Terbit *online* pada laman : http://teknosi.fti.unand.ac.id/

Jurnal Nasional Teknologi dan Sistem Informasi

| ISSN (Print) 2460-3465 | ISSN (Online) 2476-8812 |

Artikel Penelitian

Teknik Bagging pada Ensemble Learning untuk Kategorisasi Produk E-Commerce

Faskal Churniansyah [a*], *Danang Wahyu Utomo* [b]

[a,b] Universitas Dian Nuswantoro, Jl. Imam Bonjol No.207, Pendrikan Kidul, Kec. Semarang Tengah, Kota Semarang, Jawa Tengah 50131, Indonesia

INFORMASI ARTIKEL

Sejarah Artikel:
Diterima Redaksi: 01 Februari 2024
Revisi Akhir: 20 Mei 2024
Diterbitkan *Online*: 29 Mei 2024

KATA KUNCI

Kategorisasi produk,
e-commerce,
bagging,
ensemble learning

KORESPONDENSI

E-mail: 111202012616@mhs.dinus.ac.id*

A B S T R A C T

E-commerce merupakan layanan dalam jual beli yang dijalankan secara online melalui media elektronik seperti komputer dan handphone. Adanya perkembangan teknologi informasi yang lebih canggih menjadi pendorong utama dalam meningkatkan kerja e-commerce. Peningkatan yang sering dilakukan adalah menyediakan layanan sebaik – baiknya dan semudah mungkin untuk pelanggan. Banyaknya produk e-commerce yang ditawarkan ke pelanggan menjadi isu utama dalam layanan e-commerce. Tidak sedikit pelanggan yang bingung dalam menentukan pilihan produk. Bahkan beberapa penelitian menyatakan pelanggan yang awam tentang penggunaan e-commerce bingung dalam pemilihan produk. Ada deskripsi atau ulasan produk yang berbeda terhadap produk yang sama. Penelitian ini mengusulkan kategorisasi produk pada layanan e-commerce dengan tujuan menempatkan deskripsi produk sesuai dengan kategori yang telah ditentukan. Teknik bagging adalah Teknik ensemble learning yang mampu membuat beberapa sub pohon keputusan yang nantinya dapat dicari nilai akurasi yang terbaik. Pada hasil pengujian diperoleh bahwa pada pengaturan hyperparameter n_estimators 200 menghasilkan nilai akurasi terbaik dengan nilai 93,25%., precision 93%, recall 93% dan f1-score 93%.

1. PENDAHULUAN

E-commerce merupakan layanan jual beli yang ditawarkan melalui media elektronik secara online. E-commerce telah berkembang pesat dalam beberapa tahun terakhir. Perusahaan e-commerce seperti amazon, e-bay, Taobao, dan Rakuten mencantumkan jutaan produk disitus meraka yang dijual oleh ribuan pedagang. Adanya perkembangan teknologi informasi menjadi faktor pendorong utama meningkatnya pelanggan dalam menggunakan e-commerce [1]. Pelanggan mendorong para pelaku bisnis untuk meningkatkan pelayanan dan investasi pada platform jual beli online. Karena itu, Perusahaan mulai meningkatkan layanan dan keunggulan kompetitif [2]. Adanya perkembangan teknologi dan internet saat ini menjadi faktor utama bagi Perusahaan dan pelaku bisnis dalam melakukan transformasi bisnis dari konvensional ke digital. Perusahaan telah melakukan transformasi ke teknologi digital guna meningkatkan layanan bagi pelanggan dan mampu bersaing dengan kompetitor lain. Layanan jual beli dalam platform online menjadi solusi bagi pelaku bisnis dalam menawarkan produk dalam jumlah banyak, waktu yang singkat dan menjangkau pelanggan dalam lokasi yang berbeda bahkan berbeda negara. Bagi pelanggan, memudahkan dalam melakukan pencarian produk tanpa terbatas oleh ruang dan waktu.

Perkembangan layanan e-commerce ditandai dengan tingginya jumlah transaksi pada platform tersebut. Berdasarkan laman web cbcindonesia.com menunjukkan bahwa tahun 2023 akhir terdapat 361,54 juta transaksi. Dapat disimpulkan bahwa tingginya transaksi pada online menunjukkan jumlah data yang digunakan besar, Artinya transaksi tersebut melibatkan jumlah data dan produk e-commerce yang besar, menurut Oase ansharullah et al platform e-commerce menyediakan produk dalam jumlah banyak[3]. Produk yang diunggah pada platform e-commerce memiliki variasi product yang banyak [4]. Bagi pelaku bisnis, hal ini menjadi sebuah keuntungan karena dapat menampilkan semua produk tanpa adanya batas jumlah produk. Namun bagi pelanggan, terutama pelanggan yang masih awam dengan teknologi online, banyaknya variasi produk yang ditawarkan menjadi masalah dalam pengguanaan platform tersebut. Menururt Suci, banyak pilihan produk dalam platform e-commerce dapat menyebabkan kesalahan pemilihan produk [5]. Selain itu permasalahan juga dapat melibatkan Perusahaan atau pelaku bisnis. Pelaku bisnis yang masih awam dalam penggunaan platform e-commerce akan kesulitan dalam kategorisasi produk

https://doi.org/10.25077/TEKNOSI.v10i1.2024.92-80

dengan jumlah variasi yang banyak. Baik pelanggan maupun pelaku bisnis yang masih awam harus memahami deskripsi satu per satu produk e-commerce. Dari penelitian yang dilakukan oleh Ristoski, salahsatu permasalahan dalam kategorisasi produk adalah kesamaan produk yang dijual namun dengan deskripsi yang berbeda. Perlu ada ketetapan katogorisasi produk yang sesuai untuk membantu pelaku bisnis dalam menempatkan produk kategorisasi yang tepat [6].

