Detection System of Promotion Abuse Using Similarity and Risk Scoring Methods

Cut Fiarni[1], Arief Samuel Gunawan[2], Ishak Anthony[3]

***Abstract*—Offering promotion coupons is one of the most popular strategies of online marketing to attract new customers and increase customer loyalty. However, this strategy opens chances for fraud risk as the coupons are being redeemed multiple times using fake accounts. This risk becomes a burden to marketing costs and leads to failure to accomplish the intended strategic value. Therefore, this research focuses on building an automatic detection system of online promotion abuse based on its risk level. The proposed system also must work on live stream and bulk data. Therefore, in live stream data, it could alert the administrator before the transaction finished or the next process started. After conducting an exploratory factor analysis of the 24 attributes collected from four tables of data transaction, there were seven attributes indicating promotion abuse. These attributes were the user IP address, shipping address, mobile number, member email, order email, payment ID, and product name. Then, supervised machine learning of similarity algorithms was used to build models and find the hidden correlation of attributes to indicate the promotion abuse. The result from comparing five similarity methods showed that based on the workflow and performance, the most suitable methods for this case were exact match and Levenshtein edit base. The automatic risk scoring feature of the proposed system used seven attributes of online transactions as their most prominent promotion abuse parameter based on its hidden correlation. From the system performance testing, the result values of precision, recall, and F-measure are 95%, 93%, and 0.94, respectively. These results indicate that the system performance is satisfactory.**

***Keywords*—Detection System, Promotion Abuse, Levenshtein Edit Base, Exact Similarity, Risk Scoring, Exploratory Factor Analysis, E-Commerce.**

I. INTRODUCTION

During this pandemic, the use of e-commerce platforms and other digital instruments is significantly increasing. It is in line with many training programs from the Indonesian government that helps encouraging micro, small, and medium enterprises (MSMEs) to survive and develop following the e-commerce trend. Increased by 29.6% (YoY) in 2020, the nominal growth rate of Indonesian e-commerce transactions is evidence of the escalating e-commerce trend [1]. It is supported by an increase in public preference for the use of digital platforms that prevent crowds and a strategy of many e-marketplaces that offer various promotions. The use of promotional strategies aims to introduce new products, attract new customers, and maintain customer loyalty. One of the promotional methods applied is issuing an applicable promotional code to get a discount. This promotional code usually consists of unique numbers and letters related to the promotional item brand, which will generally be applied during checkout.

In using promotional codes, e-commerce must carry out transaction security processes to reduce the inherent risk. The procedure involves authenticating the coupon and making sure it has not been changed, has not been traded, and is still valid. However, in this marketing strategy implementation, risks of promotion abuse by irresponsible persons (fraudsters) persist. One of the promotion abuse modes is that fraudsters take advantage of loopholes in the promotional code mechanism for new customers, which can only be used once. In their operation, fraudsters will fake accounts to get promotional prices. This act is known as promotion abuse, which is a part of online fraud [2]. Generally, promotion abuse is conducted to gain additional benefits from price discounts for sales. The modes used by fraudsters are very diverse and constantly growing along with the development of the existing technology. These modes range from hacking to scrapping and social engineering. Companies commonly rely on fraud analyst experts to examine any transactions with an indication of promotion abuse based on records of transactions that have been done. Afterwards, the results are sent to the sales department for further evaluation. However, the drawback of this procedure lies in its weakness in preventing the risk of fraud to happen.

Research related to online fraud generally focuses on banking transactions; at the same time, many MSMEs have started to build their online business. For this reason, research on promotion abuse is very essential. In addition, MSMEs generally do not have sufficient financial resources to hire fraud analyst experts. Hence, this study focuses on modeling the fraud detection analysis process based on how the experts analyze and conduct data-driven research in utilizing machine learning to find attributes indicating frauds. The resulting model will be used in the automatic fraud detection system to prevent fraudulent transactions on the promotional code usage. The transaction data used in this study were obtained from XYZ e-marketplace, which is one of the top 10 e-commerce companies with the best performance in Indonesia in 2020 [3].

II. FRAUD DETECTION PROBLEMS OF ONLINE TRANSACTION

Automated fraud detection of online transactions becomes a primary study in this e-commerce era. Online transaction records are getting more extensive in terms of volume, speed, and variety of attributes, which is in conjunction with the shift

[1,3] *Department of Information System, Harapan Bangsa Insitute of Technology, Jl. Dipatiukur 80–84, Bandung, Indonesia (tel./fax: 022-2506636; email: cut.fiarni@ithb.ac.id, ishakantonny@yahoo.com)*

[2] *Industrial Systems Engineering and Product Design, Ghent University, Technologiepark-Zwijnaarde 46, 9052 Gent, Belgium; (email: ariefsamuel.gunawan@ugent.be)*

