Sistem Deteksi Penyalahgunaan Promosi Menggunakan Metode *Similarity* dan Penilaian Risiko

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

Intisari*—Menawarkan kupon promosi adalah salah satu strategi pemasaran *online* yang paling populer untuk menarik pelanggan baru dan meningkatkan loyalitas pelanggan. Akan tetapi, strategi ini membuka peluang risiko *fraud* karena kupon dapat ditebus berkali-kali menggunakan akun palsu. Risiko ini menjadi beban biaya pemasaran dan menyebabkan kegagalan dalam mencapai nilai strategis yang diharapkan. Oleh karena itu, penelitian ini berfokus untuk membangun sistem deteksi otomatis penyalahgunaan promosi berdasarkan tingkat risikonya. Sistem yang ditawarkan harus mampu bekerja pada data *live stream* dan data *bulk*. Maka, pada data *live stream*, sistem mampu memperingatkan administrator sebelum transaksi selesai atau sebelum langkah selanjutnya dimulai. Analisis faktor *exploratory* terhadap 24 atribut yang dikumpulkan dari empat data tabel transaksi menunjukkan bahwa terdapat tujuh atribut yang terindikasi sebagai penyalahgunaan promosi. Atribut-atribut tersebut meliputi *IP address*, *shipping address*, *mobile number*, *member email*, *order email*, *payment ID*, dan *product name*. Algoritme *similarity* dari *machine learning* terarah kemudian digunakan untuk merancang model dan menemukan korelasi tersembunyi atribut yang dapat digunakan untuk mengindikasikan penyalahgunaan promosi. Hasil perbandingan lima metode *similarity* menunjukkan bahwa berdasarkan alur kerja dan kinerjanya, metode yang paling sesuai adalah metode *exact match* dan *Levenshtein edit base*. Fitur penilaian risiko otomatis dari sistem yang diusulkan menggunakan tujuh atribut transaksi *online* sebagai parameter penyalahgunaan promosi yang paling berpengaruh berdasarkan korelasi tersembunyinya. Dari pengujian kinerja sistem, didapatkan hasil nilai presisi, *recall*, dan *F-measure* yang masing-masing sebesar 95%, 93%, dan 0,94. Hasil ini menunjukkan bahwa kinerja sistem memuaskan.

***Kata Kunci*—Sistem Deteksi, Penyalahgunaan Promosi, *Levenshtein Edit Base*, *Exact Similarity*, Penilaian Risiko, Analisis Faktor *Exploratory*, *E-Commerce*.**

I. PENDAHULUAN

Selama masa pandemi, penggunaan *platform e-commerce* dan instrumen digital lainnya semakin meningkat. Hal ini sejalan dengan banyaknya program pelatihan dari pemerintah Indonesia yang turut serta mendorong usaha mikro kecil menengah (UMKM) untuk bertahan dan berkembang mengikuti tren *e-commerce*. Pertumbuhan nominal transaksi *e-commerc*e Indonesia menjadi bukti peningkatan tren *e-commerce*, ditandai dengan peningkatan menjadi 29,6% (YoY) pada tahun 2020 [1]. Peningkatan ini juga tidak terlepas dari meningkatnya preferensi masyarakat terhadap penggunaan *platform* digital yang mampu mencegah terjadinya kerumunan serta meningkatnya strategi sejumlah *e-marketplace* yang menawarkan berbagai promosi. Penggunaan strategi promosi bertujuan untuk memperkenalkan produk baru, menarik pelanggan baru, dan mempertahankan loyalitas pelanggan. Salah satu metode promosi yang digunakan adalah dengan menerbitkan kode promosi yang dapat dipakai untuk mendapatkan diskon. Kode promosi ini biasanya terdiri atas angka dan huruf unik terkait dengan merek barang promosi, yang umumnya akan diisikan pada saat proses *checkout* transaksi.

Dalam menggunakan kode promosi, *e-commerce* harus menjalankan proses keamanan transaksi untuk menekan risiko inherennya. Prosedur keamanan ini meliputi proses autentikasi kupon dan memastikan bahwa kupon tersebut tidak dimodifikasi, belum ditukar, dan masih berlaku. Akan tetapi, risiko penyalahgunaan promosi yang dilakukan oleh oknum tidak bertanggungjawab masih terjadi pada strategi pemasaran ini. Salah satu modusnya adalah dengan memanfaatkan celah pada mekanisme kode promosi untuk pelanggan baru yang hanya bisa digunakan satu kali. Dalam operasinya, penipu akan memalsukan akun untuk mendapatkan harga promosi. Tindakan ini dikenal sebagai penyalahgunaan promosi dan merupakan bagian dari *online fraud* [2]. Penyalahgunaan promosi pada umumnya dimanfaatkan untuk mendapatkan keuntungan tambahan dari harga diskon penjualan. Modus yang digunakan oleh penipu sangat bervariasi dan semakin berkembang seiring dengan perkembangan teknologi, misalnya peretasan, *scrapping*, dan *social engineering*. Perusahan-perusahan biasanya bergantung pada pakar analisis *fraud* dalam upaya memeriksa setiap transaksi yang terindikasi sebagai penyalahgunaan promosi berdasarkan catatan transaksi yang telah dilakukan. Setelah itu, hasilnya dikirim ke departemen penjualan untuk evaluasi lebih lanjut. Namun, prosedur ini memiliki kekurangan, yang terletak pada kelemahannya dalam mencegah risiko *fraud*.

