JEPIN
(Jurnal Edukasi dan Penelitian Informatika)
ISSN(e): 2548-9364 / ISSN(p) : 2460-0741

Vol. 10
No. 1
April 2024

Perbandingan Algoritma Pohon dengan Beberapa Skenario Pelabelan untuk Analisis Sentimen pada Aplikasi Milik Pemerintah/BUMN

Silmi Annisa Rizki Manaf[#1], Anwar Fitrianto[#2], Agus Mohamad Soleh[#3]

[#]*Departemen Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, IPB University*
Jalan Raya Dramaga, Gedung Sekolah Pascasarjana IPB, Kampus IPB Dramaga Bogor, 16680 - Jawa Barat, Indonesia

[1]silmiannisa@apps.ipb.ac.id
[2]anwarstat@gmail.com
[3]agusms@apps.ipb.ac.id

***Abstrak*— Berkembangnya era digitalisasi mengakibatkan banyaknya inovasi yang diupayakan untuk mempermudah aktivitas masyarakat di berbagai bidang, salah satunya yaitu adanya aplikasi yang menunjang agar menjadi lebih efisien dan dapat diakses dari mana saja. Aplikasi milik pemerintah dan BUMN sebagai perusahaan berskala nasional cenderung belum banyak diketahui dan banyak yang memiliki *rating* rendah disertai dengan berbagai macam ulasan pengguna aplikasi. Analisis sentimen merupakan analisis yang cocok untuk menganalisis ulasan dari aplikasi yang dipilih. Data yang digunakan adalah ulasan aplikasi InfoBMKG, BPOM *Mobile, MyIndihome,* dan *MyPertamina*. Penelitian bertujuan untuk membandingkan performa algoritma *double random forest* dan algoritma berbasis pohon lain yaitu *decision tree, extra trees,* dan *random forest* berdasarkan tingkat ketepatan performa akurasi model. Pelabelan data berdasarkan *rating* aplikasi, *lexicon-based,* dan *sentiment scoring* dengan peubah prediktor dihasilkan dari tokenisasi unigram yang diberi bobot dengan TF-IDF. Setiap observasi data dikategorikan ke dalam kelas positif, netral, dan negatif. Hasil penelitian menunjukkan algoritma *extra trees* dan metode pelabelan *sentiment scoring* mampu menghasilkan performa terbaik dengan nilai rata-rata akurasi mencapai 80 – 84% pada tiap aplikasi yang dipilih.**

Kata kunci*— *Analisis Sentimen, Decision Tree, Double Random Forest, Extra Trees, Klasifikasi, Random Forest

I. PENDAHULUAN

Semakin berkembangnya era digitalisasi, banyak upaya yang dilakukan untuk mempermudah aktivitas masyarakat di berbagai bidang, salah satunya dengan adanya aplikasi. Hasil Survei Penetrasi dan Perilaku Internet menunjukkan jumlah penduduk Indonesia yang telah terkoneksi internet mencapai 78,19% atau naik 1,17% dari tahun 2022 [1]. Banyaknya pengguna internet membuktikan teknologi dan internet berkembang cukup cepat [2]. Hal ini tentunya berpengaruh pada penggunaan aplikasi dalam kehidupan sehari-hari. Adanya aplikasi menjadi lebih mudah, praktik, efisien, dan dapat di akses dari mana saja.

Penelusuran terhadap beberapa aplikasi yang dilakukan pada Desember 2022 dan didukung dari berita yang beredar, aplikasi milik perusahaan swasta cenderung lebih unggul berdasarkan *rating*. Tidak jarang aplikasi tersebut sudah cukup banyak dikenal dan memiliki penilaian baik oleh masyarakat. Aplikasi milik pemerintah atau BUMN sebagai instansi nasional cenderung belum banyak diketahui dan banyak yang memiliki *rating* rendah disertai dengan berbagai macam ulasan pengguna aplikasi sehingga dilakukan penelitian dari ulasan aplikasi yang dipilih. Analisis yang cocok digunakan adalah analisis sentimen.

Sentimen diekspresikan sebagai emosi, padangan orang, atau visi terhadap suatu isu. Analisis sentimen (*sentiment analysis*) merupakan proses identifikasi atau kategorisasi opini, emosi, sikap, penilaian acara, pelayanan jasa, atau suatu permasalahan produk dari individu ke dalam kelas kategori tertentu [3]. Penerapan sentimen pada web dapat digunakan untuk mengekspresikan opini dari khalayak dalam bentuk komentar atau status tentang berbagai topik sebagai bentuk tanggapan mengenai suatu produk secara spesifik, mempromosikan barang atau jasa, menganalisis politik pada pemilu, atau mengevaluasi kinerja pelayanan jasa yang terjadi pada suatu aplikasi [4]. Tujuan penting analisis sentimen adalah untuk mengumpulkan informasi kemudian mengklasifikasikannya dan menganalisis ulasan terkait untuk diungkapkan dalam bentuk teks terkategori menjadi kelas positif, netral, atau negatif [4].

Ulasan aplikasi yang diunduh dari *Google Play Store* dalam bentuk penilaian ataupun pendapat pengguna dapat diolah menggunakan analisis sentimen dengan berbagai algoritma statistika. Hal ini tentunya menunjukkan bahwa analisis sentimen dapat digunakan untuk mengiringi ilmu dan teknologi informasi yang terus berkembang, serta mampu berperan dalam keilmuan statistika. Setiap ulasan aplikasi memiliki penilaian bintang 1 – 5 namun tidak

jarang ditemukan kondisi yang tidak dapat dihindari dari pengguna aplikasi seperti pengguna aplikasi memberikan penilaian yang tidak sesuai dengan ulasan. Ulasan baik, memberikan kritik dan saran yang mendukung namun penilaian yang diberikan rendah, atau sebaliknya, ulasan yang buruk namun memberikan penilaian dengan bintang tinggi. Hal ini kurang dapat menggambarkan kualitas dari aplikasi [5]. Untuk mengatasi hal tersebut maka dilakukan skenario dengan metode pelabelan dalam mengategorikan ulasan dari pengguna aplikasi. Salah satu pendekatan yang bisa digunakan analisis sentimen adalah klasifikasi dengan pembelajaran mesin (*machine learning*) seperti *support vector machine* (SVM), Naive Bayes (NB), *decision tree* (DT), *logistic regression*, *K-Nearest Neighbor* (KNN), *ranfom forest* (RF), dan sebagainya [6]. Algoritma DT, RF, *extra trees* (ET) merupakan algoritma berbasis pohon yang telah banyak digunakan pada analisis sentimen dan algoritma ini terus berkembang dengan perubahan yang berbeda-beda sehingga diharapkan dapat memberikan performa model yang baik untuk penelitian. Penggunaan model komputer dalam analisis sentimen digunakan untuk mencari model dengan performa terbaik. Dengan begitu, tingkat kesalahan dalam memprediksi data rendah dan hal ini menunjukkan bahwa algoritma tertentu mampu dalam mengklasifikasikan data penelitian dengan baik.

