Entity Matching of Shop Accounts in Online Commerce Portals

D Salsabila[1], **Takdir**[1]

[1]Statistical Computing Department, Politeknik Statistika STIS, Jakarta, Indonesia

*Corresponding author's e-mail: 221709635@stis.ac.id

Abstract. Currently, online marketplace data are valuable data sources to be analyzed for various purposes. In the data collecting phases, duplication of shop accounts was found, resulting in biased analysis. This study examines the development of a mechanism to identify duplicate entities, i.e. store accounts, between different online marketplaces, or commonly known as entity matching. Word similarity algorithms were adopted as the core elements of our approach. Additionally, we present an entity matching model by examining logistic regression, naive Bayes, and random forest to find the best model for classifying store account similarities. Top online marketplaces in Indonesia are the object of our study, limited to one developing municipality, i.e. Sleman, DI Yogyakarta. The results show the best model has an accuracy value of 0.961, precision of 0.963, a recall of 0.958, and an F1-score of 0.962. Therefore, these results are acceptable for duplicate identification.

1. Introduction

The rapid development of information and communication technology has caused many changes, one of which is disseminating information using the internet. As noted in the *Asosiasi Penyelenggara Jasa Internet Indonesia* (APJII) results in surveys from the 2019 - second quarter/2020 [1], the number of users internet in Indonesia rose to 73.7% of the total population of 266.91 million Indonesians or reached 196.71 million users. This number increased by 25.5 million or 8.9% compared to 2018. The growth of the internet and information technology today certainly affects data collection, processing, and dissemination.

In addition to data collection, the internet, technology, and information also affect the economy. In the economic aspect, a new term was born, namely the digital economy. The digital economy is marked by the many changes in the sales system from offline to online. The concept of the digital economy was first introduced by Tapscott (1997), which is a social phenomenon that affects the economic system, where this phenomenon has characteristics as an intelligence space, including information, various access to information instruments, information capacity, and information processing [2]. The components of the digital economy that have been identified for the first time are the ICT industry, e-commerce activities, and the digital distribution of goods and services.

E-commerce is a part of the digital economy. According to the Organization for Economic Co-Operation and Development (OECD) 2011, e-commerce is the sale or purchase of goods/services carried out through a computer network with a method specifically designed to receive or place orders. Still, the principal payment and delivery of goods/services do not have to be made online [3]. As time goes by, with the increasing number of smartphone users and internet users, more and more people are starting to sell and shop via the internet. Sales and purchases of goods online can be done directly

between sellers and buyers through social media or websites and can use digital platforms as intermediaries between sellers and buyers. A digital platform called a marketplace is a location for buying and selling products where sellers and consumers meet on a digital marketplace/platform. In Indonesia, e-commerce is growing rapidly because 40 percent of the total population in Indonesia owning smartphones, which is consumers in Indonesia more frequently transact via smartphone applications [4]. As reported from Statista.com [57], top online e-commerce platforms in Indonesia, i.e. Shopee, Tokopedia, and Bukalapak.

A marketplace as a platform for buying and selling makes it easier for people to carry out economic activities practically and quickly. This platform makes it easy for consumers to get product information from various online stores, also choose and compare products easily and quickly. Meanwhile, from the seller's point of view, this platform makes it easy for the selling system to not think about sales strategies and others. With the various conveniences offered by marketplaces, the number of people selling online increases by registering their stores to different online marketplaces.

Online marketplace data can be adopted to support official statistics such as the number of shops selling online and sales turnover. One of the challenges in analyzing marketplace data is calculating the actual number of store accounts in the marketplace and sales turnover because from the author's search on various online stores, there are sellers who promote their products on more than one e-portal commerce. In line with the results of research by [6] who conducted market data analysis, it was found that 86 of the 120 shops interviewed or 71.67% had stores in other marketplaces. Based on the results of this study, it is concluded that one business can have more than one store in various marketplaces and cause data duplication. Therefore, we need an approach to reduce duplication of store accounts to improve the accuracy of marketplace data analysis.

