verb, adjective, etc.) to assist in understanding sentence structure and analysis based on word types.
4.	Named Entity Recognition	Searching for and classifying named entities in the text, such as the names of people, organizations, locations, etc. This is crucial for analyzing the product to be classified.
5.	TF-IDF	Transforming the text into a numerical representation that can be understood and processed by the SVM algorithm. This process assigns higher weight to words that frequently appear in one document so that the SVM algorithm can identify the most relevant words for a document.

Furthermore, the appearance before and after Data Preprocessing can be seen as in Figure 1 below where the image on the left is before and the image on the right is after.



```
             label                                         description
0        Household     Paper Plane Design Framed Wall Hanging Motivat...
1        Household         SAF 'Floral' Framed Painting (Wood, 30 inch x ...
2        Household       SAF 'UV Textured Modern Art Print Framed' Pain...
3        Household        SAF Flower Print Framed Painting (Synthetic, 1...
4        Household       Incredible Gifts India Wooden Happy Birthday U...
...            ...                                                   ...
50420  Electronics     Strontium MicroSD Class 10 8GB Memory Card (Bl...
50421  Electronics      CrossBeats Wave Waterproof Bluetooth Wireless ...
50422  Electronics     Karbonn Titanium Wind W4 (White) Karbonn Titan...
50423  Electronics    Samsung Guru FM Plus (SM-B110E/D, Black) Colou...
50424  Electronics                      Micromax Canvas Win W121 (White)
```

```
       paper  plane    design  framed      wall  hanging
0        0.0    0.0  0.000000     0.0  0.000000      0.0
1        0.0    0.0  0.000000     0.0  0.000000      0.0
2        0.0    0.0  0.111366     0.0  0.000000      0.0
3        0.0    0.0  0.000000     0.0  0.000000      0.0
4        0.0    0.0  0.000000     0.0  0.000000      0.0
...      ...    ...       ...     ...       ...      ...
22236    0.0    0.0  0.000000     0.0  0.000000      0.0
22237    0.0    0.0  0.000000     0.0  0.000000      0.0
22238    0.0    0.0  0.000000     0.0  0.000000      0.0
22239    0.0    0.0  0.000000     0.0  0.020539      0.0
22240    0.0    0.0  0.000000     0.0  0.000000      0.0
```

Fig 3. Preprocessing Result

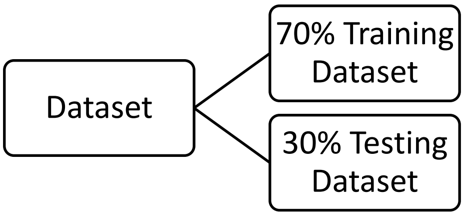


Fig 4. Data Split

As shown in the image above, the dataset was split into training and testing sets in this stage. The distribution was as follows: 80% of the data was designated for training purposes, while the remaining 20% was used for testing. It's important to note that this distribution is quite standard in machine learning research, as it allows for a robust assessment of model performance while also providing sufficient data for training.

After collecting the facial photo data, the next step is to design a classification model using the Inception V-3 architecture. Inception V-3 is a convolutional artificial neural network known for its ability to recognize complex features in images. First, the data will be trained with a dataset used for training, then a dataset used for testing will be used to produce prediction results, the accuracy of which will be calculated. The hyperparameter used in this research is as follows.

*saut.unpri@gmail.com

Table 3. Training Performance Result

Data Actual Predicted	**Predicted Book**	**Predicted Cloth & Accessories**	**Predicted Electronics**	**Predicted Household**
Actual Book	TP (1184)	FP (7)	FP (18)	FP (62)
Actual Cloth & Accessories	FP (4)	TP (1094)	FP (6)	FP (17)
Actual Electronics	FP (14)	FP (0)	TP (991)	FP (52)
Actual Household	FP (24)	FP (19)	FP (15)	TP (2054)

As shown in the table above, the TP (True Positive) and FP (False Positive) can be explained as follows:

1. TP (True Positive): True Positive refers to the scenario where the model correctly predicts the positive class. For instance, the model accurately predicts that a product is "Electronics" when the product is indeed "Electronics". So, looking at the table, we see that there are 4898 products that are actually "Electronics" and the model correctly classifies them as "Electronics".
2. FP (False Positive): False Positive refers to the scenario where the model incorrectly predicts the positive class. This means the model predicts that a product is "Electronics" but in reality, the product is not "Electronics". So, looking at the table, we see that there are 12 products that are actually not "Electronics" but the model wrongly classifies them as "Electronics".

This stage involves the process of evaluate the performance of the model using the test data. In this phase, the trained model is used to make predictions on the test data, and the results are compared with the actual values.

Table 4. Test Performance Result

Data Actual Predicted	**Predicted Book**	**Predicted Cloth & Accessories**	**Predicted Electronics**	**Predicted Household**
Actual Book	TP (1184)	FP (7)	FP (18)	FP (62)
Actual Cloth & Accessories	FP (4)	TP (1094)	FP (6)	FP (17)
Actual Electronics	FP (14)	FP (0)	TP (991)	FP (52)
Actual Household	FP (24)	FP (19)	FP (15)	TP (2054)

DISCUSSIONS

This study demonstrates how Natural Language Processing (NLP) methods, specifically TF-IDF and SVM, can be utilized to classify e-commerce product descriptions into relevant categories. Through the TF-IDF method, data is transformed into numerical vectors representing their frequency in the description and uniqueness across the dataset. Subsequently, the SVM model engineered in this study is capable of classifying these product descriptions into the appropriate categories with a high degree of accuracy, demonstrating the effectiveness of these research steps.