Penelitian yang dilakukan oleh [7] menggunakan data twitter untuk mengetahui sentimen publik mengenai kasus protes petani di India dengan membandingkan dua metode ekstraksi fitur *bag of words* (BoW) dan *term frequency-inverse document frequency* (TF-IDF). Penelitian dengan menerapkan empat algoritma klasifikasi, yakni NB, DT, RF, dan SVM. Hasilnya ditemukan bahwa BoW memiliki kinerja yang lebih baik daripada TF-IDF dengan algoritma RF memberikan hasil akurasi klasifikasi tertinggi (± 95%). Penelitian lainnya dilakukan oleh [8] yaitu mencari model terbaik untuk analisis sentimen mengenai komentar video youtube Indonesia tentang pelayanan pemerintah yang berhubungan dengan pandemi covid-19. Penelitian tersebut menggunakan lima algoritma yakni NB, SVM, DT, RF, dan ET dengan menguji beberapa tahapan penyiapan data. Hasil penelitian menunjukkan ET mampu menghasilkan akurasi maksimum hingga 89.68%.

Salah satu algoritma baru berbasis pohon, yakni *double random forest* (DRF) menggunakan *bootstrap* pada setiap node selama proses pembentukan pohon. Algoritma ini merupakan algoritma improvisasi dari RF ketika data RF mengalami *underfit* [9]. Penelitian lain tidak banyak yang membahas penggunaan DRF dan menerapkannya pada analisis sentimen. Adanya perkembangan berbasis pohon oleh [9], akan dilakukan penelitian untuk membandingkan keempat algoritma berbasis pohon, DT, RF, ET, DRF pada analisis sentimen. Sehingga dengan demikian, tujuan dari penelitian ini adalah untuk membandingkan performa klasifikasi algoritma DRF dan algoritma berbasis pohon lainnya dalam hal ini adalah DT, RF, ET berdasarkan tingkat ketepatan performa model dengan nilai akurasi. Aplikasi yang dipilih dalam penelitian ini berdasarkan kriteria pilihan peneliti dan diharapkan dapat menjadi perbaikan untuk pengembang aplikasi sedangkan algoritma yang terbaik dalam penelitian ini dapat menjadi inovasi untuk penelitian selanjutnya.

II. TINJAUAN PUSTAKA

A. Decision Tree

Decision tree (DT) merupakan algoritma berbasis pohon yang sederhana karena pohon yang terbentuk hanya pohon tunggal. Pohon keputusan algoritma DT mulai dari simpul akar memisahkan node berlanjut secara rekursif hingga mencapai simpul daun dengan label kelas tertentu [10]. Setiap node ada tahap pemisahan untuk memutuskan nilai masukkan (*input*) ke dalam subpohon kanan atau kiri. Pohon keputusan DT mampu menghasilkan pohon yang panjang dan DT dapat melakukan pemangkasan (*pruning*) jika pohon yang terbentuk sudah terlalu besar [11]. Proses identifikasi peubah prediktor serta nilai batas pemisah dalam penelitian ini menggunakan nilai *gini index* yang dapat diperoleh dari persamaan berikut [12] [13].

$$Gini\ (D) = 1 - \sum_{i=1}^{n} {p_i}^2 \quad (1)$$

$$Gini\ index\ (D,k) = \sum_{v=1}^{V} \frac{|D^v|}{|D|}\ (Gini\ (D^v)) \quad (2)$$

Setiap node akan memilih peubah dengan nilai pemisah berdasarkan nilai *gini index* minimum. Selain nilai *gini index* dapat menggunakan nilai *gain ratio, information gain,* dan *misclassification rate* [14]. Iterasi pohon akan berhenti ketika mengalami kondisi seperti sudah tidak ada peubah yang dapat digunakan, jumlah amatan dalam sub pohon sudah terlalu sedikit, atau salah satu sub pohon nilainya sudah homogen sehingga tidak dapat melanjutkan pemisah berikutnya [11].

B. Random Forest

Random forest (RF) merupakan algoritma klasifikasi *ensemble* yang menggunakan pohon keputusan sebagai konsep dasar pengklasifikasian. Algoritma RF merupakan metode pengembangan dari teknik *bagging* (*bootstrap and aggregating*) [15]. Algoritma RF disebut dengan hutan acak karena terbentuk dari kombinasi pohon yang setiap pohon bergantung pada nilai acak dan saling bebas [16]. Kelebihan RF cocok untuk data berdimensi sedang hingga besar yakni ketika jumlah peubah prediktor lebih besar dari banyaknya observasi karena RF bekerja dengan tidak menggunakan semua peubah prediktor sekaligus [17]. Semakin sedikit peubah yang digunakan, menyebabkan korelasi antar pohon menjadi lebih rendah. Terjadinya hal tersebut diharapkan akan menghasilkan akurasi prediksi akhir yang lebih baik [9].

C. Extremely Randomized Trees

Sesuai namanya, algoritma *extra trees* (ET) merupakan algoritma berbasis pohon dengan adanya pengacakan yang ekstrim saat pemilihan peubah pemisah dan titik potong. Algoritma ET akan membagi node dengan memilih titik

potong sepenuhnya secara acak dan menggunakan seluruh sampel data (tidak hasil *bootstrap*) untuk membangun pohon [18]. Situasi yang ekstrim akan menciptakan pohon yang sangat teracak dan struktur yang tidak bergantung satu sama lain. Pengacakan penuh saat membentuk pohon dari pemilihan titik potong dan pemilihan peubah mampu mengurangi keanekaragaman serta penggunaan seluruh sampel data mampu mengurangi bias [19]. Salah satu kelebihan ET yakni mampu bekerja mirip dengan RF tapi jauh lebih cepat dan mampu memberikan hasil yang lebih baik. Hal ini karena adanya pengacakan pada ET [20].

D. Double Random Forest

Double random forest (DRF) dikembangkan dari RF oleh [9] ketika data RF mengalami *underfit* [9]. Berbeda dengan RF, DRF menggunakan seluruh sampel data latih untuk membangun pohon keputusan. Seluruh pohon yang dibentuk pada DRF berasal dari data yang sama akan menghasilkan contoh unik yang banyak. Semakin banyak contoh unik yang dimiliki, akan semakin besar pohon yang dihasilkan [9]. Adanya DRF diharapkan dapat menghasilkan pohon keputusan yang lebih besar dan bisa meningkatkan performa yang maksimum sesuai dengan ukuran node minimum [15]. DRF melakukan *bootstrap* pada setiap node untuk mendapatkan pemisah terbaik dan kondisi lain ketika tidak memerlukan *bootstrap* yakni saat suatu node memiliki sampel kurang dari 10% dari total sampel. Setiap hasil prediksi suatu pohon akan digabungkan dengan menggunakan *majority vote* [9].

