

PENERAPAN ALGORITMA RANDOM FOREST DENGAN VARIASI FITUR DALAM PREDIKSI POPULARITAS BERITA DIGITAL

Dega Felix Ivanusi, Vihi Atina, Anisatul Farida

Teknik Informatika, Universitas Duta Bangsa Surakarta, Indonesia
220103053@mhs.udb.ac.id

ABSTRAK

Prediksi popularitas berita digital menjadi kebutuhan penting bagi media daring untuk memahami karakteristik konten yang berpotensi memperoleh tingkat penyebaran tinggi. Namun, banyaknya atribut yang digunakan dalam proses klasifikasi dapat meningkatkan kompleksitas komputasi dan mempengaruhi kinerja model. Penelitian ini bertujuan menganalisis pengaruh variasi fitur terhadap performa algoritma Random Forest dalam prediksi popularitas berita digital. Penelitian menggunakan dataset Online News Popularity yang diperoleh dari UCI Machine Learning Repository. Metode yang digunakan meliputi pra-pemrosesan data, pembangunan model Random Forest, pengujian empat skenario variasi fitur, serta optimasi model melalui penanganan outlier, feature engineering, hyperparameter tuning, dan threshold tuning. Hasil pengujian baseline menunjukkan bahwa penggunaan seluruh fitur menghasilkan akurasi sebesar 66,34%. Setelah dilakukan optimasi, performa model meningkat dengan hasil terbaik diperoleh pada skenario reduksi fitur noise yang menghasilkan accuracy 72,15%, precision 72,05%, recall 72,15%, dan f1-score 72,10%. Hasil penelitian menunjukkan bahwa variasi fitur dan strategi optimasi berpengaruh terhadap performa klasifikasi popularitas berita digital menggunakan Random Forest.

Kata kunci : Popularitas Berita, Random Forest, Variasi Fitur, Feature Importance, Optimasi Model

1. PENDAHULUAN

Perkembangan media digital mendorong portal berita daring untuk menghasilkan konten yang mampu menarik keterlibatan pembaca secara tinggi. Salah satu indikator keterlibatan tersebut dapat dilihat melalui jumlah *shares* yang diperoleh artikel setelah dipublikasikan pada berbagai platform digital. Nilai *shares* yang tinggi menunjukkan bahwa sebuah artikel memiliki daya tarik, relevansi informasi, serta potensi penyebaran yang luas kepada masyarakat. Kemampuan memprediksi popularitas berita sebelum dipublikasikan menjadi kebutuhan penting bagi perusahaan media untuk mendukung pengambilan keputusan konten yang lebih efektif [1][2].

Prediksi popularitas berita digital menjadi persoalan yang kompleks karena karakteristik artikel dibentuk oleh banyak atribut yang saling berinteraksi, seperti statistik linguistik, polaritas sentimen, topik laten, hingga distribusi kata kunci. Banyaknya atribut prediktor yang digunakan berpotensi meningkatkan kompleksitas komputasi apabila seluruh fitur diproses secara simultan, sekaligus menyulitkan identifikasi atribut mana yang benar-benar berkontribusi terhadap popularitas berita. Penggunaan atribut dalam jumlah besar tanpa proses penyaringan dapat menurunkan efisiensi pelatihan model dan memperbesar peluang terbentuknya fitur yang kurang informatif [3]. Meskipun sejumlah penelitian terdahulu telah menerapkan algoritma *machine learning* seperti *Random Forest* yang menunjukkan kestabilan performa cukup baik pada data berdimensi tinggi [4][5], sebagian besar penelitian masih berfokus pada pencapaian nilai akurasi akhir model, sedangkan pengujian mendalam mengenai bagaimana variasi atribut masukan dapat memengaruhi efisiensi

sekali sekaligus perilaku klasifikasinya masih belum banyak dibahas [6].

Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk menganalisis pengaruh variasi fitur terhadap performa algoritma Random Forest dalam prediksi popularitas berita digital. Analisis dilakukan untuk mengetahui bagaimana perubahan jumlah atribut memengaruhi kemampuan model dalam melakukan klasifikasi serta mempertahankan kualitas hasil prediksi.

Untuk mencapai tujuan tersebut, penelitian ini menerapkan empat skenario variasi fitur yang dibangun berdasarkan nilai *feature importance* dari Random Forest, yaitu penggunaan fitur penuh, reduksi fitur *noise*, penggunaan fitur utama, dan penghapusan fitur dominan. Selain itu, dilakukan optimasi model melalui penanganan *outlier*, *feature engineering*, *hyperparameter tuning*, dan *threshold tuning* untuk meningkatkan performa klasifikasi. Pendekatan tersebut didukung oleh penelitian sebelumnya yang menunjukkan bahwa pengelolaan fitur dan optimasi parameter dapat meningkatkan kualitas model klasifikasi pada data berdimensi tinggi. Hasil penelitian diharapkan dapat memberikan gambaran mengenai pengaruh variasi fitur terhadap performa Random Forest serta menghasilkan model yang lebih efisien dan akurat dalam memprediksi popularitas berita digital.