Penelitian terkait *online fraud* umumnya berfokus pada transaksi perbankan, sedangkan pada saat yang sama, UMKM telah mulai merambah bisnis *online*. Oleh karena itu, penelitian pada penyalahgunaan transaksi sangat penting, apalagi UMKM umumnya tidak memiliki sumber dana yang cukup untuk mempekerjakan pakar analisis *fraud*. Oleh karena itu, penelitian ini berfokus pada pemodelan proses analisis deteksi *fraud* berdasarkan cara pakar menganalisis dan melakukan penelitian berbasis data dalam memanfaatkan *machine learning* untuk menemukan atribut-atribut yang mengindikasikan *fraud*. Model yang dihasilkan akan digunakan pada sistem deteksi *fraud* otomatis untuk mencegah kecurangan transaksi pada

[1,3] *Departemen Sistem Informasi, Institut Teknologi Harapan Bangsa, 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, Belgia; (email: ariefsamuel.gunawan@ugent.be)*

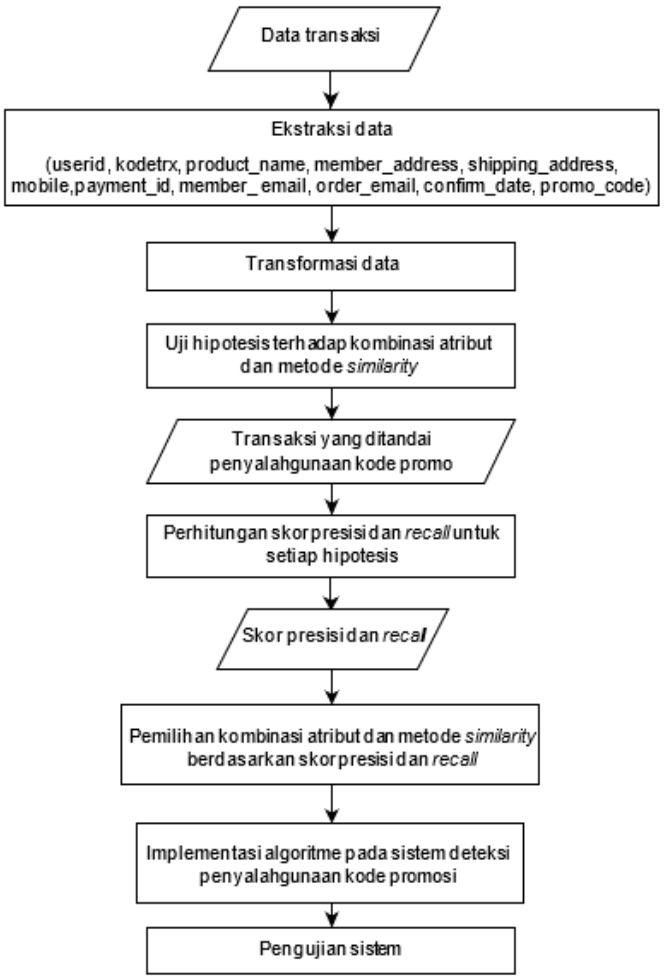
[Diterima: 1 Januari 2022, Direvisi: 8 Maret 2022]

penggunaan kode promosi. Data transaksi yang digunakan dalam penelitian ini diperoleh dari PT. XYZ yang merupakan salah satu dari sepuluh besar perusahaan *e-commerce* dengan kinerja terbaik di Indonesia pada tahun 2020 [3].

II. DETEKSI *FRAUD* PADA TRANSAKSI *ONLINE*

Deteksi *fraud* otomatis pada transaksi *online* merupakan kajian utama pada era *e-commerce*. Seiring dengan bergesernya kebiasaan orang yang lebih memilih pembayaran *online*, catatan transaksi *online* semakin meningkat dari segi volume, kecepatan, dan keragaman atribut. *Big Data* ini dapat memberikan pengetahuan yang berharga tentang transaksi pengguna, yang tidak hanya untuk menyegmentasi pasar sasaran yang dipersonalisasi, tetapi juga untuk memetakan korelasi antara *fraud* transaksi dan perilaku pengguna. Hal ini kemudian mendorong implementasi *machine learning* pada penelitian di bidang keamanan dan manajemen risiko transaksi *e-commerce*. *State-of-the-art* penelitian di bidang ini pada dasarnya bertujuan untuk mengidentifikasi pola mencurigakan dari analisis catatan transaksi [4]. Pada bidang ini, penelitian terkait penyalahgunaan promosi di *e-commerce* masih terbatas karena umumnya hanya berfokus pada transaksi perbankan dan mekanisme pembayaran menggunakan kartu kredit sebagai subjek. Misalnya, sebuah penelitian berfokus pada peningkatan efisiensi dan stabilitas *platform* deteksi *fraud* untuk kartu kredit dengan memanfaatkan *deep neural network* [5]. Sementara itu, telah dilakukan perbaikan arsitektur sistem kupon seluler menggunakan teknik kriptografi kode QR [6]. Pemodelan berbasis data yang menggunakan klasifikasi dengan *machine learning* mampu menangani data yang tidak seimbang dengan volume yang besar. Referensi [7] membandingkan algoritme *machine learning* untuk mendeteksi *fraud* kartu kredit di *e-commerce*. Hasil penelitian menunjukkan bahwa nilai akurasi tertinggi adalah *neural network* dengan nilai 96%, sedangkan *random forest* dan *naïve Bayes* menghasilkan akurasi sebesar 95%. Di sisi lain, *decision tree* menghasilkan nilai terendah, yaitu sebesar 91%. Referensi [8] menerapkan *classifier support vector machine* (SVM) untuk mendeteksi *fraud* dalam laporan keuangan berdasarkan analisis rasio. Sementara itu, telah dilakukan juga penelitian deteksi *fraud* pada *e-commerce* terorganisasi menggunakan *scalable categorical clustering* untuk *dataset bulk* [9]. Penelitian ini berhasil mendeteksi 26,2% *fraud* dan hanya menimbulkan 0,1% *false alarm* dari transaksi legal.

Penelitian ini membangun deteksi penyalahgunaan promosi dengan menggabungkan proses pakar untuk menganalisis *fraud* dan menggunakan *machine learning* untuk memodelkan pembelajaran pola data transaksi. Dalam upaya mengurangi penyalahgunaan promosi, langkah pencegahan pada dasarnya diperlukan untuk secara otomatis mengategorikan transaksi dalam sistem sebagai penyalahgunaan promosi atau bukan, sehingga transaksi-transaksi tersebut tidak menyebabkan *bottleneck* dalam siklus transaksi *online* yang pada akhirnya dapat memengaruhi kinerja dan reputasi perusahaan *e-commerce*. Untuk itu, fokus penelitian ini adalah mengembangkan sistem deteksi penyalahgunaan promosi yang dapat melakukan *tagging* untuk mencegah terjadinya *fraud*.