E. Term Frequency-Inverse Document Frequency

Term frequency-inverse document frequency (TF-IDF) dilakukan sebagai proses ekstraksi peubah untuk memberi bobot pada tiap kata dan mampu mencerminkan seberapa penting suatu kata bagi suatu dokumen. Kata yang sering muncul akan terhitung dari TF dan menghindari kata yang tidak penting yang muncul di setiap dokumen dengan IDF [21]. TF akan mewakili frekuensi suatu kata dalam suatu dokumen dan IDF menunjukkan jumlah dokumen dimana suatu kata tersebut ditemukan [22]. Perhitungan TF-IDF diperoleh melalui persamaan berikut.

$$tf_{t,d} = \frac{n_{t,d}}{Banyak\ kata\ dalam\ dokumen} \quad (3)$$

$$idf_d = log\left(\frac{N}{DF_t}\right) \quad (4)$$

$$tfidf_{t,d} = tf_{t,d} \times idf_d \quad (5)$$

Istilah “dokumen” merujuk sebagai pada data ulasan.

F. Evaluasi Performa Model Klasifikasi

Ukuran performa model dibutuhkan ketika tujuan dari suatu penelitian ingin mengevaluasi dan membandingkan model klasifikasi yang berbeda. Evaluasi performa model menggunakan matriks konfusi untuk menangkap semua informasi yang relevan terkait kinerja model [23]. Ukuran performa yang digunakan dalam penelitian ini adalah nilai akurasi. Akurasi dapat merepresentasi rasio antara yang diprediksi benar dan semua contoh dalam gugus data [24]. Nilainya berada dalam rentang 1.0 (sempurna) dan nilai kesalahan terburuk adalah 0.0 dengan perhitungan berikut.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

III. METODOLOGI

A. Data

Data dalam penelitian ini menggunakan hasil penarikan dari *Google Play Store* yang dipilih empat aplikasi untuk dikaji menggunakan keempat algoritma berbasis pohon. Empat aplikasi dipilih dari dua aplikasi milik pemerintah dan dua aplikasi BUMN. Pilihan aplikasi didasarkan dari dua kriteria yakni berdasarkan *rating* aplikasi di *Google Play* dan keragaman ulasan yang diberikan oleh pengguna. *Rating* aplikasi yang dipilih yaitu aplikasi yang terkategori tinggi (≥ bintang 4) dan terkategori rendah (< bintang 4). Tabel 1 merupakan aplikasi pilihan yang akan dilakukan analisis sentimen. Penilaian aplikasi disesuaikan pada saat penarikan data pada tanggal 5 Agustus 2023.

TABEL I
APLIKASI PILIHAN

Perusahaan	Nama Aplikasi	Penilaian (*rating*) aplikasi (1 – 5)
Pemerintah	Info BMKG	4.4
	BPOM Mobile	3.4
BUMN	MyIndihome	4.2
	MyPertamina	2.9

Tabel 2 menujukkan peubah yang akan digunakan dalam pemodelan. Data teks yang digunakan merupakan data teks dalam Bahasa Indonesia.

TABEL II
PEUBAH YANG DIGUNAKAN

Peubah	Keterangan Peubah
Y	Sentimen pada setiap ulasan 1 = Sentimen positif 2 = Sentimen netral 3 = Sentimen negatif
X_i	Bobot peubah ke-*i* untuk setiap token hasil tokenisasi dan pembobotan dengan TF-IDF *i* = 1, 2, 3, …, *k*; *k* = banyaknya token

B. Pengumpulan Data

Penarikan data dari *Google Play Store* menghasilkan data ulasan pada setiap aplikasi. Adapun informasi yang diperoleh adalah ID ulasan, nama pengguna (*username*), penilaian aplikasi, dan waktu unggah ulasan. Informasi dari aplikasi yang digunakan diambil berdasarkan ulasan paling relevan (*most relevan*) dan ulasan terbaru (*newest*). Penelitian ini berisi data dengan ID ulasan unik sehingga ulasan yang terunggah lebih dari satu kali akan dihapus tanpa menghapus ulasan pertama. Hal ini dilakukan untuk menghindari adanya duplikasi data. Banyaknya ulasan pada Tabel 3 merupakan ulasan setelah penghapusan duplikasi data.

TABEL III
BANYAKNYA ULASAN PADA SETIAP APLIKASI

Aplikasi	Banyaknya Ulasan
InfoBMKG	1020
BPOM *Mobile*	459
MyIndihome	1044
MyPertamina	1073

A. *Penyiapan Data*

Data yang digunakan dalam penelitian merupakan data teks yang sulit dibaca oleh sistem (bentuknya tidak teratur) sehingga perlu dilakukan pembersihan dan penyiapan data untuk mengubah data teks menjadi bentuk data numerik agar dapat terbaca dan mudah diproses oleh sistem. Tahap penyiapan data yang dilakukan adalah mengubah bentuk huruf kapital menjadi huruf kecil (*case folding*) agar lebih seragam, membersihkan dokumen teks dari karakter dan simbol seperti tautan, angka, *emoticon*, tagar, *enter*, spasi berlebih, dan karakter lain (*text cleans up*), mengubah kata singkatan (*slang words*) menjadi kata baku (*normalization words*), menghilangkan kata yang tidak memiliki makna (*stop words removal*), dan menghilangkan imbuhan baik di awalan, akhiran, atau keduanya (*stemming*).

TABEL IV
CONTOH ULASAN SEBELUM DAN SESUDAH PENYIAPAN DATA

Sebelum Penyiapan Data
Kadang bisa dibuka, kadang susah, Sering pake kode QR hasil tidak terdaftar. Jadi membuat hati jengkel
Sesudah Penyiapan Data
kadang buka kadang susah pakai kode qr hasil daftar hati jengkel

Sebelum Penyiapan Data
Semua produk yg ada d rumah d cobain 1 1 tidak terdaftar semua. Udah suudzon aja coba tkut produk palsu padahal aplikasinya yg error. Kasian sama penjualnya udah d komplen produk palsu
Sesudah Penyiapan Data
produk rumah coba daftar suudzon coba takut produk palsu aplikasi eror kasihan jual komplain produk palsu

Bahasa Indonesia memiliki beberapa bentuk sehari-hari yang dapat ditemukan dalam Bahasa lain. Penelitian ini menggunakan kamus leksikon khusus Bahasa Indonesia yang dapat mengobservasi tren kata-kata gaul tersebut kemudian dinormalisasi menjadi bentuk baku [25]. Kamus yang digunakan pada *stop words removal* menggunakan bantuan dari *Indonesian stop words collection*.