Pemasaran produk pada platform e-commerce memberikan identitas tambahan seperti ulasan produk atau deskripsi untuk memudahkan pelanggan dalam memilih produk yang asing atau tidak dikenal. Dalam platform e-commerce, perlu disediakan deskripsi untuk memberikan identitas yang unik untuksetiap produk. Perlu adanya kategorisasi produyk untuk menempatkan deskripsi sesuai dengan kategori[7]. Selain penempatan kategori yang sesuai, kategorisasi produuk dapat digunakan untuk rekomendasi atau prediksi produk ke pelanggan [8]. Permasalahan utama yang telah dibahas pada topik kategoroisasi produk adalah adanya ketidak tepatan produk masuk di dalam sub kategori. Kategori yang dilakukan oleh pelaku bisnis atau operator e-commerce menepatkan produk secara subyektif atau tidak tepat [9]. Kategorisasi berdasarkan deskripsi secara manual membutuhkan waktu lama [10]. Permasalahan lainya, adanya informasi yang hilang atau nama produk yang terlalu singkat dapat mempengaruhi akurasi kategorisasi produk. Faktor lain yang mempengaruhi kategorisasi produk adanya produk tambahan dalam platform e-commerce.

Solusi telah diusulkan untuk permasalahan yang dihasilkan pada kategorisasi produk: *website categorization*[11], *Prototype web*[12], pembelajaran mesin *(machine learning)* atau disingkat ML [13], [14], [15]. Solusi kategorisasi teks berbasis ML menjadi populer diusulkan dalam kategorisasi produk. Kategorisasi produk mampu memberikan akurasi terbaik berdasarkan kategori yang telah disediakan [16]. Teknik, Model, dan Algoritma yang telah diusulkan pada penelitian sebelumnya mampu melakukan kategorisasi produk dengan performa terbaik.

Berdasarkan permasalahan diatas, pada studi ini menggunakan Teknik *ensamble learning* berbasis ML dalam kategorisasi produk terutama untuk data tambahan dalam platform e-commerce. Teknik *ensamble learning* menggabungkan algoritma individu dengan performa lemah untuk mendapatkan performa lebih baik. Penggunaan *ensamble learning* Random Forest dengan XGBoost mampu menghasilkan performa lebih baik dibandingkan algoritma lainya dalam klasifikasi produk review [13]. Penelitian lainya [17] mengusulkan metode *bagging* untuk menangani klasifikasi dengan *data imbalance*. Peneliti menyatakan metode *bagging* mampu menunjukkan performa lebih baik membandingkan metode lainya.

Penelitian ini mengusulkan Teknik *bagging* untuk kategorisasi produk pada dataset e-commerce. Pada tahap eksperimen menggunakan algoritma *Random Forest* untuk membangun beberapa pohon Keputusan secara pararel. Penggunaan Teknik *voting* diterapkan untuk mendapatkan hasil terbaik dari beberapa pohon Keputusan yang digunakan. Tahap evaluasi menggunakan confusion matrix untuk menghitung nilai akurasi yang dihasilkan dari model yang diusulkan. Selain itu menggunakan analisis precision, recall dan f1-score untuk mengevaluasi ketepatan dan kesesuaian model yang diusulkan pada kategorisasi produk e-commerce.

2. METODE

2.1. Ensemble Learning

Ensemble learning adalah Teknik dalam algoritma pembelajaran mesin (ML) dengan kombinasi dua algoritma atau lebih untuk mendapatkan solusi yang lebih baik dari model sebelumnya. Menurut Dong, *ensemble learning* adalah metode yang melibatkan beberapa algoritma ML dengan hasil rendah kemudian ditingkatkan untuk mendapatkan hasil yang lebih tinggi [18]. Teknik *ensemble learning* terdiri dari *bagging, boosting* dan *stacking*. Menurut Qing-Fu Li, Teknik ensemble terdiri dari *bagging* dan *boosting* [19].

Boosting menggunakan kumpulan data untuk pelatihan guna mendapatkan pelajar yang lemah, mengamati hasil pelatihan pelajar yang lemah, menentukan sampel pelatihan yang salah, dan menggunakan set pelatihan yang disesuaikan untuk melatih pelajar lemah berikutnya.

Bagging merupakan proses yang menggunakan beberapa pelatihan *base learner* dengan melibatkan Teknik *voting* sebagai penentu hasil klasifikasi [20]. Tujuan utama adalah menghapus *base learner* yang lemah untuk mendapatkan *base learner* yang kuat. Dari tujuan ini, *base learner* yang dihasilkan oleh model berbeda dengan *base learner* lainnya.