Gbr. 1 Kerangka kerja penelitian model deteksi penyalahgunaan promosi.

Model sistem yang diusulkan menggunakan *machine learning* berbasis data untuk mencocokkan profil dan karakter pengguna yang melakukan penyalahgunaan promosi di *e-commerce* ritel B2C dan C2C.

III. DETEKSI PENYALAHGUNAAN PROMOSI

Bagian ini membahas skema dan alur metode untuk mengembangkan sistem yang diusulkan.

A. Kerangka Penelitian

Sistem yang dibuat bertujuan untuk menganalisis nilai risiko transaksi untuk menunjukkan indikasi penyalahgunaan promosi berdasarkan persamaan atribut transaksi yang menghasilkan skor. Berdasarkan Gbr. 1, skema penelitian dimulai dari memasukkan catatan data transaksi dari modul *point of sales* (POS) PT. XYZ. Kemudian, data transaksi menjadi masukan yang akan melalui proses-proses berikut.

1) Ekstraksi Data dan Transformasi: Proses ini menghasilkan data transaksi dengan atribut yang diperlukan untuk menentukan model deteksi penyalahgunaan promosi. Setelah data transaksi dengan atribut yang dipilih diambil, transformasi data akan dilakukan untuk meningkatkan akurasi deteksi dan menyamakan format data.

2) Analisis Faktor Exploratory: Tujuan dari uji faktor *exploratory* adalah untuk mengidentifikasi transaksi yang dianggap sebagai penyalahgunaan promosi berdasarkan kombinasi atribut dan metode *similarity* yang digunakan.

3) Perhitungan Skor Precision dan Recall: Hasil dari kombinasi atribut dan metode *similarity* digunakan untuk menandai penyalahgunaan promosi pada setiap transaksi. Berdasarkan hasil ini, nilai *precision* dan *recall* akan dihitung dan dibandingkan dengan ulasan manual dari pakar *fraud*.

4) Implementasi Kombinasi Algoritme Terbaik: Nilai terbaik akan dipilih dari hasil *precision* dan *recall*. Kemudian, kombinasi yang dipilih akan menjadi algoritme yang digunakan dalam sistem yang diusulkan.

5) Uji Sistem: Setelah sistem deteksi dikembangkan, sistem tersebut diuji menggunakan pengolahan data dan *dataset* baru untuk memeriksa konsistensi analisis dan perbandingan dengan ulasan manual oleh pakar.

Berdasarkan prinsip kerjanya, penelitian ini merupakan agen berbasis pengetahuan karena algoritme *machine learning* menghasilkan *solution state* berdasarkan data transaksi yang digunakan sebagai data uji pada pemodelan. Model yang dihasilkan bukan hanya hasil dari analisis data *exploratory* (*exploratory data analysis*, EDA*)* untuk mendapatkan atribut data transaksi yang paling berkorelasi dengan kasus *fraud*, melainkan juga algoritme *similarity* dengan kinerja terbaik. Diperlukannya pemahaman proses bisnis dan fungsi modul sistem *e-commerce* serta tipe atribut data menyebabkan setiap penelitian *machine learning* berbasis data menjadi unik dan perlu dilakukan evaluasi serta pengujian rutin parameter atribut yang digunakan pada sistem solusi yang diusulkan [10]. Oleh karena itu, pemodelan terdiri atas dua tahap. Tahap pertama bertujuan untuk memperoleh atribut-atribut yang dapat mengindikasikan penyalahgunaan promosi berdasarkan data transaksi yang telah dianalisis dan diindikasikan oleh pakar sebagai penyalahgunaan promosi. Kemudian, tahap kedua bertujuan menemukan algoritme *machine learning* yang paling tepat untuk diterapkan pada sistem usulan. Atribut ini akan diidentifikasi dengan menemukan korelasi dan koneksi antartransaksi. Dengan penerapan metodologi ini, indikasi *fraud* dapat dideteksi berdasarkan pola dan atributnya.

B. Eksplorasi Data

Sebagaimana telah dijelaskan pada bagian sebelumnya, penelitian ini menggunakan laporan transaksi tahunan PT. XYZ. Tabel I menunjukkan tipe pengelompokan laporan transaksi tahunan. Tahap transformasi data diterapkan pada atribut *member address* dan *shipping address.* Menggunakan kamus data, setiap kata pada kalimat diganti dengan sinonimnya. Lalu, tahap berikutnya adalah tahap penyortiran kode promosi. Penyortiran dilakukan dengan menghapus transaksi dengan kode promosi sekali pakai dari kumpulan data setelah menjalani tahap eksplorasi, praproses, dan pembersihan laporan transaksi tahunan. Hasilnya berupa *dataset* transaksi, termasuk semua atribut yang diperlukan untuk analisis data transaksi serta catatan data yang secara manual diberi label sebagai penyalahgunaan promosi dengan kolom *manual_flag*.

TABEL I
TIPE DATA DARI LAPORAN TRANSAKSI TAHUNAN

Tipe	Data
Karakter	*Name, mobile, address, ship-to address, member* (*email*), *order email, paymentID*
Karakter unik	*UserID*
Kategori	*Status, payment method, bank, subpaymentmethod, channel grouping, month, order date*
Numerik	*Total amount, discount*

TABEL II
ATRIBUT DAN SUMBER DATA

Tabel Data	Atribut	
Pengguna	● *Full name* ● *Mobile* ● *Member email*	● *Address* ● *User ID*
Produk	● *Part ID* ● *Brand* ● *Merchant name*	● *Series* ● *Price* (IDR)
Transaksi	● *Shipping address* ● *Order email* ● *Order date* ● *Promo code* ● *SO Number*	● *Code trx* ● *Quantity* ● *Total amount* ● *Discount*
Pembayaran	● *Payment ID* ● *Payment method* ● *Sub payment method*	● *Confirm date* ● *Bank name*

C. Proses Pemilihan Atribut

Terdapat empat tipe atribut data, yaitu nominal, ordinal, interval, dan rasio. Atribut ordinal dan nominal adalah jenis atribut kategoris yang digunakan untuk data kuantitatif atau pelabelan variabel. Sementara itu, interval dan rasio adalah atribut kuantitatif yang nilainya dapat dikontrol. Berdasarkan analisis dan evaluasi sistem yang berjalan, terdapat 24 atribut yang diperoleh dari empat data tabel, sebagaimana ditunjukkan pada Tabel II.