[Received: 1 January 2022, Revised: 8 March 2022]

in people's habits that favor making online payments. This big data could give valuable knowledge on user transactions, which is not only for segmenting personalized target markets, but also for mapping the correlation between transaction fraud and user behavior. It then encourages the implementation of machine learning in the security and risk management of e-commerce transactions research. The state-of-the-art research in this field basically aims to identify suspicious patterns of the transaction record analysis [4]. In this field, research on promotion abuse in e-commerce is still limited as it mainly focuses on banking transactions and payment mechanisms with credit cards as the subject. For instance, [5] focused on improving the efficiency and stability of fraud detection platforms for credit cards by utilizing deep neural networks. Meanwhile, [6] made improvements to the mobile coupons system architecture using the QR code cryptography technique. The data-driven modeling used classification which used machine learning could handle a massive volume of imbalanced data. Reference [7] compared machine learning algorithms for detecting credit card frauds in e-commerce. The result showed that the highest accuracy value was the neural network with a value of 96%, while both random forest and naïve Bayes both yielded an accuracy of 95%. On the other hand, the decision tree generated the lowest value of 91%. Reference [8] applied the support vector machine (SVM) classifier to detect fraud in financial statements based on ratio analysis. Meanwhile, [9] conducted detection research on organized e-commerce fraud using scalable categorical clustering for bulk datasets. This study succeeded in detecting 26.2% fraud and only caused 0.1% false alarms from legal transactions.

This study attempts to build a promotion abuse detection system by combining expert processes to analyze fraud and using machine learning methods to model pattern learning of transaction data. To reduce promotion abuse, preventive measures are essentially needed to automatically categorize transactions in the system as promotion abuse or not. Hence, they do not cause bottlenecks in the online transaction cycle which eventually can affect the performance and reputation of e-commerce companies. For this reason, the focus of this research is to develop a promotion abuse detection system which can perform tagging to prevent the occurrence of such fraud. The model of this proposed system used data-driven machine learning to match the users' profiles and characters committing promotion abuse in B2C and C2C retail e-commerce.

III. Promotion Abuse Detection

This section discusses the scheme and the flow of methods to develop the proposed system.

A. Research Framework

The system that will be made aims to analyze the value of transaction risk to show indications of promotion abuse based on the similarity of the transaction attributes that produce the score. Based on Fig. 1, the research scheme starts from inputting transaction data records from the XYZ e-marketplace's point of sales (POS) module. Following that,

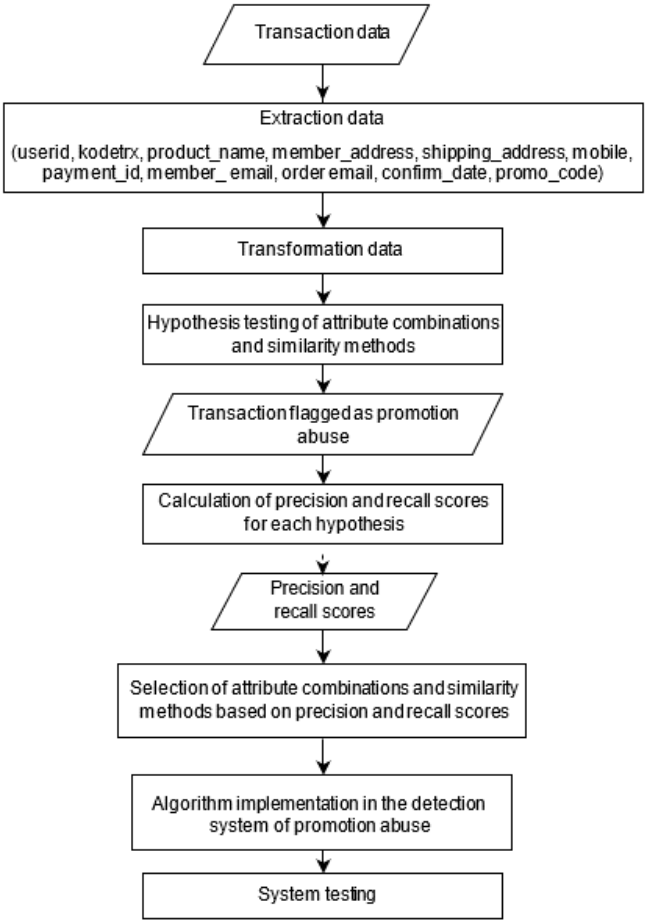

Fig. 1 Research framework of promotion abuse detection model.

TABLE I
DATA TYPE FROM THE ANNUAL TRANSACTION REPORT

Type	Data
Characters	Name, mobile, address, ship-to address, member (email), order email, payment ID
Unique character	User ID
Category	Status, payment method, bank, sub payment method, channel grouping, month, order date
Numeric	Total amount, discount

transaction data that becomes input will undergo the following process.

1) Data Extraction and Transformation: It acquires transaction data with the required attributes to define the promotion abuse detection model. After the transaction data with selected attributes are taken, data transformation will be carried out to increase detection accuracy and equalize the data format.

2) Exploratory Factor Analysis: The purpose of this exploratory factor test is to identify transactions that are considered as promotion abuse according to the combination of attributes and similarity methods used.

3) Precision and Recall Score Calculation: The results of the attribute combination test and the similarity method are

used for flagging promotion abuse in each transaction. Based on these results, precision and recall values will be calculated and are compared to manual reviews from fraud experts.

4) Best Algorithm Combination Implementation: From the results of precision and recall, the best values will be chosen. Then, the selected combination will be the algorithm used in the proposed system.

5) System Testing: After the detection system is developed, it will be tested using data processing and new datasets to check the analysis consistency and the comparison with manual reviews by experts.