Keuntungan dalam menggunakan Teknik bagging adalah mampu menangani dataset yang tidak seimbang atau dataset dalam jumlah besar. Konsep pemilihan dari beberapa pohon Keputusan yang dibuat menjadi kelebihan Teknik bagging dalam menghasilkan *base learner* yang optimal. Hal ini menjadi alasan Teknik bagging banyak diusulkan dengan topik klasifikasi atau kategorisasi teks.

Pada penelitian ini mengusulkan teknik *bagging* dalam kategorisasi produk e-commerce untuk mendapatkan nilai optimal berdasarkan pohon keputusan yang ditentukan. Menurut Giang Ngo, Teknik *bagging* merupakan Teknik yang banyak digunakan dalam membuat sub pohon Keputusan yang dilatih oleh algoritma ML [21].

2.2. Teknik Bagging

Teknik *bagging* berasal dari kata dasar *bags* yang dikenal sebagai tas yaitu dengan konsep membuat kelompok subset data untuk melatih proses algoritma ML. Dari tas – tas tersebut akan dilatih oleh model gabungan pohon keputusan (*decision tree*) untuk mendapatkan hasil yang lebih baik [22].

Konsep dari Teknik *bagging* adalah membagi data latih menjadi *base learner* yang dipilih secara acak. Dari beberapa *basae learner* tersebut dikombinasikan menggunakan Teknik voting untuk mendapatkan hasil yang optimal [19].

Teknik *bagging* mampu meningkatkan performa belajar dasar algoritma dengan performa yang belum stabil. Keuntungan dalam penggunaan *bagging* adalah mengurangi bias pada variasi subset,

mampu menangani perbedaan data, dan untuk data dalam jumlah besar *bagging* menggunakan waktu komputasi lebih kecil dibanding algoritma ML lainnya. Pada penelitian lainnya, keuntungan teknik *bagging* mampu menangani permasalahan pada data *imbalance* [23].

2.3. Algoritma Random Forest

Algoritma Random Forest

input:

p ← integer

S ← {(x1,y1), (x2,y2), (x3,y3) … , (xn, yn)}

Atribut p dan variabel kelass p

T ← jumlah tree

C ← jumlah label

p jumlah atribut

for t: 1…**to** T **do**:

1. Bangkitkan sample S_j dari data training S
2. Cocokkan *base learner* H_j menggunakan S_j dari node yang diberikan
3. Pilih secara acak atribut k.
4. Hitung fitur pemisahakan terbaik dengan metode pemilihan subset acak
5. Pisahkan node berdasarkan nilai yang diperoleh pada tahap 4.
6. Ulangi tahap 1 sampai 5. Perulangan berhenti sampai kriteria sesuai atau terpenuhi

end for

Random forest adalah salah satu algoritma ensemble learning yang menerapkan Teknik bagging untuk membuat beberapa sub pohon Keputusan. Komponen utama dalam *random forest* adalah pohon Keputusan atau biasa dikenal dengan nama *decision tree* [24]. Komponen utama dalam algoritma *random forest* adalah pembuatan beberapa pohon Keputusan selama uji latih (*base learner*) dan kombinasi aturan atau model yang diusulkan. Dari 2(dua) komponen dapat ditemukan nilai prediksi terbaik dari beberapa pohon Keputusan yang dibuat.

Dalam implementasinya pada kasus klasifikasi atau kategorisasi, algoritma dengan Teknik bagging menggunakan konsep tas (bag) dan fitur keacakan. Artinya, dalam pencarian nilai optimal, random forest membuat beberapa bag dan mengacak fitur atau datanya. Dari data acak pada bag – bag tersebut akan ditemukan nilai yang optimal. Berdasarkan urutan algoritma diatas, *random forest* menggunakan bag acak dari jumlah atribut yang ditentukan. Selanjutnya, tahap pemisahan ditentukan berdasarkan bag yang dipilih (terpilih). Dari proses tersebut, dapat dipastikan bahwa masing – masing bag adalah acak dan tidak ada keterkaitan dengan bag lain. Tujuannya, digunakan untuk menurunkan variasi jumlah model yang diusulkan sehingga dapat meningkatkan nilai akurasi dari model yang diusulkan.

Keuntungan dari algoritma *random forest* adalah dapat menurunkan jumlah variasi pohon keputusan sehingga dapat diterapkan pada dataset dalam jumlah yang banyak dan menangani data yang hilang [25].

2.4. Dataset

Uji kategorisasi produk menggunakan data e-commerce dengan data publik yang dapat diakses pada doi.org/10.5281/zenodo.3355823. Dataset tersebut terdiri 50425 baris dengan 4 kelas atau label. Format dataset adalah csv. Pada Gambar 1 dan Gambar 2 menunjukkan sampel data yang digunakan pada kategorisasi produk.

	label	desc
0	Household	Paper Plane Design Framed Wall Hanging Motivat...
1	Household	SAF 'Floral' Framed Painting (Wood, 30 inch x ...
2	Household	SAF 'UV Textured Modern Art Print Framed' Pain...
3	Household	SAF Flower Print Framed Painting (Synthetic, 1...
4	Household	Incredible Gifts India Wooden Happy Birthday U...