Berdasarkan analisis bisnis, tiga atribut tidak dapat digunakan sebagai acuan untuk mendeteksi penyalahgunaan promosi karena keunikan dan ketidakcocokannya untuk dikelompokkan, serta merupakan atribut pelengkap informasi, bukan pengidentifikasi transaksi. Atribut-atribut tersebut adalah *user ID*, *payment ID*, dan *full name*. Atribut *part ID* dan *brand* diwakili oleh atribut *series*, yang merupakan nama lengkap dari produk yang dibeli. Atribut *part ID* dan *brand* merupakan penjelasan dari series, sehingga jika terdapat kemiripan pada *series*, *part ID*, dan *brand* juga akan sama. Jika atribut ini menjadi penentu deteksi, *series* akan memiliki faktor penentu tiga kali lebih banyak daripada atribut lainnya. Untuk alasan yang sama dengan poin sebelumnya, atribut *payment method*, *bank name*, dan *sub payment method* diwakili oleh atribut *payment ID*. Sementara itu, atribut *full name* tidak dapat digunakan sebagai acuan karena terdapat kemungkinan yang tinggi terjadinya kemiripan *username*, sehingga akan memicu *false positive* (FP) yang besar. Di sisi lain, pada proses pemilihan atribut, terdapat tujuh atribut terkait penyalahgunaan promosi, yaitu *mobile*, *address*, *member email*, *series*, *shipping address* dan *order email*, *product name*, serta *payment ID*.

TABEL III
ALUR KERJA ALGORITME *SIMILARITY*

Algoritme	Alur Kerja
Jarak Hamming [11]	Mencari kemiripan dengan eksekusi waktu tercepat dibandingkan keempat metode pencari kemiripan lainnya dalam tabel, tetapi panjang *string* yang diukur haruslah sama.
Jarak Levenshtein [12]	Mencari kemiripan yang cocok digunakan untuk *string* dengan banyak kesalahan ketik.
Longest common substring (LCS) [13]	Mencari kemiripan dengan melakukan eliminasi karakter yang tidak cocok antara kedua buah *string*.
Jarak Jaro-Wrinkle [13]	Mencari kemiripan untuk *string* dengan atribut yang pendek seperti nama.
Exact match [13]	Metode tercepat untuk menemukan kemiripan antara dua *string* dengan secara berurutan mencari kemiripan antara setiap karakter dalam dua *string* yang dibandingkan. Hasil perbandingannya adalah 0 dan 1, sehingga tidak dapat digunakan untuk menemukan kemiripan.

Ketujuh atribut ini selanjutnya dikodekan menjadi A1-A7 untuk penyederhanaan penulisan.

D. Pemilihan Algoritme Deteksi Penyalahgunaan Promosi

Membandingkan metode dengan karakteristik sistem yang diusulkan sangatlah penting dalam menentukan metode *similarity* yang sesuai. Tabel III menunjukkan algoritme populer untuk membandingkan nilai *similarity* dan menjelaskan cara kerja metode *similarity*. Berdasarkan analisis kebutuhan bisnis, metode *similarity* yang diaplikasikan harus membandingkan setiap atribut antartransaksi.

Data transaksi mempunyai beberapa karakteristik, seperti yang dideskripsikan pada Tabel IV. Analisis dilakukan dengan menggabungkan kebutuhan bisnis yang diperoleh dari deteksi risiko *fraud*, metode *similarity*, dan karakteristik *e-commerce* berdasarkan data transaksi. Hasil analisis ditunjukkan pada Tabel V. Sebagaimana telah dijelaskan sebelumnya, penelitian ini berfokus pada tujuh atribut (dikodekan menjadi A1-A7) terkait dengan penyalahgunaan promosi. *Dataset* pertama menggunakan 82 transaksi dan kemudian digandakan untuk *dataset* kedua. Dalam hipotesis uji, dua digit pertama mewakili metode *similarity* yang dipilih. Digit selanjutnya menandakan atribut yang dipilih, misalnya, "S2-A1A3A6" yang berarti uji faktor *exploratory* menggunakan *atribut mobile*, *member email*, *order email*, dan metode *similarity* Levenshtein. Terdapat tiga algoritme yang dapat menentukan kemiripan yang sesuai dengan karakteristik dan kebutuhan risiko *fraud* terkait dengan transaksi menggunakan kode promosi [14].

1) Exact Match (Kode: S1): Menggunakan (1), metode ini dilakukan untuk memeriksa kemiripan antara dua buah *string* dengan menghasilkan nilai 1 atau 0 berdasarkan sama atau tidaknya kedua *string* tersebut.

$$sim(A,B) = \ if\ A = B, 1\ otherwise, 0 \quad (1)$$

2) Levenshtein Distance Similarity (LCS) (Kode: S2): Metode ini digunakan untuk memeriksa kemiripan antara dua *string* dengan mengukur jumlah langkah yang diperlukan

TABEL IV
PENGODEAN KARAKTERISTIK TRANSAKSI DATA *E-COMMERCE*

Kode	Karakteristik
K1	Terdapat banyak atribut kategoris yang merupakan masukan dari sistem.
K2	Terdapat banyak indikasi penyalahgunaan promosi pada transaksi yang disebabkan oleh deteksi pola berurutan dalam penamaan alamat email dan nomor telepon.
K3	Terdapat banyak kesalahan ketik pada kolom alamat pengguna dan alamat pengiriman.
K4	Terdapat kata-kata acak dalam kalimat untuk atribut *address* yang merujuk pada alamat yang sama.
K5	Beberapa atribut hanya perlu diperiksa kemiripan dan perbedaannya, seperti *payment ID* untuk setiap pelanggan yang melakukan pembayaran haruslah unik, sehingga satu perbedaan karakter akan dianggap berbeda.