This paper is aimed to find a mechanism to identify duplicate store accounts from different marketplaces. This research uses two variables that can determine the similarity of store accounts, i.e., the store's name and the product's name. In calculating the similarity of store accounts, *Levenshtein distance* algorithm is used for store name matching and *cosine similarity* for product name matching. In addition, this work compares classification methods in predicting store account similarity, i.e. *logistic regression*, *naïve Bayes*, and *random forest*. The results of this study have obtained a mechanism for identifying duplicate store accounts and a tool for forming a unique online store list based on entity matching so that a marketplace repository can be created for the compilation of e-commerce statistics.

There are many studies that do entity matching with text similarity algorithm. One of the text similarity algorithms used is the Levenshtein distance as done by [7], [8] and [9]. In addition, there is also a document similarity analysis using the Cosine Similarity approach [10], [11] and [12]. This study will combine the two text similarity algorithms, namely Levenshtein distance and cosine similarity. Levenshtein distance is used to measure the similarity of store names, while cosine similarity is used to measure the similarity of product names.

Based on the results of text similarity with Levenshtein distance and cosine similarity, the basis for classification is obtained to determine whether the store is the same store or a different store. This study compares 3 classification methods as has been done by previous research, namely Logistic Regression [13], Naïve Bayes [14], and Random Forest [15]. In this study, various new methods will be carried out by adopting various methods in previous studies to identify the similarity of store accounts in various marketplaces, as well as obtain a list of unique stores.

2. Methods

2.1. Scope of Research

This study focuses on entity matching of shop accounts by compiling a mechanism for duplication identification based on the similarity of store accounts between marketplaces on Bukalapak and Tokopedia in Sleman Regency, DI Yogyakarta. The variables used to identify duplication based on the similarity of the store accounts are the store's name and the product's name. The word similarity algorithm used is Levenshtein distance and cosine similarity. Next, a model is developed from the similarity of each variable by comparing various classification methods, i.e. logistic regression, naïve

Bayes, and random forest, so that the best model will be obtained to predict the similarity of store accounts between marketplaces. The results of the best classification model that have been selected will be used to form a marketplace repository of all marketplace data.

2.2. Source and Data Collecting Methods

This research uses marketplace data in Sleman Regency, which consists of store data and product data in February 2020. The marketplace data are obtained from yearly student's field projects, namely Praktik Kerja Lapangan (PKL), in Politeknik Statistika STIS Academic Year 2019/2020 that collected data by crawling from several marketplace websites. The structure of data can be seen in Table 1.

Table 1. Data Structure of a Marketplace.

Variable Name	Data Type	Description
Store Data		
storeid	int64	Store identity
storename	object	Store name
district	object	Store location (district/city)
Product Data		
storeid	int64	Store identity
prodid	int64	Product identity
prodname	object	Name of product

2.3. Analysis Methods

This experiment uses the *Jupyter Notebook* application with the *Python 3.8* programming language to compile a mechanism to identify duplicate store accounts between marketplaces. The research flow can be seen in Figure 1.

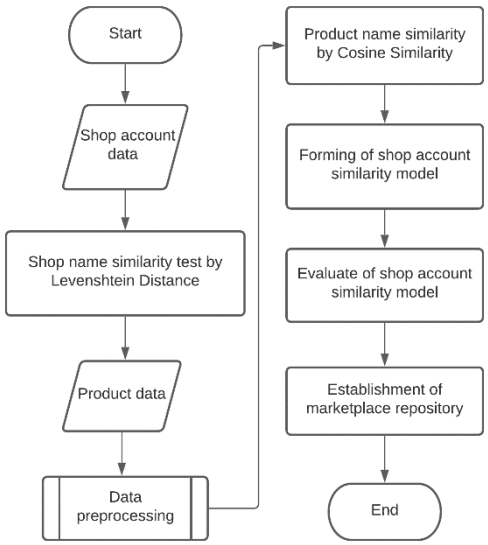

Figure 1. Research flowchart.