Based on the working principle, this research is a knowledge-based agent since the machine learning algorithm generated the solution state based on transaction data used as training data in the modeling. The resulting models is not only the exploratory data analysis (EDA) result to get the attributes of transaction data that were most correlated with the fraud case, but also the similarity algorithm with the best performance. The need to understand business processes and the functionality of e-commerce system modules and types of data attributes has caused each data-driven machine learning research to be unique, and the attribute parameters used in the proposed solution system needs to be regularly evaluated and tested [10]. Hence, the modeling consisted of two stages. The first stage aimed to obtain attributes that could indicate promotion abuse based on transaction data analyzed and indicated by the experts as promotion abuse. The second stage aimed to seek the most appropriate machine learning algorithm for the proposed system. This attribute was identified by finding correlations and connections between transactions. By applying these methods, indications of fraud could be detected based on their patterns and attributes.

B. Data Exploration

As explained in the previous section, this study used the XYZ e-marketplace annual transaction report. Table I shows the grouping types of the annual transaction reports. The data transformation stage was applied on the member address and shipping address attributes. Each word in the sentence was replaced with its synonym using a data dictionary. After data transformation, promotion code sorting was carried out, where transactions with single-use promotional codes were removed from the data set after undergoing the exploration, preprocessing, and cleaning stages of the annual transaction report. The result was a transaction dataset including all the attributes needed for the analysis of transaction data as well as data record that was manually labeled as promotion abuse with the manual_flag column.

C. Attribute Selection Process

There are four types of data attributes, namely nominal, ordinal, interval, and ratio. Ordinal and nominal attribute types are categorical attribute types used for qualitative data or variable labeling. Meanwhile, interval and ratio are quantitative attributes which values can be sorted. Based on the analysis and evaluation of the ongoing system, there were 24 attributes derived from four data tables, as shown in Table II.

TABLE II
ATTRIBUTE AND DATA SOURCE

Data Table	Attribute	
User	● Full name ● Mobile ● Member email	● Address ● User ID
Product	● Part ID ● Brand ● Merchant name	● Series ● Price (IDR)
Transaction	● Shipping address ● Order email ● Order date ● Promo code ● SO Number	● Code trx ● Quantity ● Total amount ● Discount
Payment	● Payment ID ● Payment method ● Sub payment method	● Confirm date ● Bank name

TABLE III
WORKFLOW OF SIMILARITY ALGORITHM

Algorithm	Workflow
Hamming distance [11]	It seeks for similarities with the fastest execution time compared to the other four similarity-searching methods, but the measured length strings must be the same.
Levenshtein distance [12]	It seeks a proper match for strings with many typos occurring.
Longest common substring (LCS) [13]	It seeks for similarities by eliminating uncommon characters between two strings.
Jaro-Wrinkle distance [13]	It examines similarity for strings with short attributes, such as names.
Exact match [13]	It is the fastest method to find the similarity between two strings by sequentially looking for the similarity between each character in the two strings compared. The result of the comparison is 0 and 1, so it cannot be used to find similarities.

Based on business analysis, three attributes could not be used as a reference to detect promotion abuse due to their uniqueness and unsuitableness to be grouped. Additionally, they were not classified as transaction identifiers, but information complementary attributes. These attributes were user ID, payment ID, and full name. The part ID and brand attributes were represented by the series attribute, which was the full name of the product purchased. Part ID and brand were the explanation of the series, so that if there were similarities in the series, the part ID and brand will also be the same. If this attribute was the determinant of the detection, the series would have three times more determinants than other attributes. For the same reason as previous points, the payment ID attribute represented the payment method, bank name, and sub payment method attributes. Meanwhile, the full name attribute could not be used as a reference because there was a high possibility of username similarities, so that it would trigger a significant false positive (FP). As for the attribute selection process, there were seven attributes related to promotion abuse, namely mobile, address, member email, series, shipping address and order email, product name, and the payment ID. For writing simplification, these seven attributes were encoded into A1-A7.

TABLE IV
CODIFICATION OF CHARACTERISTIC DATA TRANSACTION OF E-COMMERCE

Code	Characteristic
K1	There are categorical attributes that are input from the system.
K2	There are many indications of promotion abuse in transactions caused by the detection of sequential patterns in naming email addresses and phone numbers.
K3	There are many typos in the user address and shipping address columns.
K4	There are random words in the sentence for the address attribute that refer to the same address.
K5	Some attributes only need to be checked for similarities and dissimilarities, such as payment ID, which should be unique for each customer who makes payment so that even one different character will be considered dissimilar.

TABLE V
COMPARISON OF SIMILARITY METHODS TO TRANSACTION DATA

Method	Code				
	K1	*K2*	*K3*	*K4*	*K5*
Hamming	✓				
Levenshtein	✓	✓	✓		
LCS	✓	✓		✓	
Jaro-Wrinkle	✓				

D. Selection of Algorithm of Promotion Abuse Detection

In determining the appropriate similarity method, it was necessary to compare the method to the characteristics of the proposed system. Table III shows the popular algorithms for finding similarity values and explains how the similarity method works. Based on the analysis of business needs, the similarity method applied must compare each attribute between transactions.