TABEL V
BANYAKNYA ULASAN SESUDAH TAHAPAN PENYIAPAN DATA

Aplikasi	Banyaknya Ulasan	
	Sebelum	**Sesudah**
InfoBMKG	1020	992
BPOM *Mobile*	459	451
MyIndihome	1044	1035
MyPertamina	1073	1053

B. *Pelabelan Sentimen*

Setiap ulasan menjadi observasi yang dikategorikan ke dalam kelas pelabelan positif, netral, dan negatif. Proses pelabelan sentimen pada penelitian ini membandingkan 3 metode pelabelan, yakni:

1) Pelabelan pertama (1) berdasarkan *rating* dari setiap pengguna aplikasi dengan ketentuan bintang 1 – 2 akan terkategori label negatif, bintang 3 label netral, dan bintang 4 – 5 terkategori label positif.
2) Pelabelan kedua (2) dengan *sentiment scoring* yakni pelabelan yang memperhitungkan setiap kata dalam ulasan berdasarkan kata sentimen, *booster words*, dan kata negasi yang terkandung dalam kalimat [26]. Skor sentimen dihitung dengan menjumlahkan keseluruhan skor yang terdapat pada kalimat.
3) Pelabelan ketiga (3) dengan *lexicon-based* yakni setiap kata pada ulasan akan mendapatkan bobot jika terdapat kata sentimen positif dan sentimen negatif yang ada di dalam kamus [27]. Perhitungan sebagai berikut.

$$Skor = \sum kata\ positif + \sum kata\ negatif \quad (7)$$

Ulasan yang memiliki skor sentimen kurang dari 0 akan terkategori label **negatif**, skor sentimen sama dengan 0 terkategori label **netral**, dan ulasan yang memiliki skor sentimen lebih dari 0 terkategori label **positif**.

$$Skor\ sentimen \begin{cases} Y = 1, & Positif, skor\ sentimen > 0 \\ Y = 2, & Netral, skor\ sentimen = 0 \\ Y = 3, & Negatif, skor\ sentimen < 0 \end{cases}$$

Setiap kata pada kata sentimen, *booster words*, kata negasi, kata positif, dan kata negatif yang digunakan dalam metode pelabelan didasarkan dari kamus pendukung hasil pengembangan penelitian oleh [26]. Pelabelan sentimen menjadi peubah respon Y yang digunakan dalam analisis.

TABEL VI
JUMLAH DATA PADA SETIAP KELAS KATEGORI

Aplikasi	Kelas Kategori	Pelabelan		
		(1)	**(2)**	**(3)**
InfoBMKG	$Y = 1$	**349**	**459**	215
	$Y = 2$	324	303	376
	$Y = 3$	319	230	**401**
BPOM *Mobile*	$Y = 1$	**209**	**200**	106
	$Y = 2$	39	168	**235**
	$Y = 3$	203	83	110
MyIndihome	$Y = 1$	**370**	**449**	300
	$Y = 2$	308	304	**381**
	$Y = 3$	357	282	354
MyPertamina	$Y = 1$	**364**	**443**	214
	$Y = 2$	351	338	**445**
	$Y = 3$	338	272	394

Tabel 6 menunjukkan banyaknya observasi atau jumlah data untuk setiap kelas kategori, setiap metode pelabelan, dan pada setiap aplikasi. Hasil pelabelan sentimen untuk setiap kelas kategori menunjukkan banyaknya observasi cenderung lebih banyak mengarah ke dalam label positif

untuk pelabelan (1) dan (2). Pelabelan (3) cenderung lebih banyak pada kelas label netral untuk 3 aplikasi dan label negatif untuk aplikasi InfoBMKG. Tidak dilakukannya penanganan ketakseimbangan data karena keputusan data tak seimbang bergantung pada karakteristik khusus untuk setiap masalah dan diharapkan hal ini dapat menunjukkan distribusi alami untuk setiap peubah respon dari data.

Beberapa metode pelabelan dilakukan guna mengatasi kondisi yang tidak dapat dihindari yang telah disampaikan pada pendahuluan. Penelitian ini menambahkan pelabelan validasi untuk mengidentifikasi hasil ketiga pelabelan lain sehingga dapat meningkatkan keakuratan hasil sentimen dan ulasan. Pelabelan validasi (4) dilakukan secara manual dan subjektif peneliti yakni membandingkan hasil ketiga pelabelan sebelumnya. Ketika ketiga pelabelan tersebut memberikan sentimen yang berbeda pada suatu ulasan, maka akan dilakukan pengecekkan ulang dan memberikan sentimen baru yang sesuai terhadap ulasan tersebut.

TABEL VII
JUMLAH DATA SETIAP KATEGORI DENGAN PELABELAN VALIDASI

Aplikasi	Kelas Kategori			Total Observasi
	$Y = 1$	$Y = 2$	$Y = 3$	
InfoBMKG	337	122	**533**	992
BPOM *Mobile*	**214**	30	207	451
MyIndihome	346	140	**549**	1035
MyPertamina	288	198	**567**	1053

Berbeda dengan pelabelan (1), (2), dan (3), hasil pelabelan dengan pelabelan (4) cenderung banyak berisi sentimen negatif dan netral. Hal ini mungkin terjadi karena dalam prosesnya membaca ulasan secara subjektivitas.

C. *Tokenisasi dan Pembobotan TF-IDF*

Peubah prediktor didapatkan dari token hasil tokenisasi unigram dan melalui proses TF-IDF. Tokenisasi adalah proses segmentasi atau pemisahan kata dalam teks ke unit yang lebih kecil. Proses tokenisasi membagi teks menjadi potongan kata dengan memisahkan berdasarkan spasi [28]. Tokenisasi unigram artinya potongan kata yang digunakan hanya satu kata. Setiap kata memiliki bobot nilai hasil dari TF-IDF agar dapat dilakukan pemodelan.

TABEL VIII
CONTOH HASIL TOKENISASI DAN PEMBOBOTAN TF-IDF

Y	aktif	aplikasi	baik	…	warga
1	0	0.4438	0.2365	…	0
3	0	0.0771	0	…	0.2731

Tabel 8 contoh bentuk data yang akan digunakan untuk analisis. Ulasan pertama sebagai observasi pertama artinya tidak mengandung kata “aktif” dan kata “warga” sehingga bobotnya 0. Ulasan kedua mengandung kata “aplikasi” dengan bobot 0.0771 dan kata “warga” 0.2731, artinya kedua kata tersebut terkandung pada ulasan kedua yang bersentimen negatif. Kata berbobot 0 dapat disebabkan karena hasil salah satu TF atau IDF yang bernilai 0 akibat berbagai macam faktor.