2.3.1. Data Preprocessing. Data preprocessing is a process where data is cleaned and prepared before analysis. This step is carried out on product data. The flow of data preprocessing in this study begins with case folding, replacing all letters with lowercase. Next, remove punctuation, non-ASCII encoding, and excess spaces. Then remove stopwords, namely deleting unnecessary words using the *Sastrawi* stopword remover module from Python. Finally, tokenization or the process of dividing the text into tokens. More detail of preprocessing marketplace data is discussed by Bustaman et al., 2020 [16].

2.3.2. Similarity Measurements. In calculating the similarity of store accounts based on store names, the Levenshtein distance algorithm is used. The Levenshtein distance algorithm is an algorithm that measures the similarity of words based on the number of operations performed to convert a word into another word, including changing, deleting, or adding a character [7]. This algorithm was chosen to measure the similarity of store accounts based on store names because store names tend to be short, so it is suitable to identify similarities with the Levenshtein distance algorithm. The calculation of Levenshtein distance can be seen in the following formula.

$$f(i,j) = \min \begin{cases} f(i-1,j)+1 & ,deletion \\ f(i,j-1)+1 & ,addition \\ f(i-1,j-1)+1 & ,replacement \end{cases} \quad (1)$$

After obtaining the distance value between two strings, the similarity calculation is carried out with the following formula.

$$Similarity\ weight = 1 - \frac{dist}{\max(S,T)} \quad (2)$$

Information:
dist: Levenshtein distance value between strings 1 and 2
max(S, T): String length largest between string 1 and 2

The similarity weight is assumed to be in the range 0-1, where the closer to 1, the more similar the two strings are. Meanwhile, closer to 0 means the two strings are increasingly dissimilar [9].

The identification of store account similarities based on product names is carried out using the cosine similarity algorithm. The Cosine similarity algorithm measures the similarity between texts by assuming a text to be a vector. The advantage of this algorithm is that it is not affected by the short length of a document and has a high level of accuracy [10]. The calculation of cosine similarity can be seen in the following formula.

$$Cosine\ similarity = \frac{A.B}{||A||||B||} = \frac{\sum_{i=1}^{n} A_i\ x\ B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2 x \sum_{i=1}^{n}(B_i)^2}} \quad (3)$$

Information:
A : vector A to compare the similarity
B : vector B to be compared the similarity
Ai : weight term i in block Ai
Bi : weight term i in block Bi
i : number of terms in a sentence

2.3.3. Similarity Modeling. The modeling to identify similarity of store accounts was carried out using three classification methods, namely logistic regression, naïve Bayes, and random forest. This comparison of classification methods aims to see which model gives the best results in predicting the similarity of store accounts. The independent variable used results from the similarity of the shop name and the similarity of the product name. The dependent variable used is store account similarities, with 0 being a different store and 1 being the same store.

The logistic regression used is binary logistic regression, a data analysis model to find the relationship or tendency of each independent variable X with a categorical or interval scale to the response variable Y, which is binary [17]. Logistic regression was chosen because the algorithm is simple to implement, effective, and can classify any problem, preferably binary data [18].

Naïve Bayes is a classification algorithm that calculates a set of Bayesian theorem probabilities by adding up the frequency and combination of values from a given dataset [19]. Naïve Bayes was chosen because the algorithm gives real-time predictions, scalable with large datasets, insensitive to irrelevant features, and can be used in text classification or predict on binary/multiple classes [18].

Random forest is a classification model carried out by developing several decision trees based on the random selection of data and variables. The resulting model is a selected sound model from all

trees [20]. This algorithm was chosen to reduce error, handle a massive amount of data, and perform well on imbalanced datasets [18].

The three classification models are run on data with k-fold cross-validation with k = 10, dividing the data into several k subsets. One of the subsets is used as data testing, and the remaining k-1 subsets are combined to form the training data. This classification will then be compared based on performance measures, i.e. accuracy, recall, precision, and f1-score. Accuracy is the standard and simple parameter to evaluate the performance of a classification algorithm by showing what level or percentage of prediction truth is. Recall, precision and f1-score are often used in information retrieval, where recall is the level of sensitivity to the relevant part of the data, precision is the accuracy of the prediction results, while f1-score is the average harmony of precision and recall [15]. The classification results from this best model will be used to form a marketplace repository to prepare e-commerce statistics.