TABEL V
PERBANDINGAN METODE *SIMILARITY* DENGAN DATA TRANSAKSI

Metode	Kode				
	K1	***K2***	***K3***	***K4***	***K5***
Hamming	✓				
Levensthein	✓	✓	✓		
LCS	✓	✓		✓	
Jaro-Wrinkle	✓				
Exact	✓				✓

sehingga satu *string* sama dengan *string* lainnya. Metode ini dilakukan dengan menggunakan tiga langkah, yaitu penjumlahan, penggantian, dan pengurangan untuk setiap karakter dengan rumus sebagai berikut.

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

3) Longest Common Substring (Kode: S3): Metode ini digunakan untuk memeriksa kemiripan antara dua *string* dengan melihat *substring* terpanjang yang sama antara dua *string* dengan menggunakan (3).

$$\begin{aligned} LCSubstr(A,B) = \ & max\ 1 \leq i \leq m, \\ & 1 \leq j \leq n\ \ LCSuff(A1mi, B1nj) \end{aligned} \quad (3)$$

dengan

A = *string* pertama,
B = *string* kedua,
i = jumlah karakter pada *string* pertama,
j = jumlah karakter pada *string* kedua.

E. Penyusunan Hipotesis dari Kombinasi Atribut dan Metode Similarity

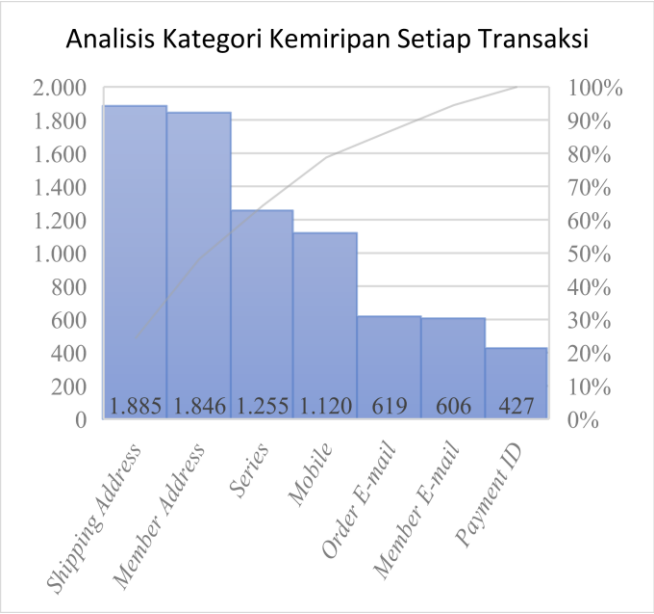
Tahap selanjutnya adalah menguji faktor *exploratory* dari kombinasi atribut dan metode *similarity*. Uji hipotesis pada tahap ini bertujuan untuk mendapatkan model terbaik berdasarkan nilai kinerjanya. Uji hipotesis ini dilakukan dengan

TABEL VI
HASIL PERBANDINGAN FAKTOR *EXPLORATORY* KOMBINASI ATRIBUT

Metode	Atribut	*Dataset* Pertama			*Dataset* Kedua		
		P (%)	*R* (%)	*F1*	*P* (%)	*R* (%)	*F1*
LV	A2A5A1A3A6 A7A5	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	A2A5A1A3A6 A5	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	A2A5A1A3A4 A7	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

Ket.: LV = Levenshtein

mengembangkan fungsi menggunakan bahasa pemrograman PHP dan dimulai dengan memilih uji hipotesis. Misalnya, untuk uji hipotesis 1 (H1), H1 menggunakan metode *exact match* dan kombinasi atribut *member address*. Fungsi ini memeriksa kemiripan menggunakan metode *exact match*. Fungsi *exact match* menerima dua parameter yang disimpan sebagai variabel "$first" dan "$second", kemudian nilainya akan diperiksa. Fungsi tersebut menghasilkan nilai 0 ketika tidak ditemukan kemiripan antara kedua variabel. Hal ini dapat terjadi karena atribut *payment ID* sering bernilai *null*. Langkah yang sama juga berlaku untuk fungsi Levenshtein. Fungsi ini menerima dua parameter yang kemudian disimpan pada dua variabel sebelum dilakukan praproses. Setelah itu, dilakukan



Gbr. 2 Bagan Pareto atribut transaksi data yang menunjukkan penyalahgunaan promosi.