D. *Pemilihan Peubah*

Data teks sesudah melalui tahap tokenisasi menghasilkan banyak peubah prediktor. Penelitian ini menggunakan dua rangkaian pemilihan peubah, secara manual dan dengan metode *information gain* (IG). Setiap kata akan dihitung nilai IG. Pemilihan peubah berdasarkan nilai IG terbesar, artinya peubah tersebut merupakan peubah yang paling signifikan [14]. Pemilihan peubah dengan IG dipilih ketika suatu kata memiliki nilai IG yang lebih besar dari batas nilai yang ditentukan. Perhitungan nilai IG sebagai berikut.

$$IG = Entropy\,(L) - \sum_{v=1}^{V} \frac{|L_i|}{|L|}\,(Entropy\,(L_i)) \quad (8)$$

$$Entropy\,(L) = -\sum_{i=1}^{j} p_i log_2(p_i) \quad (9)$$

Pemilihan peubah secara manual dilakukan dengan menelusuri tiap kata yang tidak sesuai dan memperbaiki kata tersebut pada ulasan. Hal ini akan mengurangi dan menghilangkan kata yang salah dan tidak sesuai sehingga berdampak berkurangnya kata atau peubah yang akan digunakan. Pemilihan peubah secara manual berkurang hingga 747 token pada InfoBMKG dan 175 token pada BPOM *Mobile* sedangkan dengan metode IG, peubah yang berkurang cukup ekstrem.

TABEL IX
BANYAKNYA TOKEN HASIL TOKENISASI

Aplikasi	Hasil Tokenisasi	Pemilihan Peubah	
		Manual	IG
InfoBMKG	1803	1056	267
BPOM *Mobile*	708	533	160
MyIndihome	1896	-	-
MyPertamina	2232	-	-

Pemilihan peubah dilakukan hanya pada aplikasi BPOM *Mobile* dan InfoBMKG dikarenakan meninjau hasil dari pemilihan peubah yang telah dilakukan.

E. *Pemodelan*

Data dibagi menjadi dua bagian, data latih dan data uji. Data latih akan digunakan untuk membangun pohon dan data uji digunakan untuk mengevaluasi performa model. Pada pemodelan dilakukan proses validasi silang dengan teknik *k-folds cross validation* dengan k = 10.

	Gugus Data									
	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10
Iterasi 1	Test	Train	Train	Train	Train	Train	Train	Train	Train	Train
Iterasi 2	Train	Test	Train	Train	Train	Train	Train	Train	Train	Train
Iterasi 3	Train	Train	Test	Train	Train	Train	Train	Train	Train	Train
Iterasi 4	Train	Train	Train	Test	Train	Train	Train	Train	Train	Train
Iterasi 5	Train	Train	Train	Train	Test	Train	Train	Train	Train	Train
Iterasi 6	Train	Train	Train	Train	Train	Test	Train	Train	Train	Train
Iterasi 7	Train	Train	Train	Train	Train	Train	Test	Train	Train	Train
Iterasi 8	Train	Train	Train	Train	Train	Train	Train	Test	Train	Train
Iterasi 9	Train	Train	Train	Train	Train	Train	Train	Train	Test	Train
Iterasi 10	Train	Train	Train	Train	Train	Train	Train	Train	Train	Test

Gambar. 1 Ilustrasi 10-*folds cross validation*

Pendekatan ini akan melakukan pengacakan pada data dan membagi data menjadi k (10) bagian yang sama. Sebagian besar data (90%) akan menjadi data latih dan

sisanya akan diuji untuk mengetahui performa model yang dihasilkan [29], adanya hal ini artinya setiap data memiliki peluang yang sama untuk di validasi. Pemilihan nilai k = 10 dapat menunjukkan performoa model yang lebih efisien secara komputasi dan mampu meningkatkan ketepatan estimasi secara efektif dengan tetap mempertahankan bias atau kesalahan prediksi yang rendah [30].

IV. Hasil dan Pembahasan

Model pohon yang telah dibentuk menggunakan data latih dan diterapkan validasi silang, 10-*folds cross validation* selanjutnya dievaluasi dengan data uji. Setiap *folds* akan menghasilkan nilai akurasi sehingga sebagai nilai perbandingan antar algoritma, hasil akhir pemodelan diperoleh berdasarkan rata-rata akurasi dari kesepuluh *fold*.

TABEL X
RATA-RATA NILAI AKURASI (%) MODEL DENGAN PEMILIHAN PEUBAH

Algoritma	InfoBMKG		
	Tanpa	Manual	IG
RF	80.55	80.25	71,91
DRF	81.24	80.85	**72,60**
ET	**85.71**	**84.37**	70,51
DT	78.85	77.02	70,02

Algoritma	BPOM *Mobile*		
	Tanpa	Manual	IG
RF	80.62	80.06	75,12
DRF	81.07	79.84	**75,55**
ET	**83.94**	**83.40**	74,93
DT	77.53	77.83	71,82

Tabel 10 merupakan pemodelan untuk pemilihan peubah terhadap InfoBMKG dan BPOM *Mobile* dengan peubah respon yang digunakan merupakan hasil dari pelabelan (2).

Pemilihan peubah secara manual cenderung memakan waktu yang lama karena prosesnya dalam memperbaiki kata, namun hasil token yang digunakan dapat terjamin baik. Meninjau hasil rata-rata akurasi dengan pemilihan peubah secara manual, hasilnya tidak jauh berbeda dengan model tanpa pemilihan peubah. Pemilihan peubah dengan metode IG, hasil rata-rata akurasi jauh di bawah pemodelan tanpa dan manual pemilihan peubah. Selain itu, dengan metode IG pada kasus ini hasil peubah yang terseleksi terpangkas hampir 80 – 85% token. Dalam kasus ini, terlihat bahwa akan banyak token yang teseleksi jika menggunakan metode IG. Sehingga setelah meninjau performa dengan 2 metode pemilihan peubah dan pertimbangan risiko tersebut, pemodelan selanjutnya dilakukan tanpa pemilihan peubah apapun.

Aplikasi InfoBMKG dengan pelabelan (1) menghasilkan rentang rata-rata akurasi dari 45.05 – 52.52% dengan DRF memberikan hasil terbaik. Pelabelan (2) memberikan hasil rata-rata akurasi 77.02 – 84.37% dengan algoritma ET yang memberikan hasil terbaik sama pada dengan pelabelan (3) dengan performa tertinggi 77.63%. Aplikasi BPOM *Mobile* dengan menggunakan keempat metode pelabelan hasilnya algoritma ET memberikan performa model terbaik hingga 83.39% dengan pelabelan (2). Dari keempat pelabelan pada dua aplikasi menunjukkan DT menghasilkan performa model paling rendah dibandingkan algoritma lainnya.