Gambar 1. Sampel 5 data awal

Pada Gambar 1 menunjukan sampel 5 data awal dataset yaitu pada urutan 0 sampai 4. Secara berurutan deskripsi menunjukkan pada label *household.*

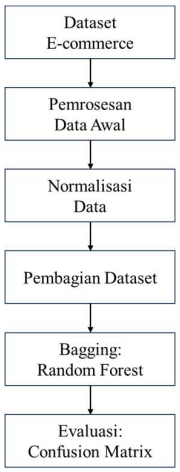
	label	desc
50420	Electronics	Strontium MicroSD Class 10 8GB Memory Card (Bl...
50421	Electronics	CrossBeats Wave Waterproof Bluetooth Wireless ...
50422	Electronics	Karbonn Titanium Wind W4 (White) Karbonn Titan...
50423	Electronics	Samsung Guru FM Plus (SM-B110E/D, Black) Colou...
50424	Electronics	Micromax Canvas Win W121 (White)

Gambar 2. Sampel 5 data akhir

Pada Gambar 2 menunjukkan sampel 5 data akhir (terakhir) dengan urutan 50420 sampai 50424. Dari deskripsi yang ditampilkan menunjukkan deskripsi pada label *electronics.*

e-commerce dataset terdiri dari 4 label *household, books, electronics, clothing & accessories*. Masing – masing label memiliki deskripsi yang ditempatkan pada kolom *desc*.

2.5. Eksperimen



Gambar 3. Alur Penelitian

https://doi.org/10.25077/TEKNOSI.v10i1.2024.92-99

Pada Gambar 3, diawali dari persiapan dataset e-commerce. Dataset diunggah pada alat *google collaboratory*. Tahap selanjutnya terdiri dari pemrosesan data awal, normalisasi data, pembagian dataset, implementasi *bagging* yaitu *random forest* dan evaluasi menggunakan *confusion matrix*. Analisis evaluasi dilengkapi dengan analisis precision, recall, dan f1-score.

2.5.1. Pemrosesan Data Awal

Pemrosesan data awal atau biasa dikenal dengan istial *data preprocessing* adalah tahap awal melakukan pembersihan data dari tanda baca, simbol atau karakter selain huruf abjad, dan penentuan *lowercase*. Tahap pemrosesan data awal terdiri dari: konversi huruf ke lowercase, menghapus tanda baca seperti titik, koma, tanda seru, tanda tanya, petik satu, petik dua, sama dengan dan simbol – simbol selain abjad. Selanjutnya menerapkan *stopword removal* yaitu menghapus kata – kata yang tidak relevan dalam library yang digunakan.

Setelah proses diatas, dilakukan proses tokenisasi yaitu memisahkan teks menjadi terpisah per kata. Sebagai contoh:

Deskripsi dataset:
"*paper plane design frame wall hang motiv offic decor art print set paint make synthet frame uv textur print give multi effect attract toward special seri paint make wall beauti give royal touch paint readi hang would proud possess uniqu paint nich apart use modern effici print technolog print ink precis epson roland hp printer innov hd print techniqu result durabl spectacular*"

hasil tokenisasi:
"paper", "plane", "design", "frame", "wall", "hang", "motiv", "offic", "décor", "art", "print", "set", "paint", "make", "synthet", "frame", "uv", "texture", "print", "give", "multi", "effect", "attract", "toward", "special", "seri", "royal", "touch", "readi", "hang", "would", "possess", "uniqu", "precis", "epson", "hp", "innov","technolog", "use", "hd", "durabl", "spectacular".

Dari hasil tokenisasi diatas, akan dihitung frekuensi kemunculan per kata tersebut. Penghitungan tersebut menggunakan persamaan TF-IDF.

2.5.2. Transformasi Data

$$Tf_{ij} = \frac{f_{i,j}}{\sum k\, f_{i,j}} \quad (1)$$

$$IDf_i = \log\left(\frac{N}{n_i}\right) + 1 \quad (2)$$

$$TFIDF_{ij} = TF_{ij} \times IDf_i \quad (3)$$

Dataset e-commerce merupakan dataset dalam bentuk teks yang artinya dataset tidak dapat langsung diproses untuk dihitung nilai akurasinya. Diperlukan transformasi data untuk konversi data dari bentuk teks ke bentuk numerik. Berdasarkan persamaan (1), (2), dan (3) menggunakan persamaan TF-IDF untuk menentukan frekuensi dari kemunculan teks dalam dataset. Pada uji dengan alat bantu matplotlib menggunakan library *tfidfvectorizer* untuk melakukan konversi dari data teks ke data numerik.

2.6. Skenario Pengujian

Skenario pengujian menggunakan pembagian dataset (*split data*) sebesar 0.2 atau 20% untuk data uji dan 0.8 atau 80% untuk data latih. Data latih dipilih secara acak dari dataset yang digunakan.

Skenario pengujian menggunakan pengaturan hyperparameter (*hyperparameter tunning*) pada algoritma *random forest* yaitu menggunakan parameter *n_estimators* dan *max_depth*. Berikut adalah pengaturan hyperparameter:

Tabel 1. *Hyperparameter tunning*

n_estimator	***max_depth***
[50, 100, 200]	[None, 10, 30, 50]

Pada Tabel 1, dipilih n_estimator yaitu 50, 100, dan 200 yang digunakan untuk menentukan batas jumlah pohon parameter max_depth dipilih adalah None, 10, 30 dan 50. Parameter max_depth dipilih untuk menentukan jumlah pemisahan dari setiap pohon Keputusan. Untuk menentukan hasil terbaik berdasarkan parameter tersebut menggunakan teknik *grid search*.