The transaction data has several characteristics as described in Table IV. The analysis was conducted by combining business needs taken from fraud risk detection, similarity methods, and e-commerce characteristics based on transaction data. The results of the analysis are shown in Table V. As explained in the previous section, this study focused on seven attributes (which were encoded into A1-A7) related to online promotion abuse. The first dataset used 82 transactions, and they were doubled for the second dataset. In the test hypothesis, the first two digits represented the chosen similarity method. The next digit signified the selected attribute, for example, "S2-A1A3A6", meaning that exploratory factor test used mobile, member email, order email attributes and the Levenshtein similarity method. There are three algorithms which can determine the similarity according to the characteristics and the needs of fraud risk related to transactions using promotion code [14].

1) Exact Match (Code: S1): Using (1), it checks the similarity between two strings by producing a value of 1 or 0 based on whether the two strings are equal.

$$sim(A,B) = \ if\ A = B, 1\ otherwise, 0 \tag{1}$$

2) Levensthein Distance Similarity (Code: S2): It checks the similarity between two strings by measuring the number of steps needed so that one string is the same as the other string. It was carried out using three steps, namely addition, replacement, and subtraction for each character with the following formula:

$$\begin{aligned} lev_{a,b}(i,j) = \ & \{max\ (i,j)\ min\ \{lev_{a,b}(i-1,j) + \\ & 1\ lev_{a,b}(i,j-1) + 1\ lev_{a,b}(i-1,j-1) + \\ & 1_{(a_i \neq b_j)} \\ lev\ \ a,b(i,j) = \ & \{(i,j)\ \ min\{lev\ a,b(i-1,j) + \\ & 1\ lev\ a,b\ (i,j-1) + 1\ \ lev\ a,b\ (i-1,j- \\ & 1) + \quad 1\ (ai \neq bj) \end{aligned} \tag{2}$$

3) Longest Common Substring (Code: S3): Using (3), it checks the similarity between two strings by looking at the longest substring that is the same between the two strings.

$$LCSubstr(A,B) = \ max\ 1 \leq i \leq m, \\ 1 \leq j \leq n\ \ LCSuff(A1mi, B1nj) \tag{3}$$

where,
A = first string,
B = second string,
i = number of characters in the first string,
j = number of characters in second string.

E. Hypothesis Building from Combinations of Attributes and Similarity Methods

The next phase was testing the exploratory factor from the attribute's combination and the similarity methods. The hypothesis testing in this stage was aimed to obtain the best model based on the performance values and was conducted by developing functions using the PHP programming language. This stage was begun by selecting the test hypothesis. For example, for testing hypothesis 1 (H1), H1 used the exact match method and the member address attribute combination. This function checked the similarity using the exact match method. The exact match function accepted two parameters stored as "$first" and "$second" variables, then the values were checked. The function returned a value of 0 when there was no similarity between the two variables. It was possible because the payment ID attribute was often null. Similar steps worked on the Levenshtein function. This function accepted two parameters which were then stored in two variables prior to conduct preprocessing. After that, the distance between the two variables with the default Levenshtein function was checked. In this function, the distance of two variables was divided by the longest variable to convert the distance into a decimal value between 0 to 1. For instance, "Bandung" and "Surabaya" variables each has seven and eight characters. In this example, the divisor is 8 since "Surabaya" has longer characters. Table VI shows the test results of the 35 combinations using two different datasets.

F. Result

At this phase, each hypothesis was tested based on their precision, recall, and F-measure values for each combination by following these steps [15], [16]. First, matching process

TABLE VI
RESULTS OF EXPLORATORY FACTOR COMPARISON OF ATTRIBUTES COMBINATION

Method	Attributes	First Dataset			Second Dataset		
		P (%)	*R* (%)	*F1*	*P* (%)	*R* (%)	*F1*
LV	A2A5A1A3A6A7A5	100	100	1.0	91	73	0.81
LV	A2	100	100	1.0	90	63	0.74
LV	A2A5	100	100	1.0	90	63	0.74
LV	A5	100	100	1.0	89	59	0.71
LV	A2A1A4	100	100	1.0	89	59	0.71
LV	A2A5A4	100	100	1.0	89	59	0.71
LV	A2A5A1A3A4	100	100	1.0	89	59	0.71
LCS	A5A4	100	100	1.0	92	54	0.68
LV	A5A1A5	100	100	1.0	88	51	0.65
LV	A2A5A3A6	100	100	1.0	87	49	0.63
LV	A2A5A1A7	100	100	1.0	87	49	0.63
LV	A2A5A1A6A4	100	100	1.0	87	49	0.63
LV	A2A5A1A3A6A5	100	100	1.0	87	49	0.63
LV	A2A5A1A3A6	100	100	1.0	86	46	0.60
LV	A2A5A1A6A7	100	100	1.0	86	46	0.60
LV	A2A5A6A4	100	80	0.9	86	46	0.60
LV	A5A1A6A4	100	80	0.9	86	46	0.60
LV	A2A1A3A6A4	100	80	0.9	86	46	0.60
LV	A2A5A3	100	100	1.0	86	44	0.58
LV	A2A5A6	100	100	1.0	86	44	0.58
LV	A2A5A1A3A7	100	100	1.0	86	44	0.58
LV	A2A1A3A4	100	80	0.9	85	41	0.55
LV	A2A1A6A4	100	80	0.9	85	41	0.55
LV	A2A3	100	100	1.0	84	39	0.53
LV	A2A6	100	100	1.0	84	39	0.53
LV	A5A3	100	100	1.0	84	39	0.53
LV	A5A6	100	100	1.0	84	39	0.53
LV	A2A5A1A4	100	100	1.0	84	39	0.53
LV	A2A1A3A6A4	100	80	0.9	84	39	0.53
LV	A2A5A1A3A4A7	100	80	0.9	84	39	0.53
LV	A2A3A6	100	100	1.0	83	37	0.51
LV	A2A5A3A4	100	80	0.9	83	37	0.51
LV	A5A1A3A4	100	80	0.9	83	37	0.51
LV	A5A3A6	100	100	1.0	82	34	0.48
LV	A2A5A3A6A4	100	80	0.9	79	27	0.40