3. Results

3.1. Text Similarity Measurement Results

This research is important to do, especially in the application of marketplace data analysis. By identifying duplicate store accounts and generating a unique store list, marketplace data analysis can have even better accuracy. The first step in this work is to calculate the similarity of store accounts based on store names with the Levenshtein distance algorithm. Illustration of Levenshtein distance can be seen in Table 2.

Table 2. Illustration of Levenshtein Distance Algorithm.

Storename_1	Storename_2	Levenshtein distance
onlineshop	onlineshop	0
onlineshop2	onlineshop	1
online.shop2	onlineshop	2
online.shop23	onlineshop	3

This study only takes a list of stores with different store names with a Levenshtein distance value limited to a distance from 0 (zero) to 2 (two). From 3682 store accounts in the first marketplace and 57411 store accounts in the second marketplace, 1553 store accounts had the same store name with a maximum distance of two characters. Next, the Levenshtein distance value is transformed into similarity weights in formula (2).

The second step is to do manual labeling for the similarity status of store accounts. Manual labeling aims to mark stores that are considered the same or different stores by checking store accounts on both marketplace websites. Things to consider at this stage are the name of the store, storefront list, store description, product type, and visual designs such as store logos and store banners. The results of the manual labeling found that 759 store accounts (49.87%) are different stores labeled as 0, and 794 store accounts (51.13%) are the same stores and marked as 1.

The next step is to calculate the cosine similarity for the product name which aims to see the similarity of the store from the name of the product being sold. First, we need to combine the dataset between the product data and the data from the Levenshtein distance calculation which was carried out in the first stage. The results of calculating the similarity of product names with cosine similarity found that from 794 same stores there were 55.92% stores with a similarity more than 0.5. Meanwhile, from 759 different stores, there are 99.21% of stores have the same product name under 0.5.

3.2. Similarity Modeling Results

In compiling the modeling of store account similarity variables, namely store name, and product name, a comparison of three classification models was carried out: logistic regression, naïve Bayes, and random forest. The following are the results of a comparison of classifications based on performance measures.

Table 3. Comparison of classification results based on the performance measure.

Algorithm	Performance Measure			
	Precision	Recall	F1-scores	Accuracy
Logistics regression	0.976	0.927	0.951	0.952
Naive Bayes	0.986	0.891	0.935	0.938
Random forest	0.963	0.958	0.962	0.961

Based on the results of the performance measure in Table 3, it can be seen that random forest is the best classification model for classifying the similarity status of store accounts with an accuracy value of 0.961, an F1-score of 0.962, a recall of 0.958, and a precision of 0.963. In addition, it can also be seen in the results of the confusion matrix in Table 4 below.

Table 4. Confusion matrix on the k-fold cross-validation model.

Algorithm	Confusion Matrix	
Logistics regression	741	18
	57	737
Naive Bayes	749	10
	86	708
Random forest	731	28
	35	759

From the table above, it can be seen that the number of shops that are misclassified is the smallest with the random forest model, followed by logistic regression and Naive Bayes. Based on these results, we choose a random forest model as the best model to classify store account similarities. Furthermore, the establishment of a marketplace repository is based on the classification results that have been obtained. The following is a summary of the establishment of a marketplace repository in Sleman regency.

Table 5. The results of the establishment of a marketplace repository.

Marketplace	Raw data	Same shop account	Different shop accounts
1	3682	787	2895
2	57411		56624
Total	61093	60306	

Table 5 shows that the actual number of stores from two marketplaces in Sleman regency is 60306 from the total raw data of 61093 store accounts. The result is based on the assumption that the similarity of store accounts is identified by the similarity of store names with a maximum difference of two characters and similarity of product names through the analysis step. The results of this marketplace repository can be used for the preparation of e-commerce statistics.