2. TINJAUAN PUSTAKA

2.1. Penelitian Terdahulu

Penelitian mengenai prediksi popularitas konten digital telah banyak dilakukan dengan memanfaatkan algoritma *machine learning*. Zhang dan Lin [1] menerapkan algoritma Random Forest pada dataset

Online News Popularity untuk memprediksi tingkat popularitas berita daring. Hasil pengujian memperlihatkan bahwa Random Forest dapat menghasilkan performa klasifikasi yang konsisten pada data berita berdimensi tinggi. Namun, penelitian tersebut masih menitikberatkan pada pencapaian akurasi model tanpa melakukan pengujian lebih lanjut terhadap perubahan performa ketika jumlah fitur dimodifikasi.

Penelitian lain dilakukan oleh Azarin et al. [7] yang menerapkan Random Forest untuk memprediksi popularitas novel berbasis fitur-fitur teks. Hasil eksperimen menunjukkan bahwa atribut teks memiliki kontribusi terhadap tingkat popularitas suatu konten digital. Meskipun demikian, penelitian tersebut masih menggunakan susunan fitur yang relatif tetap sehingga belum menunjukkan bagaimana sensitivitas model ketika komposisi atribut mengalami pengurangan atau penghapusan [8].

Selain Random Forest, beberapa penelitian juga menunjukkan bahwa algoritma ensemble boosting seperti XGBoost mampu meningkatkan performa klasifikasi pada data berdimensi tinggi melalui proses hyperparameter tuning dan optimasi model [9].

Pada sisi optimasi model, Agustina et al. [3] menunjukkan bahwa penerapan feature selection pada Random Forest mampu meningkatkan efisiensi komputasi serta menjaga kestabilan klasifikasi. Temuan tersebut diperkuat oleh Rodicky et al. [6] yang menyatakan bahwa pengurangan atribut yang kurang informatif dapat membantu Random Forest bekerja lebih optimal. Selain itu, Ding et al. [10] juga menegaskan bahwa perubahan jumlah atribut masukan dapat mempengaruhi perilaku klasifikasi serta efisiensi pemrosesan pada model Random Forest. Penelitian lain oleh Saputra et al. [11] membuktikan bahwa algoritma Random Forest dapat memberikan hasil klasifikasi yang cukup optimal pada kasus prediksi kredit melalui tahapan preprocessing dan evaluasi model berbasis confusion matrix. Penelitian tersebut menunjukkan bahwa kualitas pra-pemrosesan data dan pemilihan atribut memberikan dampak terhadap hasil klasifikasi yang dihasilkan oleh model Random Forest.

Penelitian lain oleh Farida et al. [12] menunjukkan bahwa penerapan sistem prediksi berbasis machine learning mampu membantu proses pengambilan keputusan melalui analisis pola data secara otomatis. Penelitian tersebut menunjukkan bahwa proses pemodelan dan optimasi algoritma memiliki pengaruh terhadap peningkatan performa sistem prediksi.

Berdasarkan beberapa penelitian terdahulu tersebut, dapat diketahui bahwa algoritma Random Forest mampu mempertahankan kestabilan performa pada proses klasifikasi data berdimensi tinggi, sedangkan seleksi fitur terbukti berpengaruh terhadap efisiensi model. Akan tetapi, penelitian terdahulu masih cenderung menempatkan seleksi fitur sebagai langkah optimasi tunggal dan belum banyak menguji

ketahanan performa Random Forest melalui beberapa skenario variasi fitur secara bertahap pada domain prediksi popularitas berita digital. Oleh karena itu, penelitian ini menempatkan variasi fitur sebagai fokus eksperimen untuk melihat perubahan akurasi dan efisiensi model secara lebih komprehensif.

2.2. Random Forest dan Variasi Fitur

Random Forest adalah metode klasifikasi berbasis ensemble yang bekerja dengan membentuk banyak decision tree dari subset data yang berbeda, kemudian menghasilkan keputusan akhir berdasarkan suara mayoritas dari seluruh pohon. Metode ini banyak digunakan pada data berdimensi tinggi karena mampu menjaga kestabilan model serta mengurangi risiko overfitting, serta mampu mempertahankan kestabilan performa pada distribusi atribut yang kompleks [13]. Selain itu, Random Forest menyediakan *feature importance* yang dapat dimanfaatkan untuk menganalisis kontribusi tiap atribut terhadap performa model klasifikasi.