2.7. Confusion Matrix

Penentu evaluasi atau pengujian menggunakan *confusion matrix* atau matriks kebingungan (CM). Teknik CM digunakan sebagai evaluasi akurasi hasil kategorisasi produk. CM terdiri dari *true positive (TP), false positive (FP), true negative (TN), false negative (FN)*. Dari komposisi tersebut dapat ditentukan untuk penghitungan akurasi, presisi, recall, dan f1-score sebagai analisis hasil kategorisasi terhadap model yang digunakan.

3. HASIL

Tahap pertama adalah pemrosesan data awal (*preprocessing)*, yaitu tahap persiapan dataset selanjutnya dilakukan *cleaning* atau pembersihan data seperti penghilangan tanda baca, simbol, dan perubahan ukuran teks seperti ke *lowercase*. Hasil pembersihan data seperti pada contoh berikut:

"*paper plane design frame wall hang motiv offic decor art print set paint make synthet frame uv textur print give multi effect attract toward special seri paint make wall beauti give royal touch paint readi hang would proud possess uniqu paint nich apart use modern effici print technolog print ink precis epson roland hp printer innov hd print techniqu result durabl spectacular*"

Gambar 4. Contoh Pembersihan Data

Pada Gambar 4, sampel data menunjukkan bahwa teks bersih dari tanda baca, simbol, dan huruf besar. Proses pembersihan data dilakukan untuk memudahkan dalam kategorisasi produk dalam memunculkan frekuensi kata dalam deskripsi tersebut.

Tahap selanjutnya adalah stemming yaitu ditujukan terhadap kolom deskripsi pada dataset. Sebagai contoh hasil stemming pada Tabel 2 berikut:

Tabel 2. Contoh Stemming

Teks	hasil *stemming*
motivate	*motiv*
texture	*textur*
synthetic	*synthet*
office	*offic*
decorate	*decor*
romanctic	*romant*
painting	*paint*
beautiful	*beauty*
durable	*durabl*
technique	*technique*
efficient	*effici*
technologi	*technolog*

Dari tabel 2 diatas, menunjukkan beberapa contoh kata atau teks yang diubah ke kata dasar. Hanya saja, beberapa kata menjadi aneh atau asing karena ada bagian yang sudah menjadi kata dasar tapi karakternya dihilangkan dari kata tersebut. Sebagai contoh adalah texture menjadi textur.

Setelah tahap stemming dilakukan proses transformasi data dengan *melakukan Term Frequency - Inverse Document Frequency* (TF-IDF). TF-IDF digunakan sebagai penghitungan bobot pada frekuensi kata dalam dataset. Hasil dari TF-IDF digunakan pada model ensemble yang digunakan.

```
(0, 48584)    0.05186195347310335
(0, 48252)    0.10871126031952381
(0, 47945)    0.12149689587959564
(0, 47754)    0.1164408763808501
(0, 46641)    0.09625043644532293
(0, 46275)    0.05405729194706496
(0, 44609)    0.07441479635127325
(0, 43888)    0.07823917338019379
```

Gambar 4. Contoh TF-IDF

Pada Gambar 4, sampel hasil TF – IDF menunjukkan nilai frekuensi dari kemunculkan masing masing kata pada dataset. Nilai TF-IDF yang dihasilkan akan digunakan pada *random forest classifier* untuk menentukan skor dalam evaluasi masing – masing pohon Keputusan. Berdasarkan skenario pengujian, evaluasi dilakukan pada pengaturan *hyperparameter* yang diterapkan pada fungsi *random forest*.

4. PEMBAHASAN

Tabel 3. Hasil Pengujian

Hyperparameter		Akurasi
N_estimator	**Max_depth**	
50	None	92,60%
	10	59,75%
	30	84,71%
	50	91,13%
100	None	92,71%
	10	63,17%
	30	84,62%
	50	91,20%
200	None	93,25%
	10	61,75%
	30	85,09%
	50	91,76%

Berdasarkan penelitian dari [26] dan [27] menunjukkan bahwa Random Forest menghasilkan akurasi terbaik dengan algoritma lainnya meskipun dalam beberapa hasil uji menunjukkan penurunan dari beberapa uji coba. Pada penelitian ini, uji coba dilakukan dalam beberapa kondisi untuk mengetahui tingkat akurasi algoritma Random Forest dalam Teknik bagging. Dengan pengaturan parameter tersebut dapat diketahui tingginya akurasi yang dihasilkan oleh algoritma Random Forest.

Uji coba pertama menggunakan n_estimators 50 dengan masing – masing pengaturan parameter max_depth adalah None, 10, 30, dan 50. Hasil uji menunjukkan bahwa nilai akurasi tertinggi pada max_depth None sebesar 92.60% dan nilai akurasi terendah pada max_depth 10 yaitu sebesar 59,75%. Adanya pembatasan di max_depth terlalu rendah juga dapat menyebabkan terjadinya underfitting.