In this study, word vector representation through TF-IDF and data preprocessing was conducted through the following steps:

1. Text Cleaning (Stopword Removal): Eliminating insignificant common words.
2. Tokenization: Breaking down the text into smaller word units.
3. Part-of-Speech Tagging: Marking the grammatical role of words.
4. Named Entity Recognition: Identifying and classifying specific entities in the text.
5. TF-IDF: Converting text into a numerical representation, weighting words based on their frequency and uniqueness within the dataset.

In this study, a reasonably good accuracy was achieved in the training process, and knowing the performance results as shown in Table 3 when predicting the training data, it can be calculated with the following accuracy formula:

*saut.unpri@gmail.com

$$Training\ Accuracy = \frac{The\ number\ of\ correctly\ classified\ products}{Total\ number\ of\ Classifications} = \frac{(4898 + 4535 + 4177 + 8390)}{(22171\ )} = 99.2\% \quad (2)$$

Then, by observing the testing performance result as shown in the Table 4 the accuracy of the testing result is as follows:

$$Test\ Accuracy = \frac{The\ number\ of\ correctly\ classified\ products}{Total\ number\ of\ Classifications} = \frac{(1184 + 1094 + 991 + 2054)}{5561} = 95.7\% \quad (3)$$

These results demonstrate that the combined model of the TF-IDF method and the SVM machine learning algorithm in this study is effective for e-commerce product description classification tasks.

CONCLUSION

This research has successfully demonstrated how Natural Language Processing (NLP) methods, specifically TF-IDF and SVM, can be applied to classify e-commerce product descriptions into relevant categories. With an accuracy rate of 99.2% in the training phase and 95.7% in the testing phase, this method proves to be effective and reliable. Therefore, this study paves the way for enhancements in research related to product data management and organization on e-commerce platforms, thereby assisting users in finding the products they want. E-commerce platform companies can, as a result, provide better service to their users.

REFERENCES

Alhamra, A. R. (n.d.). *Rancang Bangun Aplikasi Pencarian Produk Sneakers Dengan Harga Terbaik Dari Beberapa E-commerce Di Indonesia Menggunakan Algoritma Boyer Moore Berbasis Website (Studi Kasus: Ourdailydose. net & Ncrsport. com)*. Fakultas Sains dan Teknologi UIN Syarif Hidayatullah Jakarta.

Amrulloh, A., & Adam, I. F. (2021). Sistem Pencarian Similaritas Judul Tugas Akhir Menggunakan Metode TF-IDF. *Jurnal CoreIT*, *7*(2).

Arifin, N., Enri, U., & Sulistiyowati, N. (2021). Penerapan Algoritma Support Vector Machine (SVM) dengan TF-IDF N-Gram untuk Text Classification. *STRING (Satuan Tulisan Riset Dan Inovasi Teknologi)*, *6*(2), 129–136.

Isnain, A. R., Sakti, A. I., Alita, D., & Marga, N. S. (2021). Sentimen Analisis Publik Terhadap Kebijakan Lockdown Pemerintah Jakarta Menggunakan Algoritma Svm. *Jurnal Data Mining Dan Sistem Informasi*, *2*(1), 31–37.

Kurniawan, S., Gata, W., Puspitawati, D. A., Parthama, I. K. S., Setiawan, H., & Hartini, S. (2020). Text Mining Pre-Processing Using Gata Framework and RapidMiner for Indonesian Sentiment Analysis. *IOP Conference Series: Materials Science and Engineering*, *835*(1), 012057.

Mutawalli, L., Zaen, M. T. A., & Bagye, W. (2019). Klasifikasi Teks Sosial Media Twitter Menggunakan Support Vector Machine (Studi Kasus Penusukan Wiranto). *Jurnal Informatika Dan Rekayasa Elektronik*, *2*(2), 43–51.

Naraloka, T. (2021). Pemanfaatan Teknologi Untuk Menganalisa Sentimen Masyarakatdalam Membantu Peningkatan Ekonomi Kreatif Di Era New Normal. *Proceeding Seminar Nasional Ilmu Komputer*, *1*(1), 121–137.

Nasution, E. Y., Hariani, P., Hasibuan, L. S., & Pradita, W. (2020). Perkembangan Transaksi Bisnis E-Commerce terhadap Pertumbuhan Ekonomi di Indonesia. *Jesya (Jurnal Ekonomi Dan Ekonomi Syariah)*, *3*(2), 506–519.

Prastyo, P. H., Ardiyanto, I., & Hidayat, R. (2020). Indonesian Sentiment Analysis: An Experimental Study of Four Kernel Functions on SVM Algorithm with TF-IDF. *2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI)*, 1–6.

Pratiwi, R. W., Dairoh, D., & Af'idah, D. I. (2021). Analisis Sentimen Pada Review Skincare Female Daily Menggunakan Metode Support Vector Machine (SVM). *Journal of Informatics Information System Software Engineering and Applications (INISTA)*, *4*(1), 40–46.

RAHMAWATI, Y., & others. (2019). Penerapan Strategi Pemasaran Website E-Commerce Untuk Meningkatkan Volume Penjualan Produk *(Studi Kasus Pada Website www. furnitureanakonline. com, www. ukirjepara. com dan www. jeparaheritage. id)*. UNISNU Jepara.

Ristiano, T. A. (2022). Implementasi Klasifikasi Teks Sms Spam Menggunakan Metode Svm Dengan Ekstraksi Fitur Fasttext.

Suryati, E., Styawati, S., & Aldino, A. A. (2023). Analisis Sentimen Transportasi Online Menggunakan Ekstraksi Fitur Model Word2vec Text Embedding Dan Algoritma Support Vector Machine (SVM). *Jurnal Teknologi Dan Sistem Informasi*, *4*(1), 96–106.

*saut.unpri@gmail.com