Note: LV = Levenshtein

between the results of algorithms was carried out using the calculation model method. Then, the results were calculated to determine the values of true positive (TN), FP, true negative (TN), and false negative (FN). Last, the values of precision,

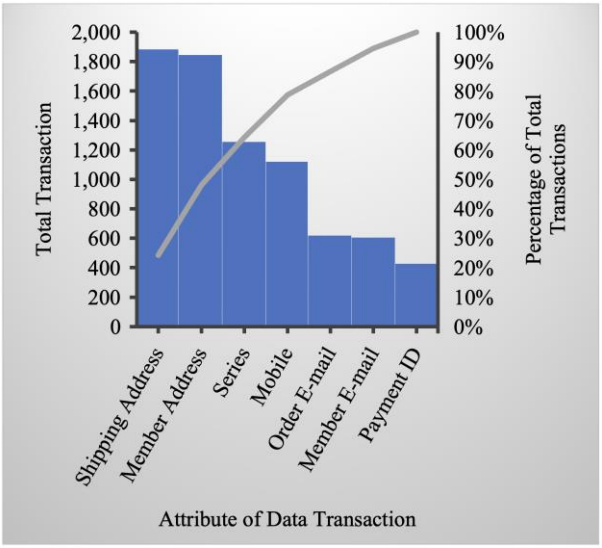

Fig. 2 Pareto chart of attributes of data transaction indicating promotion abuse.

TABLE VII
ATTRIBUTES AND SCORINGS

Attribute	Similarity Count	Percentage	Weight
Payment ID	427	5.45%	55
Member address	1,846	23.57%	236
Shipping address	1,885	24.06%	241
Mobile	1,120	14.30%	143
Member email	606	7.74%	77
Order email	619	7.90%	79
Series	1,330	16.98%	170
Total		100.00%	1,000

recall, accuracy, and F-measure results were calculated and compared.

Table VI shows the results of the pathfinder test of the similarity method. From the results of the exploratory factor test on each factor hypothesis, the combination of methods and attributes would produce a flag which was then compared with manual flag and produced TP, FP, TN, or FN values. The value obtained became the basis of the accuracy and recall calculation for each exploratory factor test [16]. Subsequently, the factor with a F-measure value greater than 0.8 was reported as meeting the hypothesis criterion.

The comparison of the result of hypothesis performance values showed that the best combination of all 35 hypotheses is the Levenshtein function with attributes A2A5A1A3A6A7A5. It indicates that it will generate the best promotion abuse detection model since it has the highest F-measure value. The application of the Levenshtein algorithm to the system model used the edit base distance. It was applied to check the similarity between two strings by comparing the number of edit distances using adding, changing, and subtracting characters in the first string. The results of the edit distance obtained were divided by the number of longest characters between two strings. The difference of the results and 1 was calculated to obtain the similarity percentage [12]. The combination of these attributes included member address, shipping address, mobile