Uji coba kedua menggunakan n_estimator 100 dengan pengaturan max_depth yang sama pada uji coba pertama. Hasil uji selaras dengan uji coba pertama yaitu nilai akurasi tertinggi pada max_depth None dan nilai akurasi terendah pada max_depth 10. Pada uji coba ini nilai akurasi mengalami peningkatan dibandingkan dengan ujicoba pertama. Nilai akurasi meningkat sebesar 0.11% yaitu menjadi 92.71%. meskipun peningkatan tidak terlalu besar atau signifikan, pengaturan hyperparameter terbukti memberikan peningkatan dalam akurasi klasifikasi.

Uji coba ketiga menggunakan n_estimator 200, hasilnya menunjukkan peningkatan yang lebih baik dari parameter sebelumnya. Dari masing – masing hasil nilai akurasi yang dihasilkan oleh masing – masing parameter menunjukkan peningkatan dari penggunaan parameter sebelumnya. Meskipun pada max_depth 10 nilai akurasinya lebih rendah dari uji coba sebelumnya.

Hasil evaluasi dari Tabel 3 menunjukkan bahwa akurasi terbaik pada n_estimators 200 max_depth None. Hasil akurasinya adalah 93,25%. Dan akurasi terendahnya terdapat di n estimator 50 dengan max_depth 10 dengan hasil akurasi adalah 61,75%. Dari hasil tersebut dapat ditarik kesimpulan bahwa semakin tinggi n_estimator dapat mempengaruhi hasil akurasi menjadi lebih tinggi.

Berdasarkan hasil uji coba ini, pengaturan hyperparameter max_depth perlu dipertimbangkan lagi dalam pemilihannya karena dapat menentukan perubahan akurasi yang signifikan. Pada max_depth None ke max_depth terjadi penurunan akurasi yang signifikan. Adanya batasan kedalaman sub pohon keputusan menjadi faktor pengurang nilai akurasi.

Tabel 4. Precision Recall

Label	Precision	Recall	F1-Score
0	0.94	0.92	0.93
1	0.96	0.97	0.96
2	0.95	0.86	0.90
3	0.90	0.95	0.93
Accuracy			0.93
Macro avg	0.94	0.92	0.93
Weighted avg	0.93	0.93	0.93

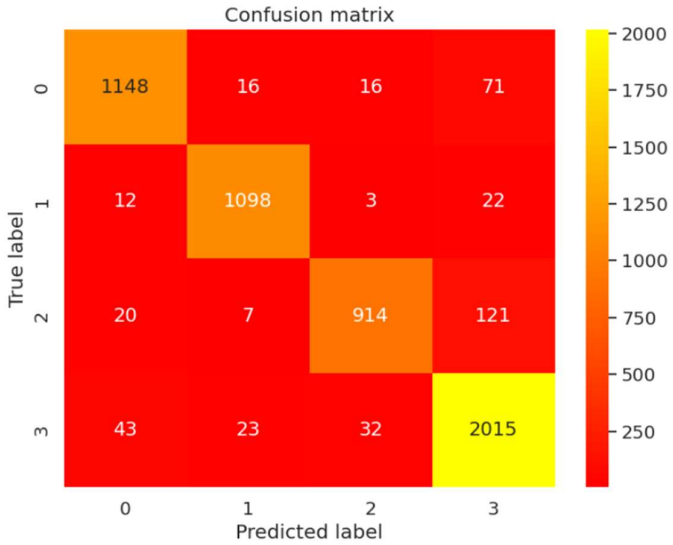
https://doi.org/10.25077/TEKNOSI.v10i1.2024.92-99

Pada Tabel 4, menunjukkan hasil analisis berdasarkan laporan pengujian precision, recall dan f1-score. Hasil dari tabel tersebut menunjukkan secara rata – rata precision menghasilkan 0.93 atau 93%. Precision merupakan perbandingan antara TP dengan TP + FP. Artinya dari hasil precision menunjukkan bahwa kategorisasi produk yang dilakukan tepat dalam mengidentifikasi kelas positif.

Untuk menentukan kategorisasi ulasan atau deskripsi sesuai dengan label menggunakan analisis recall. Hasil menunjukkan bahwa recall 0.93 atau 93% dapat menentukan deskripsi yang sesuai pada label.

Berdasarkan nilai f1-score, diperoleh nilai 0.93 atau 93% yang dapat diartikan bahwa 93% model yang diusulkan mampu mengkategorikan dan menemukan kelas positif. Model yang diusulkan tepat dalam menangani kasus kategorisasi produk.

Hasil evaluasi selanjutnya ditunjukkan pada grafik confusion matrix. Grafik confusion matrix memudahkan dalam analisis data berdasarkan hasil prediksi yang dihasilkan pada perhitungan dengan *random forest classifier.*



Gambar 5. Grafik Confusion Matrix

Berdasarkan Grafik yang ditampilkan pada gambar 6, pada diagonal utama menunjukkan TP dimana label yang diprediksi sesuai dengan labelnya. Sedangkan untuk FN terletak pada baris 0 dan dikolom 1, 2 atau kolom 3. Hasil menunjukkan bahwa nilai pada masing – masing posisi di diagonal utama menunjukkan nilai lebih besar dibandingkan nilai pada kolom yang lain. Hal ini menunjukkan bahwa akurasi yang ditampilkan pada confusion matrix adalah optimal untuk kategorisasi produk. Dapat disimpulkan bahwa model yang diusulkan tepat dan sesuai dalam menangani kasus kategorisasi produk berdasarkan ulasan atau deskripsi produk terhadap label yang sudah ditentukan.