Nilai *feature importance* menjadi salah satu indikator penting dalam proses evaluasi atribut karena mampu menunjukkan fitur mana yang memiliki pengaruh dominan dan fitur mana yang relatif kurang informatif. Subandono dan Ariatmanto [14] menjelaskan bahwa pemanfaatan seleksi atribut secara bertahap dapat membantu model mempertahankan kestabilan performa sekaligus meningkatkan efisiensi komputasi. Temuan tersebut selaras dengan penelitian Wibowo et al. [8] yang menekankan bahwa reduksi atribut yang terukur dapat menurunkan dimensi data tanpa merusak kualitas klasifikasi secara signifikan.

Identifikasi fitur-fitur dominan juga menjadi langkah penting untuk mengetahui atribut mana yang paling berkontribusi dalam pembentukan keputusan model. Kouassi et al. [15] menunjukkan bahwa pendekatan *top-k feature selection* mampu membantu penentuan subset atribut yang paling informatif pada data berdimensi tinggi. Oleh karena itu, dalam penelitian ini nilai *feature importance* dimanfaatkan sebagai dasar untuk membangun beberapa skenario variasi fitur, yaitu reduksi fitur *noise*, pemilihan fitur utama, dan penghapusan fitur dominan.

Variasi fitur dilakukan untuk menguji apakah Random Forest tetap mampu mempertahankan performa prediksi ketika jumlah atribut dikurangi atau ketika fitur dengan kontribusi tertinggi dihilangkan. Dengan demikian, evaluasi yang dilakukan tidak hanya menilai akurasi model, tetapi juga menilai ketahanan dan efisiensi komputasi model terhadap perubahan dimensi data.

3. METODE PENELITIAN

3.1. Dataset dan Pra-pemrosesan Data

Penelitian ini menggunakan dataset *Online News Popularity* yang diperoleh dari UCI Machine Learning Repository. Dataset tersebut berisi 39.644 data artikel berita digital yang dipublikasikan dalam rentang Januari 2013 hingga Januari 2015 dengan total 61

atribut awal. Setiap data merepresentasikan karakteristik artikel berdasarkan berbagai indikator numerik, seperti jumlah kata, distribusi kata kunci, polaritas sentimen, topik laten, serta nilai *shares* sebagai indikator tingkat popularitas.

Pada tahap pra-pemrosesan, atribut *url* dan *timedelta* dihapus karena tidak memiliki kontribusi prediktif terhadap proses klasifikasi. Selanjutnya, variabel target berupa jumlah *shares* didiskritisasi menjadi dua kelas, yaitu tidak populer dan populer, menggunakan metode kuartil ketiga (Q3 threshold) sebagai batas pemisah kelas. Pendekatan ini dilakukan untuk memperoleh komposisi kelas yang lebih seimbang sehingga model klasifikasi dapat dilatih secara lebih stabil.

Setelah pembentukan label kelas, dataset dipisahkan menjadi data latih sebesar 80% dan data uji sebesar 20% menggunakan *random state* 42 agar pembagian data tetap konsisten pada setiap eksperimen. Tahap ini menghasilkan 31.715 data latih dan 7.929 data uji yang digunakan pada seluruh skenario pengujian.

3.2. Penerapan Algoritma Random Forest

Metode klasifikasi yang diterapkan pada penelitian ini adalah algoritma Random Forest. Metode Random Forest bekerja dengan membentuk banyak decision tree dari beberapa subset data hasil bootstrap sampling untuk menghasilkan keputusan klasifikasi yang lebih stabil, kemudian menentukan hasil klasifikasi akhir berdasarkan mekanisme *majority voting* [13].

Algoritma ini dipilih karena mampu mengolah data tabular berdimensi tinggi, tahan terhadap *overfitting*, serta mampu menghasilkan nilai *feature importance* untuk menganalisis pengaruh setiap atribut terhadap model.

Pada penelitian ini, model Random Forest dibangun menggunakan parameter dasar sebanyak 100 pohon keputusan ($n_estimators = 100$) dengan *random state* 42 untuk menjaga konsistensi hasil. Proses pelatihan dilakukan pada data latih, kemudian model digunakan untuk memprediksi kelas popularitas pada data uji.