4. Conclusion

Based on our experiment, two variables can be used to identify duplication based on the similarity of shop accounts between different marketplaces, namely, store names and product names. Matching results based on store names, it was found that from 1553 store accounts with a maximum Levenshtein distance value of 2, 51.13% were the same store. The matching results based on the product name with

the cosine similarity algorithm have a similarity value greater than 0.5 of 28.98%. After compiling the model with logistic regression, naïve Bayes, and random forest, it was found that the random forest was the best model to predict the similarity of store account. The best model has the highest accuracy value of 0.961, precision of 0.963, recall of 0.958, and F1-score of 0.962. Furthermore, from the classification results with random forest model, a marketplace repository is formed, which contains a list of unique stores to prepare e-commerce statistics.

Acknowledgments
The authors would like to thank Politeknik Statistika STIS, PKL 59, and the Directorate of Analysis and Statistics Development BPS Statistics Indonesia for their support regarding this research.

References

[1] APJII, *Survei Internet APJII 2019-2020 [Q2]*. https://www.apjii.or.id. [accessed November 2020]
[2] Tapscott D 1997 *The Digital Economy: Promise and Peril in the Age of Networked Intelligence* (New York: McGraw-Hill Inc)
[3] OECD 2011 *OECD Guide to Measuring the Information Society 2011* (Paris: OECD Publishing)
[4] Pramana S, Mariyah S and Takdir 2021 *Stat. J. of the IAOS.* **37** pp 415-27
[5] Statista, *Top 10 e-commerce sites in Indonesia as of 4th quarter 2020 by monthly traffic.* https://www.statista.com/statistics/869700/indonesia-top-10-e-commerce-sites [accessed November 2020]
[6] PKL STIS 2020 *Pemanfaatan BIG DATA dalam Mengetahui Pertumbuhan Jumlah Akun Toko, Jumlah Barang Terjual dan Besaran Omzet Penjualan Marketplace* (Jakarta: Politeknik Statistika STIS) p 50
[7] Ariyani N H, Sutardi and Ramadhan R 2016 *J. SemanTIK*. **2** pp 279-86
[8] Fauzan R, Riadi J and Sholihin F 2018 *Proc. SNRT* pp 1-6
[9] Pratama B and Pamungkas S 2016 *J. LOG!K@*. **6** pp 131-43
[10] Riyani A, Naf'an M Z and Burhanuddin A 2019 *J. Ling. Komp*. **2** pp 23-27
[11] Habibi M and Sumarsono 2018 *JISKa*. **3** pp 110-18
[12] Iriananda S W, Muslim M A and Dachlan H S 2018 *MATICS J. Ilm. Komp. dan Tek. Inf*. **10** pp 30-38
[13] Nishadi A S T 2019 *Int. J. of Advanced Research and Publications*. **3** pp 69-74
[14] Syaputri A W, Irwandi E and Mustakim 2020 *J. Int. Comp. & He Inf*. **1** pp 15-19
[15] Primajaya A and Sari B N 2018 *Indonesian J. of Artificial Intelligence and Data Mining*. **1** pp 27-31
[16] Bustaman U, Larasati D N, Putri Z H S, Mariyah S, Takdir, Pramana S 2020 *Proc. on 10th Int. Conf. on Information Technology and Electrical Engineering* pp 186-91
[17] Hosmer D W and Lemeshow S 2000 *Applied Logistic Regression 2nd Edition* (USA: John Wiley & Sons Inc)
[18] Gupta S, *Pros and Cons of Various Machine Learning Algorithms.* https://towardsdatascience.com/pros-and-cons-of-various-classification-ml-algorithms-3b5bfb3c87d6 [accessed August 2021]
[19] Wibisono A B and Fahrurozi A 2019 *J. Ilm. Tek. dan Rek*. **24** pp 161-70
[20] Manthovani A N 2018 *Analisis Perbandingan Klasifikasi Metode Regresi Logistik Biner dan Random Forest pada Big Data* (Yogyakarta: UII)