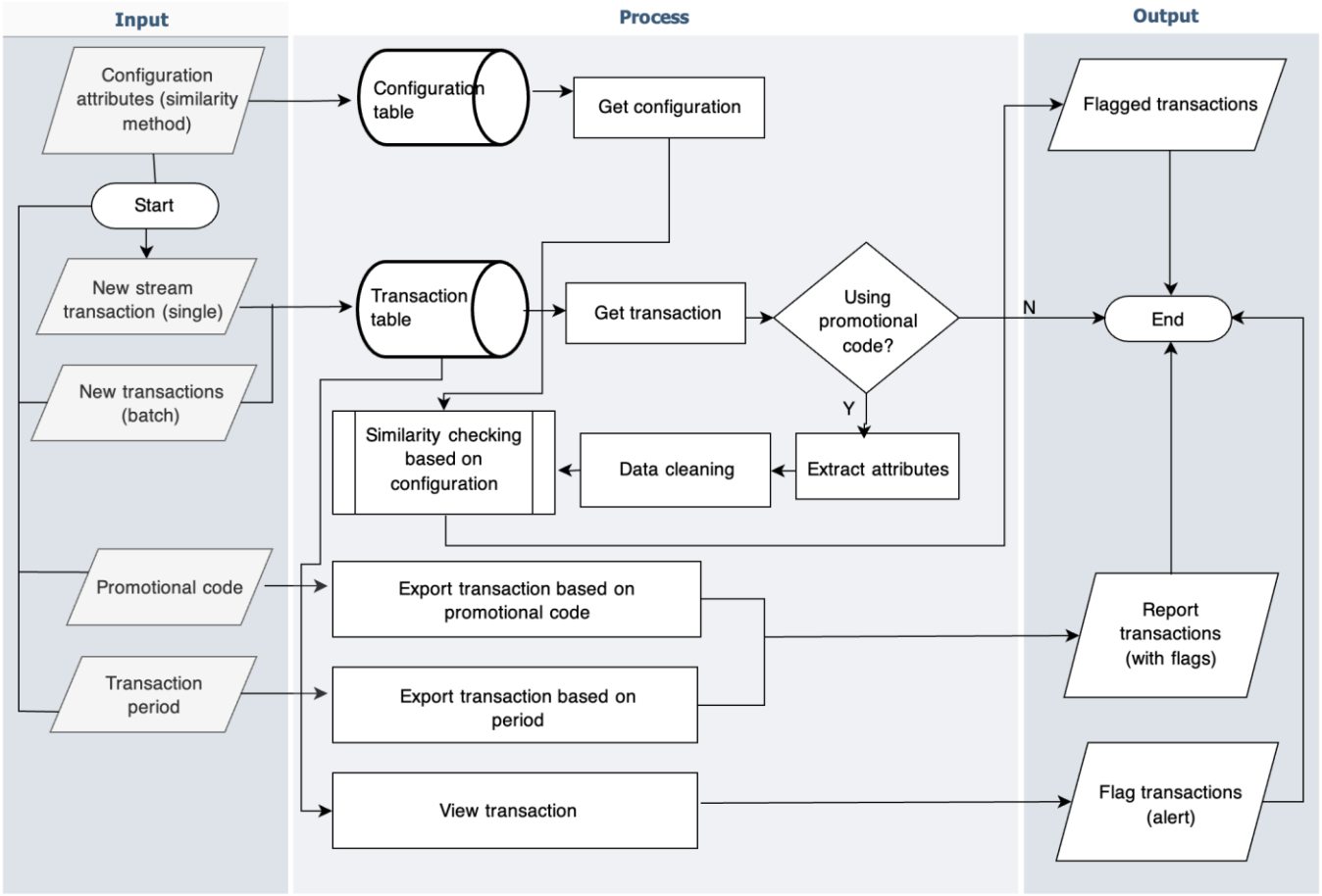

Fig. 3 Schematic of promotion abuse detection system process.

number, member email, order email, payment ID, and product name. However, since each attribute had the same weight, the risk level of one attribute with another was the same. Therefore, the risk level of these attributes for the proposed system must be determined through further analysis.

G. Risk Level of the Attributes of Online Promotion Abuse

Fig. 2 displays the Pareto chart of the relationship between attributes and similarities of manual labeling transaction data that has been labelled as promotion abuse. The chart suggests that 80% of promotion abuse indications were found in the shipping address, member address, series, and mobile attributes. To get the level or the weights for scoring risk transactions, the values of the percentage of attributes was used, as shown in Table VII. Table VII suggests that the most dominant indications of fraud are the Member Address and Shipping Address attributes. The sum of the results of the similarity ratings was multiplied by the weight of each attribute to compute the value of the transaction risk. This would apply to a risk assessment in the proposed system.

IV. System Implementation and Result

A. Design of Promotion Abuse Detection System

Fig. 3 illustrates the input-process-output interaction of the proposed system based on the explanation of the data processing that was carried out with the results of attributes selection as well as similarity and scoring methods. The input was a bulk transaction dataset. Subsequent verification and query transactions were carried out using promotional codes. Transaction records were extracted into seven transaction attributes for analyzing promotion misuses. Next, data clean-up replaced the words in the phrase with the created data dictionary. After that, similarity checking was carried out as in the factor testing stage. It, therefore, generated indications as to whether the promotion was abused or not in the processed transactions.

Fig. 3 describes that there are two modes of analysis for transactions, namely streams, and bulk data. Analysis of the live stream data was conducted when the user inputs the promotional code. The system would verify the potential for promotion abuse based on the member address, delivery address, member email, order email, mobile phone number, product name and payment ID. Subsequently, the similarity method would verify the risk according to the risk attributes in the database. The risk points would be added according to the weighting indicated in Table VI.

The bulk transaction analysis was carried out by importing transaction data into the database system. After that, the use of promotional code in the existing transaction was checked. If the promotion code was not applied, promotion abuse detection

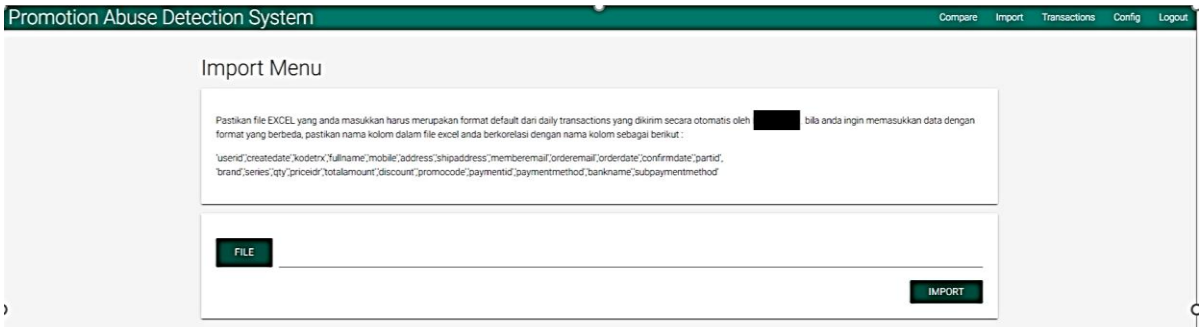

Fig. 4 Page of imported transactions.