Salah contoh Teknik bagging lainnya adalah Naive Bayes Classifier. Pada penelitian dilakukan ujicoba terhadap algoritma yang digunakan pada penelitian klasifikasi produk dengan Naïve Bayes [28]. Hasilnya berdasarkan Tabel 5 menunjukkan bahwa Random Forest terbukti menghasilkan akurasi lebih baik dari Naïve Bayes dengan akurasi yaitu 93.25%. implementasi Teknik bagging dan pengaturan hyperparameter pada *n_estimator* dan *max depth* memberikan hasil terbaik bagi algoritma Random Forest.

Tabel 5. Komparasi Algoritma dalam Teknik Bagging

Algoritma	Akurasi
Random Forest	93.25%
Naive Bayes	92.35%

5. KESIMPULAN

Permasalahan utama dalam e-commerce adalah bagaimana menempatkan deskripsi produk sesuai dengan label atau kategori yang ditentukan. Bagi pengguna awam, penempatan deskripsi produk sesuai dengan kategorisasinya membantu dalam pemilihan produk yang sesuai dengan kebutuhannya.

Penelitian ini mengusulkan kategorisasi produk dengan Teknik bagging dengan tujuan menemukan hasil atau model yang optimal dalam kategorisasi produk. Algoritma random forest mampu menangani permasalahan kategorisasi produk dalam jumlah data yang besar.

Setelah dilakukan pengujian, pada Teknik *Bagging* dengan algoritma *random forest*, diperoleh hasil akurasi 93.25% pada *n_estimators* 200 dan *max_depth* None. Hasil akurasi ini lebih tinggi dari pengaturan lainnya. Dari analisis precision, recall dan f1-score menunjukkan bahwa 0.93 atau 93% uji model mampu melakukan kategorisasi yang tepat pada kelas positive (TP). Selain itu, model yang diusulkan benar – benar tepat dalam melakukan kategorisasi produk berdasarkan deskripsi produk dengan label yang ditentukan. Dari hasil ini dapat disimpulkan bahwa teknik yang diusulkan mampu mengkategorisasikan produk e-commerce dengan tepat.

DAFTAR PUSTAKA

[1] H. Al Mashalah, E. Hassini, A. Gunasekaran, and D. Bhatt (Mishra), "The impact of digital transformation on supply chains through e-commerce: Literature review and a conceptual framework," *Transp Res E Logist Transp Rev*, vol. 165, p. 102837, Sep. 2022, doi: 10.1016/j.tre.2022.102837.

[2] R. E. Bawack, S. F. Wamba, K. D. A. Carillo, and S. Akter, "Artificial intelligence in E-Commerce: a bibliometric study and literature review," *Electronic Markets*, vol. 32, no. 1, pp. 297–338, Mar. 2022, doi: 10.1007/s12525-022-00537-z.

[3] M. Oase Ansharullah, W. Agustin, L. Lusiana, J. Junadhi, S. Erlinda, and F. Zoromi, "Product Classification Based on Categories and Customer Interests on the Shopee Marketplace Using the Naïve Bayes Method," *JAIA-Journal Of Artificial Intelligence And Applications*, vol. 2, no. 2, pp. 15–22, 2022.

[4] L. Donati, E. Iotti, G. Mordonini, and A. Prati, "Fashion Product Classification through Deep Learning and Computer Vision," *Applied Sciences*, vol. 9, no. 7, p. 1385, Apr. 2019, doi: 10.3390/app9071385.

[5] S. Suci Indasari and A. Tjahyanto, "Automatic Categorization of Multi Marketplace FMCGs Products using TF-IDF and PCA Features," *Jurnal SISFOKOM (Sistem Informasi dan Komputer)*, vol. 12, pp. 198–204, 2023.

[6] P. Ristoski, P. Petrovski, P. Mika, and H. Paulheim, "A machine learning approach for product matching and categorization," *Semant Web*, vol. 9, no. 5, pp. 707–728, Aug. 2018, doi: 10.3233/SW-180300.

[7] L. Tan, M. Y. Li, and S. Kok, "E-Commerce Product Categorization via Machine Translation," in *ACM Transactions on Management Information Systems*, Association for Computing Machinery, Aug. 2020. doi: 10.1145/3382189.

[8] S. Jain and V. Kumar, "Garment categorization using data mining techniques," *Symmetry (Basel)*, vol. 12, no. 6, Jun. 2020, doi: 10.3390/SYM12060984.

[9] H. Kim, G. Joo, and H. Im, "Product Category Classification using Word Embedding and GRUs," *The Journal of Korean Institute of Information Technology*, vol. 19, no. 4, pp. 11–18, Apr. 2021, doi: 10.14801/jkiit.2021.19.4.11.

[10] H. Jahanshahi *et al.*, "Text Classification for Predicting Multi-level Product Categories," Sep. 2021.