TABEL XI
RATA-RATA NILAI AKURASI (%) APLIKASI MILIK PEMERINTAH

Algoritma	InfoBMKG			
	P (1)	P (2)	P (3)	P (4)
RF	51.12	80.25	72.09	71.78
DRF	**52.52**	80.85	72.29	**71.78**
ET	51.61	**84.37**	**77.63**	71.47
DT	45.05	77.02	67.15	67.05

Algoritma	BPOM *Mobile*			
	P (1)	P (2)	P (3)	P (4)
RF	76.48	80.06	74.06	79.14
DRF	76.71	79.84	74.29	79.14
ET	**77.81**	**83.39**	**79.83**	**81.57**
DT	74.05	77.83	71.83	75.59

TABEL XII
RATA-RATA NILAI AKURASI (%) APLIKASI MILIK BUMN

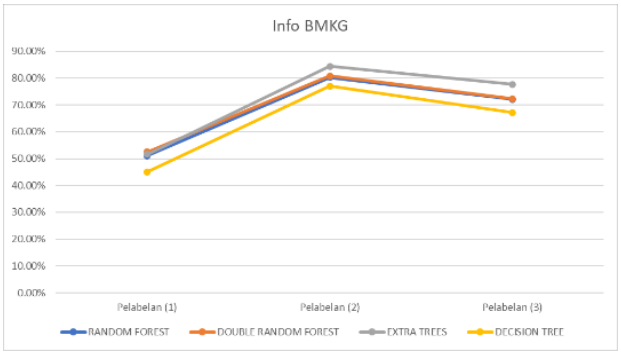
Algoritma	*MyIndihome*			
	P (1)	P (2)	P (3)	P (4)
RF	58.55	79.42	73.73	75.55
DRF	58.36	79.99	73.82	75.65
ET	**58.74**	**82.31**	**78.17**	**76.81**
DT	53.33	75.46	68.32	70.43

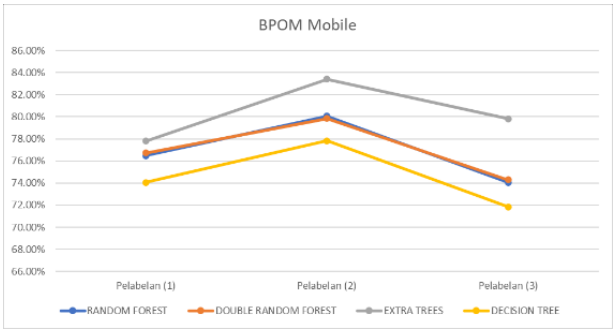
Algoritma	*MyPertamina*			
	P (1)	P (2)	P (3)	P (4)
RF	51.76	77.39	72.93	73.60
DRF	52.23	78.34	73.87	73.98
ET	**52.42**	**79.96**	**76.63**	**74.83**
DT	47.39	75.69	63.15	64.76

Hasil akurasi performa model pada *MyIndihome* dan *MyPertamina*, keduanya menunjukkan bahwa algoritma ET dapat menghasilkan performa model terbaik dengan seluruh metode pelabelan. Sama halnya pada InfoBMKG dan BPOM *M*obile, DT menjadi algoritma yang memberikan performa model paling rendah. Aplikasi *MyIndihome* dan *MyPertamina* dengan menggunakan pelabelan (1) hasilnya paling rendah dibandingkan ketiga pelabelan lainnya pada kasus ini. Hasil rata-rata akurasi antara kedua aplikasi tidak berbeda jauh dan juga nilai *gap* yang kecil untuk performa antar algoritma.

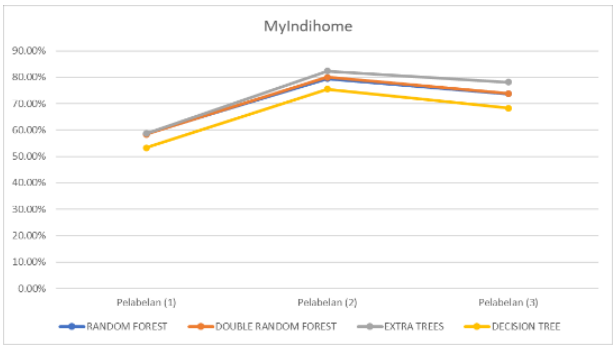
Hasil analisis secara keseluruhan dari empat aplikasi menunjukkan bahwa pelabelan dengan validasi dapat meningkatkan performa model setelah melakukan evaluasi ulang terhadap ulasan yang memberikan hasil berbeda. Jika dibandingkan berdasarkan penilaian *rating*, pelabelan (4) dalam kasus ini mampu mengatasi permasalahan kondisi yang tidak dapat dihindari sebelumnya pada pelabelan (1) karena mampu meningkat performa hingga ±20% yang artinya cukup memberikan pengaruh yang signifikan. Jika dibandingkan berdasarkan metode *scoring*, pelabelan (2) dalam kasus ini mampu memberikan hasil yang lebih baik dibandingkan dengan pelabelan (3) serta pelabelan lainnya.

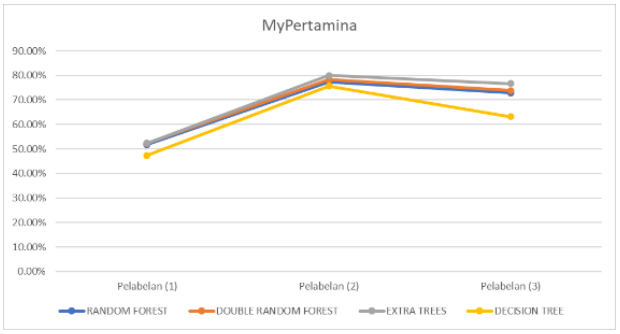
Secara keseluruhan, hasil analisis terhadap 4 aplikasi milik pemerintah dan BUMN dengan 4 metode pelabelan menunjukkan bahwa dari 22 kali pemodelan, 18 diantaranya menunjukkan algoritma ET memberikan performa model terbaik dalam kasus ini dibandingkan dengan ketiga algoritma berbasis pohon lainnya dan 4 pemodelan lainnya menunjukkan DRF yang memberikan performa terbaik. Sedangkan algoritma DT berada di posisi paling akhir. RF dan DRF cenderung menghasilkan rata-rata nilai akurasi yang tidak jauh berbeda antar satu sama lain. Hal ini dapat dilihat pada grafik di bawah ini.