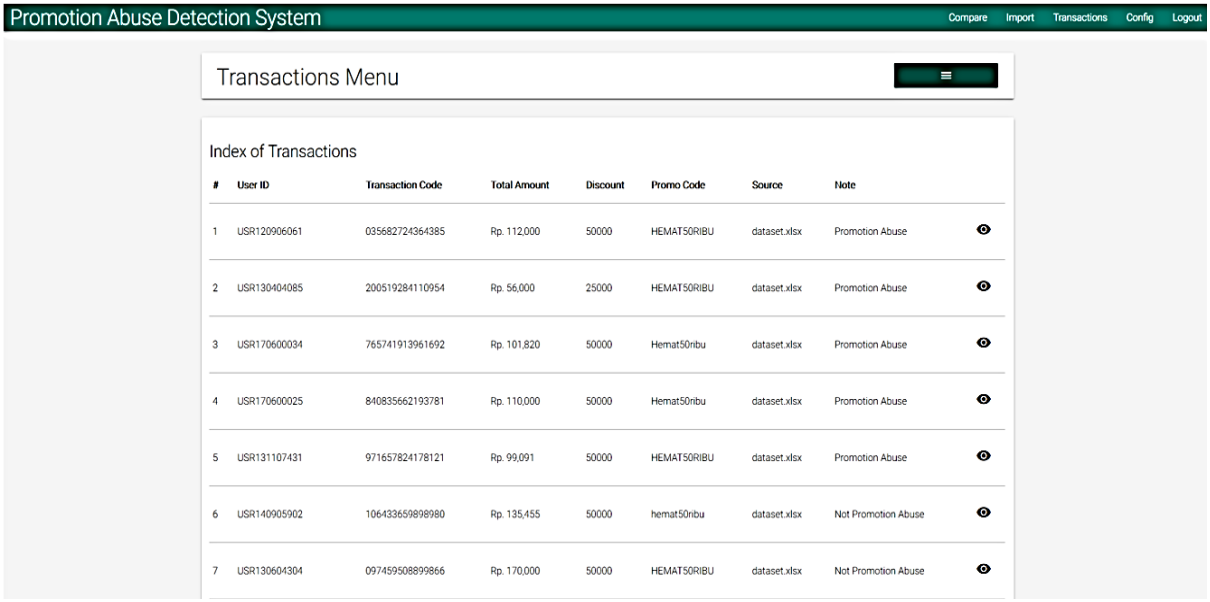

Fig. 5 Page of promotion abuse detection.

would not be conducted; if otherwise, an extraction of the transaction attributes needed for promotion abuse analysis was carried out. Afterward, data cleaning was carried out by replacing the words in the sentence with the data dictionary created. Then, similarity checking was conducted as in the factor testing stage and generated indications whether promotion abuse occurred during the transaction process. The purpose of this bulk transaction mode is to adjust the risk weight, and to ensure whether this method is still applicable for the live stream data detection given that the online fraud mode is potentially changing and developing. The designed system was then developed into an application, as shown in Fig. 4 and Fig. 5.

B. Performance Evaluation of the Proposed System

Fig. 5 shows the import function page of bulk transactions into the database with the required attributes query format. The system would reject the imported file and display an error message if the format or attributes for promotion abuse analysis did not match.

C. System Evaluation

The testing of the proposed system using actual data was conducted to determine the capability of promotion abuse detection done. This testing used existing transaction data in June 2019, with 577 successful transactions using promotional codes. It indicated that the transaction met the promotion abuse detection criteria. Table VIII shows the testing results using the promotion abuse detection system.

TABLE VIII
AUTOMATIC DETECTION TEST OF PROMOTION ABUSE

No.	Criteria	Value
1	TP	38
2	TN	534
3	FP	2
4	FN	3

Based on Table VIII, three test classifications can be calculated, namely:

$$Precision = \frac{tp}{tp + fp} = \frac{38}{38 + 2} = 95\%$$

$$Recall = \frac{tp}{tp + fn} = \frac{38}{38 + 3} = 93\%$$

$$F - Score = 2 * \frac{p * r}{p + r} = \frac{0.95 * 0.93}{0.95 + 0.93} = 0.938272.$$

The test results proved that the proposed algorithm designed to detect promotion abuse was in accordance with similarity and scoring calculations performed on the hypothesis testing. Thus, compared to other methods in [2], [7], [9], the proposed system was proven to be more effective given two mechanisms to detect fraud or promotion abuse, both on streaming data and on bulk data. From the testing, the algorithm implemented to the proposed system had an accuracy score of 0.94. Since the score closes to 1, which is the highest possible value of an F-score, the proposed system is declared feasible to detect promotion abuses on e-commerce.