TABEL VII
ATRIBUT DAN SKOR

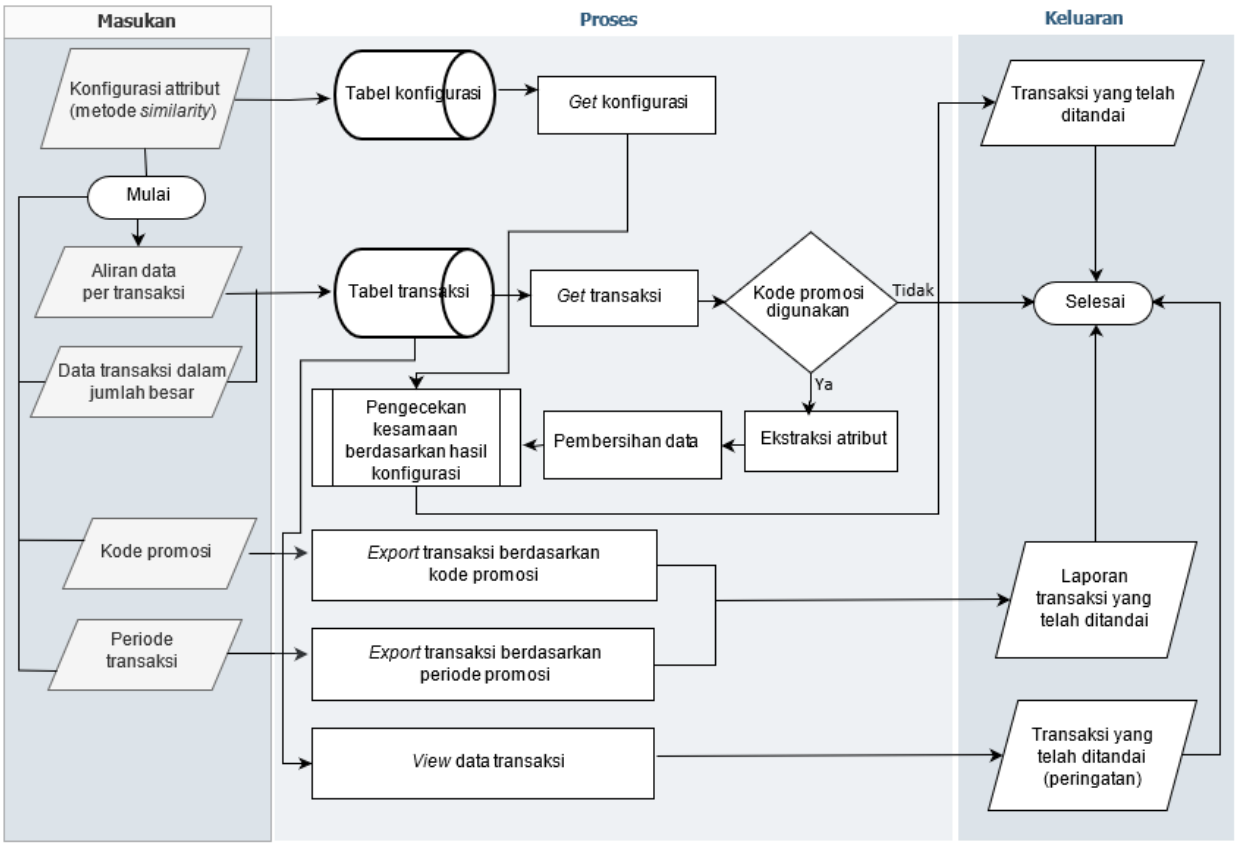
Atribut	Jumlah Kemiripan	Persentase	Bobot
Payment ID	427	5,45%	55
Member address	1.846	23,57%	236
Shipping address	1.885	24,06%	241
Mobile	1.120	14,30%	143
Member email	606	7,74%	77
Order email	619	7,90%	79
Series	1.330	16,98%	170
Total		100,00%	1.000

pengecekan jarak antara dua variabel menggunakan fungsi Levenshtein *default*. Pada fungsi ini, variabel terpanjang digunakan untuk membagi jarak antara dua variabel agar jarak tersebut dapat diubah menjadi nilai desimal antara 0 dan 1. Sebagai contoh, variabel "Bandung" memiliki panjang tujuh karakter dan "Surabaya" memiliki delapan karakter. Pada contoh tersebut, pembaginya adalah delapan karena "Surabaya" memiliki karakter yang lebih panjang. Tabel VI menunjukkan hasil pengujian dari 35 kombinasi menggunakan dua *dataset* yang berbeda.

F. Hasil

Pada tahap ini, setiap hipotesis diuji berdasarkan nilai *precision*, *recall*, dan *F-measure* setiap kombinasi dengan mengikuti langkah berikut [15], [16]. Langkah pertama adalah mencocokkan hasil algoritme menggunakan metode model perhitungan. Kemudian, hasilnya dihitung untuk menentukan nilai *true positive* (TN), FP, *true negative* (TN), dan *false negative* (FN). Terakhir, nilai *precision*, *recall*, dan *F-measure* dihitung dan dibandingkan.

Tabel VI menunjukkan hasil uji *pathfinder* dari metode *similarity*. Berdasarkan hasil uji faktor *exploratory* yang



Gbr. 3 Skema proses sistem deteksi penyalahgunaan promosi.

diperoleh pada setiap hipotesis faktor, metode kombinasi dan atribut akan menghasilkan *flag* yang kemudian dibandingkan dengan *flag* manual, sehingga menghasilkan nilai TP, FP, TN, dan FN. Nilai yang diperoleh menjadi dasar perhitungan akurasi dan *recall* untuk setiap uji faktor *exploratory* [16]. Selanjutnya, faktor dengan nilai *F-measure* lebih besar dari 0,8 dianggap memenuhi kriteria hipotesis.

Perbandingan hasil nilai kinerja hipotesis menunjukkan bahwa fungsi Levenshtein dengan atribut A2A5A1A3A6A7A5 merupakan kombinasi terbaik di antara 35 hipotesis yang ada. Temuan ini menunjukkan bahwa fungsi Levenshtein akan menghasilkan model deteksi penyalahgunaan promosi terbaik, mengingat fungsi ini memiliki nilai *F-measure* tertinggi. Penerapan algoritme Levenshtein pada model sistem dilakukan menggunakan jarak *edit base*. Jarak *edit base* diterapkan untuk mengetahui kemiripan dua *string* dengan membandingkan jumlah jarak *edit* menggunakan penambahan, pengubahan, dan pengurangan karakter pada *string* pertama. Perbedaan jarak *edit* yang diperoleh kemudian dibagi dengan jumlah karakter terpanjang di antara dua *string* dan hasilnya diselisihkan dengan 1 untuk memperoleh persentase kemiripan [12]. Kombinasi atribut ini terdiri atas *member address*, *shipping address*, *mobile number*, *member email*, *order email*, *payment ID*, dan *product name*. Akan tetapi, mengingat setiap atribut memiliki bobot yang sama, maka atribut yang satu dengan yang lain memiliki tingkat risiko yang sama. Oleh karena itu, tingkat risiko atribut-atribut untuk sistem yang diusulkan harus ditentukan melalui analisis lebih lanjut.