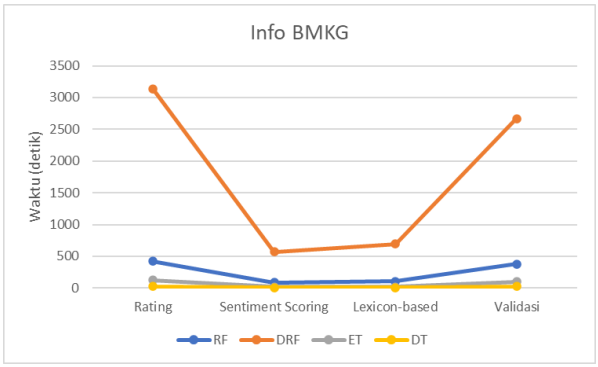
Gambar. 2 Rata-rata nilai akurasi setiap aplikasi milik pemerintah

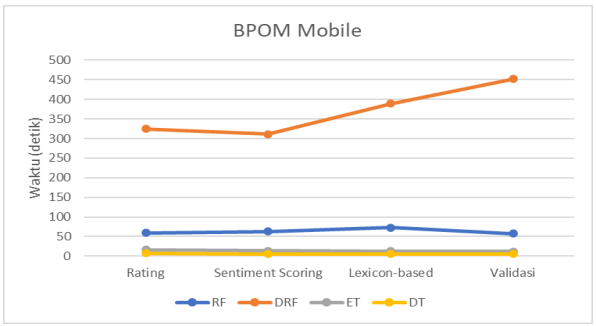


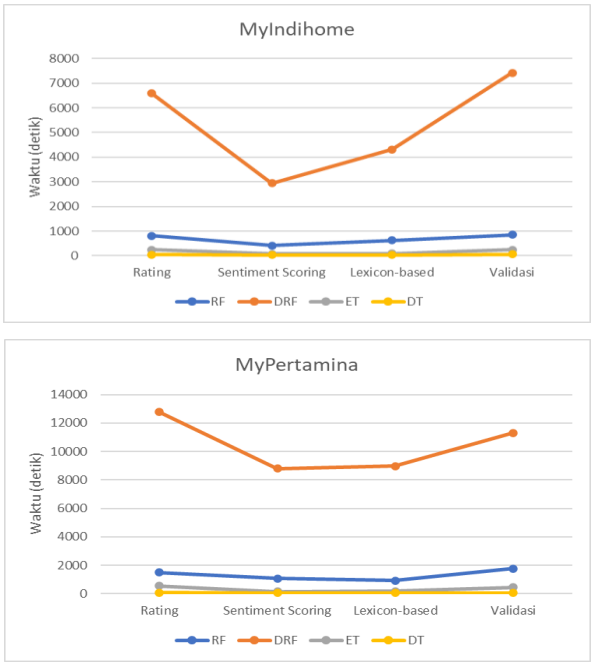


Gambar. 3 Rata-rata nilai akurasi setiap aplikasi milik BUMN

Semakin terlihat jelas melalui grafik pada Gambar 2 dan Gambar 3 menunjukkan bahwa algoritma ET (abu-abu) berada di posisi paling atas, artinya memiliki rata-rata akurasi yang paling tinggi sedangkan DT (kuning) berada di posisi paling bawah, artinya rata-rata akurasi paling rendah. Untuk RF dan DRF (biru dan jingga), dari keempat grafik terlihat bahwa keduanya cukup tumpang tindih, artinya performanya tidak jauh berbeda antar satu sama lain. Berdasarkan jenis metode pelabelan, pelabelan (2) yaitu *sentiment scoring* yang memerhatikan kata negasi, kata sentimen, dan *booster words* cenderung memberikan rata-rata nilai akurasi yang baik (80 – 84%). Hal ini juga terlihat pada Gambar 2 dan Gambar 3, dari pelabelan (1) mengalami kenaikan ketika menggunakan pelabelan (2) dan kemudian mengalami penurunan kembali dengan menggunakan pelabelan (3).







Gambar. 4 Waktu komputasi setiap aplikasi

Berdasarkan waktu komputasi dari keempat algoritma berbasis pohon, algoritma DT merupakan algoritma yang memberikan waktu komputasi tercepat dilanjutkan dengan ET, RF, dan terakhir DRF yang membutuhkan waktu yang cukup lama. Hal ini seiring dengan teori dari DRF karena data yang digunakan semakin besar akan membuat pohon yang terbentuk juga akan semakin rimbun sehingga waktu komputasi yang dibutuhkan juga akan semakin lama. Gambar 4 menunjukkan bahwa algoritma DRF, semakin banyak peubah yang digunakan, waktu komputasi juga akan berbanding lurus yang mengakibatkan semakin lama waktu yang dipelrukan. Dalam kasus ini, DRF mampu memberikan performa yang baik namun memerlukan waktu yang cukup jika diterapkan pada analisis sentimen. Sebagai tambahan informasi, penelitian ini menggunakan perangkat keras berupa laptop dengan spesifikasi *processor* AMD Ryzen 5 3500U *with* Radeon Vega Mobile Gfx 2.10 Ghz serta *random access memory* 8 GB.

V. KESIMPULAN DAN SARAN

Hasil analisis terhadap empat aplikasi milik pemerintah dan BUMN, InfoBMKG, BPOM *Mobile*, *MyIndihome*, dan *MyPertamina* berdasarkan ulasan yang paling relevan dan terbaru menggunakan empat algoritma berbasis pohon (DT, RF, ET, dan DRF) dan empat metode pelabelan (*rating*, *sentiment scoring*, *lexicon-based*, dan validasi), hasilnya menunjukkan arah sentimen ke label positif karena cenderung banyak ulasan pada kategori tersebut dengan pelabelan (1) dan (2) sedangkan pelabelan (3) dan (4) cenderung ke arah netral dan negatif. Berdasarkan metode pelabelan, selain pelabelan (4) yang mampu memberikan pengaruh signifikan hingga meningkatkan akurasi ±20%, pelabelan (2), *sentiment scoring* cenderung memperoleh rata-rata akurasi terbaik dalam kasus ini. Hasil klasifikasi sentimen menggunakan empat algoritma berbasis pohon menunjukkan bahwa algoritma *extra trees* (ET) dalam kasus ini mampu menghasilkan performa model terbaik (sekitar 80 – 84%). Sehingga penerapan DRF pada analisis sentimen dibandingkan algoritma lain, DRF cenderung membutuhkan waktu yang lebih lama dibandingkan ET dengan perolehan nilai akurasi yang tidak jauh berbeda antar keduanya.