V. CONCLUSION

Based on the data processing, the factors influencing the detection of promotion abuse are the similarity of attributes in transaction data. Of the four data tables and a total of 24 attributes that were generally used in e-commerce transactions, there were only seven indications of promotion abuse. It is in accordance with the concept of Pareto analysis. Based on the results of factor testing, the combination of the attributes with the most significant effect on the detection of promotion abuse was member address, shipping address, mobile, member email, order email, product name, and payment ID. Based on the exploratory factor test to check similarity, the Levenshtein algorithm had the best performance compared to the exact and LCS algorithms.

The proposed system had two detection modes for real-time transaction live stream data with triggers for the use of promotional codes redeemed. Then, the risk level calculation was carried out with additional iterations of risk based on the weight of each attribute bearing similarities. Meanwhile, for bulk data, transaction data records were used. Based on the tests conducted on 577 transactions, the precision score was 95%, and the recall was 93%, with an F-measure of 0.938272. These results indicate that the detection model applied is appropriate for detecting promotion abuse. Other detection models will be conducted for further research, not only from the promotional code based on user segmentation, but also from demographics and the addition of accuracy detection with the IP address attribute.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTION

Conceptualization, methodology, and resources, Cut Fiarni and Arief S. Gunawan; software, Ishak Anthony; validation, Cut Fiarni, Arief S. Gunawan, and Ishak Anthony; formal analysis, investigation, and data curation, Ishak Anthony; writing Cut Fiarni and Ishak Anthony.

REFERENCES

[1] Bank Indonesia, "Synergize to Build Optimism for Economic Recovery," 2020, [Online], https://www.bi.go.id/en/publikasi/laporan/Documents/2020_LTBI.pdf.

[2] European Consumer Centres Network, "Fraud in Cross Border E-Commerce," 2017, [Online], https://ec.europa.eu/info/sites/default/files/online_fraud_2017.pdf.

[3] A.S. Putri and R. Zakaria, "Analisis Pemetaan E-Commerce Terbesar di Indonesia Berdasarkan Model Kekuatan Ekonomi Digital," *Sem., Konf. Nas. IDEC 2020, 2020*, pp. C06.1–14.

[4] U. Fiore, *et al*., "Using Generative Adversarial Networks for Improving Classification Effectiveness in Credit Card Fraud Detection," *Inf. Sci.*, Vol. 479, pp. 448–455, Apr. 2019.

[5] T. Amarasinghe, A. Aponso, and N. Krishnarajah, "Critical Analysis of Machine Learning Based Approaches for Fraud Detection in Financial Transactions," *Proc. 2018 Int. Conf. Mach. Learn. Technol.*, 2018, pp. 12–17.

[6] A. Bartoli and E. Medvet, "An Architecture for Anonymous Mobile Coupons in a Large Network," *J. Comput. Netw., Commun.*, Vol. 2016, pp. 1–10, Dec. 2016.

[7] A. Saputra and Suharjito, "Fraud Detection Using Machine Learning in E-Commerce," *Int. J. Adv. Comput. Sci., Appl. (IJACSA)*, Vol. 10, No. 9, pp. 332–339, 2019.

[8] Y. Sibaroni, M. Ekaputra, and S. Prasetiyowati, "Detection of Fraudulent Financial Statement based on Ratio Analysis in Indonesia Banking Using Support Vector Machine," *J. Online Inf.*, Vol. 5, No. 2, pp. 185-194, Dec. 2020.

[9] S. Marchal and S. Szyller, "Detecting Organized Ecommerce Fraud Using Scalable Categorical Clustering," *Proc. 35th Annu. Comput. Secur. Appl. Conf.*, 2019, pp. 215–228.

[10] E. Sipayung, C. Fiarni, and R. Tanudjaya, "Modeling Data Mining Dynamic Code Attributes with Scheme Definition Technique," *Proc. Elect. Eng. Comput. Sci., Inform.*, 2014, pp. 25–28.

[11] J. Wang, H.T. Shen, J. Song, and J. Ji, "Hashing for Similarity Search: A Survey," 2014, arXiv:1408.2927.

[12] A. Niewiarowski, "Short Text Similarity Algorithm Based on the Edit Distance and Thesaurus," *Tech. Trans. Fundam. Sci.*, No. 1-NP, pp. 159–173, Dec. 2016.

[13] Y. Wang, J. Qin, and W. Wang, "Efficient Approximate Entity Matching Using Jaro-Winkler Distance," *Int. Conf. Web Inf. Syst. Eng.*, 2017, pp. 231–239.

[14] P. Christen, "A Comparison of Personal Name Matching: Techniques and Practical Issues," *IEEE Int. Conf. Data Mining-Workshop (ICDMW'06)*, 2006, pp. 290–294.

[15] C. Fiarni, H. Maharani, and C. Nathania, "Product Recommendation System Design Using Cosine Similarity and Content-Based Filtering Methods," *Int. J. Inf. Technol. Elect. Eng.*, Vol. 3, No. 2, pp. 42–48, Jun. 2019.

[16] D.M.W. Powers, "Evaluation: From Precision, Recall, and F-Measure to ROC, Informedness, Markedness, and Correlation," *J. Mach. Learn. Technol.*, Vol. 2, No. 1, pp. 37–63, 2011.