[11] R. Bruni and G. Bianchi, "Website categorization: A formal approach and robustness analysis in the case of e-commerce detection," *Expert Syst Appl*, vol. 142, p. 113001, Mar. 2020, doi: 10.1016/j.eswa.2019.113001.

[12] V. Gomero-Fanny, A. Ruiz Bengy, and L. Andrade-Arenas, "Prototype of Web System for Organizations Dedicated to e-Commerce under the SCRUM Methodology," 2021. [Online]. Available: www.ijacsa.thesai.org

[13] D. M. Alghazzawi, A. G. A. Alquraishee, S. K. Badri, and S. H. Hasan, "ERF-XGB: Ensemble Random Forest-Based XG Boost for Accurate Prediction and Classification of E-Commerce Product Review," *Sustainability (Switzerland)*, vol. 15, no. 9, May 2023, doi: 10.3390/su15097076.

[14] P. Kalaivani, "Machine Learning Approach to Analyse Ensemble Models and Neural Network Model for E-Commerce Application," *Indian J Sci Technol*, vol. 13, no. 28, pp. 2849–2857, Jul. 2020, doi: 10.17485/IJST/v13i28.927.

[15] M. Pawłowski, "Machine Learning Based Product Classification for eCommerce," *Journal of Computer Information Systems*, vol. 62, no. 4, pp. 730–739, 2022, doi: 10.1080/08874417.2021.1910880.

[16] K. POTHUGANTI, "Open-World Classification Algorithm to Product Identification," *SSRN Electronic Journal*, 2019, doi: 10.2139/ssrn.3719055.

[17] B. Sun, H. Chen, J. Wang, and H. Xie, "Evolutionary under-sampling based bagging ensemble method for imbalanced data classification," *Front Comput Sci*, vol. 12, no. 2, pp. 331–350, Apr. 2018, doi: 10.1007/s11704-016-5306-z.

[18] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, "A survey on ensemble learning," *Front Comput Sci*, vol. 14, no. 2, pp. 241–258, Apr. 2020, doi: 10.1007/s11704-019-8208-z.

[19] Q.-F. Li and Z.-M. Song, "High-performance concrete strength prediction based on ensemble learning," *Constr Build Mater*, vol. 324, p. 126694, Mar. 2022, doi: 10.1016/j.conbuildmat.2022.126694.

[20] Y. Liu, "High-Performance Concrete Strength Prediction Based on Machine Learning," *Comput Intell Neurosci*, vol. 2022, pp. 1–7, May 2022, doi: 10.1155/2022/5802217.

[21] G. Ngo, R. Beard, and R. Chandra, "Evolutionary bagging for ensemble learning," *Neurocomputing*, vol. 510, pp. 1–14, Oct. 2022, doi: 10.1016/j.neucom.2022.08.055.

[22] Q.-F. Li and Z.-M. Song, "High-performance concrete strength prediction based on ensemble learning," *Constr Build Mater*, vol. 324, p. 126694, Mar. 2022, doi: 10.1016/j.conbuildmat.2022.126694.

[23] B. Sun, H. Chen, J. Wang, and H. Xie, "Evolutionary under-sampling based bagging ensemble method for imbalanced data classification," *Front Comput Sci*, vol. 12, no. 2, pp. 331–350, Apr. 2018, doi: 10.1007/s11704-016-5306-z.

[24] P. D. Caie, N. Dimitriou, and O. Arandjelović, "Precision medicine in digital pathology via image analysis and machine learning," in *Artificial Intelligence and Deep Learning in Pathology*, Elsevier, 2021, pp. 149–173. doi: 10.1016/B978-0-323-67538-3.00008-7.

[25] S. Hong and H. S. Lynn, "Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction," *BMC Med Res Methodol*, vol. 20, no. 1, p. 199, Dec. 2020, doi: 10.1186/s12874-020-01080-1.

[26] S. Amien, P. Perdana, T. Bharata Aji, and R. Ferdiana, "Aspect Category Classification dengan Pendekatan Machine Learning Menggunakan Dataset Bahasa Indonesia (Aspect Category Classification with Machine

Learning Approach Using Indonesian Language Dataset)," 2021.

[27] S. Jain and V. Kumar, "Garment categorization using data mining techniques," *Symmetry (Basel)*, vol. 12, no. 6, Jun. 2020, doi: 10.3390/SYM12060984.

[28] M. Oase Ansharullah, W. Agustin, L. Lusiana, J. Junadhi, S. Erlinda, and F. Zoromi, "Product Classification Based on Categories and Customer Interests on the Shopee Marketplace Using the Naïve Bayes Method," *JAIA-Journal Of Artificial Intelligence And Applications*, vol. 2, no. 2, pp. 15–22, 2022.

BIODATA PENULIS

Faskal Churniansyah
Biodata penulis, dituliskan secara singkat (maksimal 100 kata) dan boleh disertai dengan foto. Tinggi maksimal foto adalah 2,5 cm dengan lebar menyesuaikan.

Danang Wahyu Utomo
Dosen pada program studi Teknik informatika Fakultas Ilmu Komputer Universitas Dian Nuswantoro, Semarang. Bidang kajian penelitian adalah rekayasa perangkat lunak, pengolahan citra digital, data mining, dan kecerdasan buatan khusus kesehatan