Saran yang dapat diberikan kepada peneliti selanjutnya yakni dapat lebih mengoptimalkan proses penyiapan data dalam hal ini tahap penyiapan data dengan memperbaiki maupun mengembangkan kamus kata yang digunakan, lainnya dapat lebih memerhatikan jumlah kapitalisasi teks atau jumlah huruf besar, tanda baca, dan tanda lainnya dalam teks agar dapat lebih memahami makna ulasan yang diberikan. Penerjemahan emoji pada teks dapat dilakukan dengan tetap memerhatikan konteks ulasan sehingga dapat lebih menggambarkan sentimen dari pengguna aplikasi. Selain itu, metode seleksi peubah dapat digunakan agar dapat membantu mengurangi waktu komputasi dalam suatu penerapan algoritma serta memilih metode seleksi peubah yang mampu menyeleksi dengan baik tanpa harus kehilangan banyaknya token secara ekstrem.

REFERENSI

[1] Asosiasi Penyelenggara Jasa Internet Indonesia (APJII). Survei Penetrasi & Perilaku Internet 2023. Jakarta, 2023.

[2] Hendriyanto, M. D., Ridha, A. A., and Enri, U, "Analisis Sentimen Ulasan Aplikasi Mola Pada Google Play Store Menggunakan Algoritma Support Vector Machine". *INTECOMS: Journal of Information Technology and Computer Science*, *5*(1), 1-7, 2022.

[3] P. Mehta and Dharnil Pandya, "A review on sentiment analysis methodologies, practices, and applications", *International Journal of Scientific and Technology Research*, vol. 2, pp. 601–609, 2020.

[4] A. Yadav and D. K. Vishwakarma, "Sentiment analysis using deep learning architectures: a review", *Artificial Intelligence Review*, vol. 53, no. 6, pp. 4335–4385, 2020.

[5] Y. Asri, W. N. Suliyanti, D. Kuswardani, M. Fajri, "Pelabelan otomatis lexicon vader dan klasifikasi Naïve Bayes dalam menganalisis sentimen data ulasan PLN Mobile", *PETIR: Jurnal Pengkajian dan Penerapan Teknik Informatika*, vol. 15, no. 2, pp. 264–275, 2022.

[6] J. A. Shathik and K. K. Prasad, "A literature review on application of sentiment analysis using machine learning techniques", *International Journal of Applied Engineering and Management Letters (IJAEML)*, vol. 4, no. 2, pp. 41–77, 2020.

[7] A. S. Neogi, K. A. Garg, R. K. Mishra, and Y. K. Dwivedi, "Sentiment analysis and classification of Indian farmers' protect using twitter data", *International Journal of Information Management Data Insights*, vol. 1, no. 2, 2021.

[8] A. S. Aribowo, H. Basiron, N. S. Herman, and S. Khomsah, "An evaluation of preprocessing steps and tree-based ensemble machine learning for analysing sentiment on Indonesian youtube comments", *International Journal of Advanced Trends in Computer Science and Engineering*, vol 9, no. 5, pp. 7078–7086, 2020.

[9] S. Han, H. Kim, and Y. S. Lee, "Double random forest", *Machine Learning*, vol. 198, pp. 1569–1586, 2020.

[10] B. Charbuty and A. Abdulazeez, "Classification based on decision tree algorithm for machine learning", *Journal of Applied Science and Technology Trends*, vol. 2, no. 1, pp. 20–28, 2021.

[11] I. Tamara. "Kajian kinerja algoritma klasifikasi extra-trees pada permasalahan data kelas tak seimbang", thesis, *Institut Pertanian Bogor*, Bogor, Indonesia, 2022.

[12] T. Daniya, M. Geetha, and K. S. Kumar, "Classification and regression trees with GINI index", *Advances in Mathematics: Scientific Journal*, vol. 9, no. 10, pp. 8237–8247, 2020.

[13] Yuan Y, Wu L, Zhang X. Gini-impurity index analysis. *IEEE Transactions on Information Forensics and Security*. 16:3154–3169, 2021.
[14] S. Tangirala, "Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm", *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 2, pp. 612 – 619, 2020.
[15] M. A. Ganaie, M. Tanveer, P. N. Suganthan, V. Snásel, "Oblique and rotation double random forest", *Neural Networks*, vol. 153, pp. 496–517, 2022.
[16] L. Breiman, "Random forests", *Machine Learning*, pp. 5–32, 2001.
[17] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning", *The Stata Journal*, vol. 20, no. 1, pp. 3–29, 2020.
[18] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees", *Machine Learning*, pp. 3–42, 2006.
[19] E. K. Ampomah, Z. Qin, and G. Nyame, "Evaluation of tree-based ensemble machine learning models in predicting stock price direction of movement", *Information*, vol. 11, no. 6, pp. 332, 2020.
[20] M. R. C. Acosta, S. Ahmed, C. E. Garcia, and I. Koo, "Extremely randomized trees-based scheme for stealthy cyber-attack detection in smart grid network", *IEEE Access*, vol. 8, pp. 19921–19933, 2020.
[21] T. Wang, K. Lu, K. P. Chow, and Q. Zhu, "COVID-19 sensing: negative sentiment analysis on social media in China via BERT model", *IEEE Access*, vol. 8, pp. 138162–138169, 2020.
[22] A. Onan, "Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks", *Concurrency and Computation: Practice and Experience*, vol. 33, no. 23, pp. e5909, 2021.
[23] M. Grandini, E. Bagli, and G. Visani, "Metrics for multi-class classification: an overview", *arXiv preprint arXiv:2008.05756*, 2020.
[24] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation", *BMC genomics*, vol. 21, no. 1, pp. 1–13, 2020.
[25] N. A. Salsabila, Y. A. Winatmoko, A. A. Septiandri, and A. Jamal, "Colloquial Indonesian lexicon", *2018 International Conference on Asian Language Processing (IALP)*, *IEEE*, pp. 226–229, 2018.
[26] D. H. Wahid and S. N. Azhari, "Peringkasan sentimen ekstraktif di twitter menggunakan hybrid TF-IDF dan cosine similarity", *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 10, no. 2, pp. 207–218, 2016.
[27] T. H. Pudjiantoro and F. R. Umbara, "Analisis sentimen terhadap e-commerce pada media sosial twitter menggunakan metode Naïve Bayes", *Seminar Nasional Informatika dan Aplikasinya (SNIA)*, 2021.
[28] A. Mash, "The impact of tokenization on gender bias in machine translation", Universitat Pompeu Fabra, Barcelona, 2023.
[29] B. G. Marcot and A. M. Hanea, "What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis?", *Computational Statistics*, col. 36, no. 3, pp. 2009–2031, 2021.
[30] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. New York, United States: Springer, 2013, vol. 26.