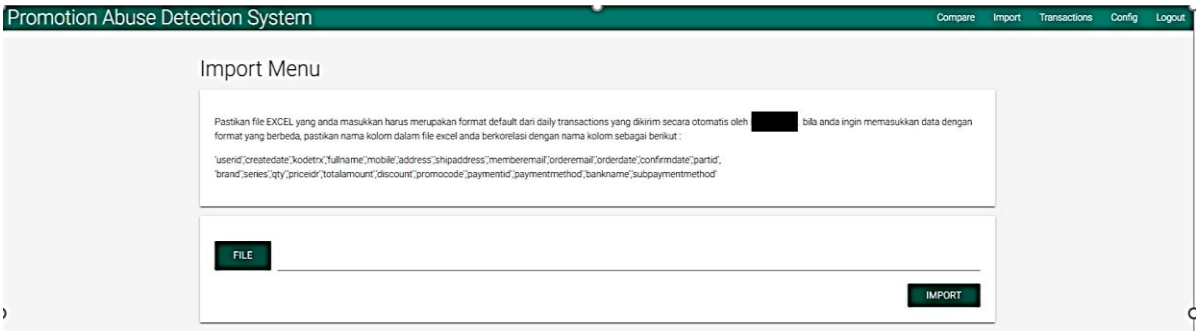
G. Tingkat Risiko Atribut Penyalahgunaan Promosi

Gbr. 2 menampilkan bagan Pareto tentang hubungan antara atribut dan kemiripan data transaksi pelabelan manual yang telah diberi label sebagai penyalahgunaan promosi. Bagan tersebut menunjukkan bahwa 80% indikasi penyalahgunaan promosi ditemukan di atribut *address*, *member address*, *series*, dan *mobile*. Kemudian, nilai persentase atribut digunakan untuk mendapatkan tingkat atau bobot untuk menilai transaksi risiko, seperti yang ditunjukkan pada Tabel VII. Tabel VII menunjukkan bahwa indikasi *fraud* yang paling dominan adalah atribut *member address* dan *shipping address*. Jumlah hasil *rating similarity* dikalikan dengan bobot setiap atribut untuk menghitung nilai risiko transaksi. Hal ini juga berlaku untuk penilaian risiko dalam sistem yang diusulkan.

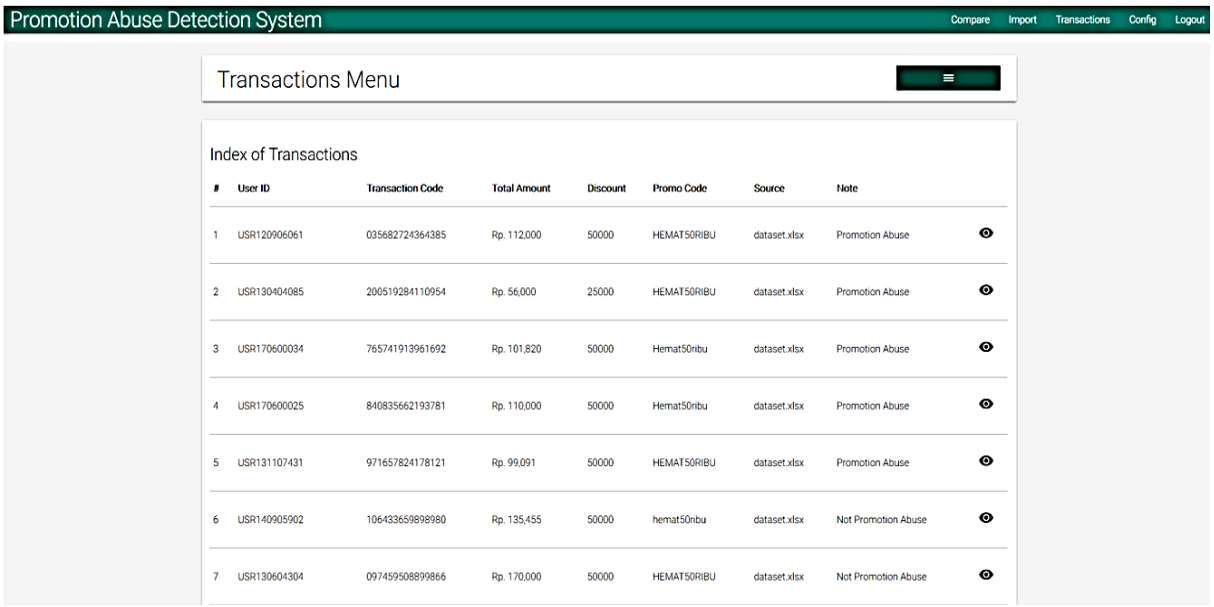
IV. IMPLEMENTASI SISTEM DAN HASIL

A. Desain Sistem Deteksi Penyalahgunaan Promosi

Gbr. 3 menunjukkan interaksi masukan-proses-keluaran dari sistem yang diusulkan berdasarkan penjelasan pemrosesan data yang dilakukan dengan hasil pemilihan atribut serta kemiripan dan metode *scoring*. Masukannya berupa *dataset* transaksi *bulk*. Verifikasi selanjutnya dan transaksi kueri dilakukan dengan menggunakan kode promosi. Catatan transaksi diekstraksi menjadi tujuh atribut transaksi untuk menganalisis penyalahgunaan promosi. Selanjutnya, pada pembersihan data kata-kata dalam frasa digantikan dengan kamus data yang dibuat. Setelah itu, dilakukan pengecekan kemiripan seperti



Gbr. 4 Laman transaksi yang diimpor.



Gbr. 5 Laman deteksi penyalahgunaan promosi.

pada tahap uji faktor. Tahap ini menghasilkan indikasi terjadinya penyalahgunaan promosi.

Gbr. 3 menunjukkan bahwa terdapat dua model analisis untuk transaksi, yaitu data *stream*, dan *bulk.* Analisis data *live stream* dilakukan ketika pengguna memasukkan kode promosi. Sistem kemudian memverifikasi adanya potensi penyalahgunaan kode berdasarkan alamat *member*, alamat pengiriman, alamat *email*, *email* pemesanan, nomor ponsel, nama produk, dan *ID* pembayaran. Kemudian, metode *similarity* akan memverifikasi risiko sesuai dengan atribut risiko pada basis data. Poin risiko akan ditambahkan sesuai dengan pembobotan yang ditunjukkan pada Tabel VI.

Analisis transaksi *bulk* dilakukan dengan mengimpor data transaksi ke dalam sistem basis data. Setelah itu, penggunaan kode promosi dalam transaksi akan diperiksa. Jika kode promosi tidak digunakan, deteksi penyalahgunaan promosi tidak akan dilakukan; jika terjadi sebaliknya, ekstraksi atribut transaksi yang diperlukan untuk analisis penyalahgunaan promosi dilakukan. Setelah itu, pembersihan data dilakukan dengan mengganti kata-kata dalam kalimat dengan kamus data yang telah dibuat. Kemudian, pengecekan kemiripan dilakukan seperti pada tahap uji faktor dan dihasilkan indikasi terjadinya penyalahgunaan promosi selama proses transaksi. Tujuan dari mode transaksi *bulk* ini adalah untuk menyesuaikan bobot risiko serta memastikan berlakunya metode dalam deteksi data *live stream*, mengingat modus *online fraud* dimungkinkan berubah dan berkembang. Sistem yang dirancang kemudian dikembangkan menjadi aplikasi, seperti yang ditunjukkan pada Gbr. 4 dan Gbr. 5

B. Kinerja Evaluasi Sistem yang Diusulkan

Gbr. 5 menunjukkan laman fungsi impor transaksi *bulk* ke dalam basis data dengan format kueri atribut yang diperlukan. Sistem akan menolak berkas (*file*) yang diimpor dan menampilkan pesan galat jika format atau atribut untuk analisis penyalahgunaan promosi tidak sesuai.

C. Evaluasi Sistem

Sistem yang diusulkan diuji menggunakan data aktual. Hal ini dilakukan untuk mengetahui kemampuan deteksi penyalahgunaan promosi. Pengujian ini menggunakan data transaksi pada Juni 2019, dengan 577 transaksi tercatat berhasil menggunakan kode promosi. Hal ini menunjukkan bahwa transaksi tersebut memenuhi kriteria deteksi penyalahgunaan promosi. Tabel VIII menyajikan hasil pengujian menggunakan sistem deteksi penyalahgunaan promosi.

TABEL VIII
UJI DETEKSI OTOMATIS PENYALAHGUNAAN PROMOSI

No.	Kriteria	Nilai
1	TP	38
2	TN	534
3	FP	2
4	FN	3

Berdasarkan Tabel VIII, tiga klasifikasi uji dapat dihitung, yaitu:

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

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

$$F - Score = 2 \times \frac{p \times r}{p + r} = \frac{0{,}95 \times 0{,}93}{0{,}95 + 0{,}93} = 0{,}938272.$$

Hasil uji membuktikan bahwa algoritme yang diusulkan yang dirancang untuk mendeteksi penyalahgunaan promosi telah sesuai dengan kemiripan dan perhitungan penilaian yang dilakukan pada uji hipotesis. Dengan demikian, dibandingkan dengan metode lain [2], [7], [9], sistem yang diusulkan terbukti lebih efektif dengan dua mekanisme untuk mendeteksi *fraud* atau penyalahgunaan promosi, baik pada data *streaming* maupun pada data *bulk*. Dari pengujian tersebut, algoritme yang diimplementasikan ke sistem yang diusulkan memiliki skor akurasi 0,94. Mengingat skor mendekati 1, yang merupakan nilai tertinggi dari *F-score*, sistem yang diusulkan dinyatakan layak untuk mendeteksi penyalahgunaan promosi pada *e-commerce*.

V. KESIMPULAN

Berdasarkan pengolahan data, faktor-faktor yang memengaruhi deteksi penyalahgunaan promosi adalah kemiripan atribut dalam data transaksi. Dari empat tabel data dan total 24 atribut yang umumnya digunakan dalam transaksi *e-commerce*, hanya terdapat tujuh indikasi penyalahgunaan promosi. Hal ini sesuai dengan konsep analisis Pareto. Berdasarkan hasil uji faktor, kombinasi atribut dengan efek paling signifikan pada deteksi penyalahgunaan promosi adalah *member address*, *shipping address*, *mobile*, *member email*, *order email*, *product name*, dan *payment ID*. Berdasarkan uji faktor *exploratory* untuk menganalisis kemiripan, algoritme Levenshtein memiliki kinerja terbaik dibandingkan dengan algoritme *exact* dan LCS.

Sistem yang diusulkan memiliki dua mode deteksi untuk transaksi data *live stream* secara *real-time* dengan pemicu penggunaan kode promosi yang ditukarkan. Kemudian, perhitungan tingkat risiko dilakukan dengan tambahan iterasi risiko berdasarkan bobot masing-masing atribut yang memiliki kemiripan. Sementara itu, untuk data *bulk* digunakan catatan data transaksi. Berdasarkan pengujian yang dilakukan pada 577 transaksi, skor *precision* adalah 95% dan *recall* adalah 93%, dengan *F-measure* sebesar 0,938272. Hasil ini menunjukkan bahwa model deteksi yang diterapkan layak untuk mendeteksi penyalahgunaan promosi. Model deteksi lainnya akan dikembangkan pada penelitian lebih lanjut, tidak hanya dari kode promosi berdasarkan segmentasi pengguna, tetapi juga dari demografi dan penambahan deteksi akurasi dengan atribut alamat IP.

KONFLIK KEPENTINGAN

Penulis menyatakan tidak ada konflik kepentingan.

KONTRIBUSI PENULIS

Konseptualisasi, metodologi, dan sumber daya, Cut Fiarni dan Arief S. Gunawan; perangkat lunak, Ishak Anthony; validasi, Cut Fiarni, Arief S. Gunawan, dan Ishak Anthony; analisis formal, investigasi, dan kurasi data, Ishak Anthony; penulisan, Cut Fiarni dan Ishak Anthony.

REFERENSI

[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 dan R. Zakaria, "Analisis Pemetaan E-Commerce Terbesar di Indonesia Berdasarkan Model Kekuatan Ekonomi Digital," *Sem., Konf. Nas. IDEC 2020, 2020*, hal. C06.1–14.

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

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

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

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

[8] Y. Sibaroni, M. Ekaputra, dan 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, hal. 185-194, Des. 2020.

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

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

[11] J. Wang, H.T. Shen, J. Song, dan 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, hal. 159–173, Des. 2016.

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

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

[15] C. Fiarni, H. Maharani, dan C. Nathania, "Product Recommendation System Design Using Cosine Similarity and Content-Based Filtering Methods," *Int. J. Inf. Technol. Elect. Eng.*, Vol. 3, No. 2, hal. 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, hal. 37–63, 2011.