3.3. Skenario Variasi Fitur

Untuk menganalisis pengaruh perubahan jumlah atribut terhadap performa model, penelitian ini menerapkan empat skenario variasi fitur yang seluruhnya dibangun berdasarkan nilai *feature importance* dari model baseline.

a. Skenario 1 (Fitur Penuh)

Model dilatih menggunakan seluruh 58 atribut prediktor sebagai model baseline untuk memperoleh acuan performa awal.

b. Skenario 2 (Reduksi Noise)

Atribut dengan nilai *feature importance* kurang dari 1% dikategorikan sebagai fitur berkontribusi rendah dan dieliminasi dari proses pelatihan.

Hasil seleksi ini menyisakan 41 atribut yang digunakan pada model kedua.

c. Skenario 3 (Fitur Utama)

Model dilatih menggunakan 15 atribut dengan nilai *feature importance* tertinggi untuk menguji kestabilan performa pada dimensi fitur yang jauh lebih ringkas.

d. Skenario 4 (Penghapusan Fitur Dominan)

Lima atribut dengan nilai *feature importance* tertinggi dihapus dari dataset, kemudian model dilatih menggunakan 53 atribut tersisa. Skenario ini bertujuan menguji sensitivitas Random Forest ketika fitur-fitur dominan dieliminasi.

3.4. Evaluasi Model

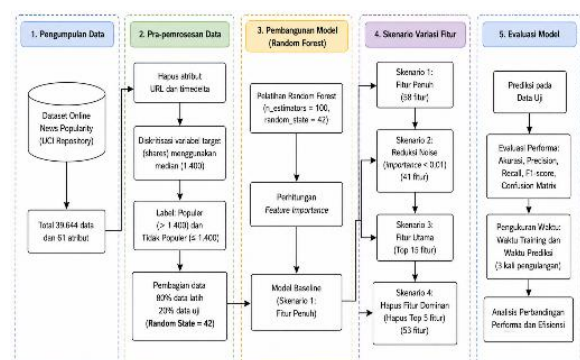
Performa model dievaluasi menggunakan metrik *accuracy*, *precision*, *recall*, dan *f1-score* yang diperoleh dari *classification report*. Confusion matrix juga digunakan untuk menganalisis sebaran hasil prediksi benar maupun kesalahan klasifikasi pada setiap kelas [5].

Untuk menilai efisiensi komputasi, setiap skenario diuji sebanyak tiga kali pengulangan dan dihitung rata-rata waktu pelatihan serta waktu prediksi model.

Melalui kombinasi evaluasi performa klasifikasi dan waktu komputasi tersebut, penelitian ini dapat menganalisis tidak hanya tingkat ketepatan prediksi, tetapi juga kestabilan dan efisiensi Random Forest terhadap perubahan jumlah atribut.

3.5. Metodologi Penelitian

Penelitian ini dilakukan melalui beberapa tahapan yang saling berkaitan, mulai dari proses pengumpulan data hingga analisis hasil klasifikasi. Setiap tahapan dirancang untuk mendukung proses pembangunan model Random Forest dalam memprediksi popularitas berita digital berdasarkan beberapa skenario variasi fitur dan proses optimasi model. Adapun alur metodologi penelitian yang digunakan dapat dilihat pada Gambar 1.



Gambar 1. Metodologi Penelitian

Dapat dilihat keterangan alur Metodologi penelitian sebagai berikut:

- a. Pengumpulan Data.
Dataset *Online News Popularity* diambil dari *UCI Machine Learning Repository* yang berisi 39.644 data artikel berita digital dengan total 61 atribut awal. Dataset digunakan untuk memprediksi tingkat popularitas berita berdasarkan jumlah *shares* yang diperoleh artikel setelah dipublikasikan pada media digital.
- b. Pra-pemrosesan Data
Tahap pra-pemrosesan bertujuan untuk membersihkan dan mempersiapkan data sebelum digunakan dalam tahapan pelatihan model. Tahap tersebut mencakup penghapusan atribut url dan timedelta karena tidak berpengaruh langsung terhadap proses klasifikasi. Selanjutnya dilakukan penanganan *outlier* pada atribut *shares* menggunakan batas persentil ke-99 untuk mengurangi pengaruh data ekstrem. Setelah itu dilakukan proses *feature engineering* dengan menambahkan beberapa atribut turunan berbasis rasio *keyword*, *token density*, dan *self-reference*. Pada tahap berikutnya, variabel target *shares* didiskritisasi menggunakan metode kuartil ketiga (Q3 threshold) sebagai batas pembentukan dua kelas popularitas berita. Dataset kemudian dibagi menjadi data latih sebesar 80% dan data uji sebesar 20% menggunakan *random state* 42.
- c. Pembangunan Model Random Forest
Model Random Forest dibangun menggunakan data latih dengan parameter dasar *n_estimators* = 100. Selanjutnya dihitung nilai *feature importance* untuk mengetahui tingkat kontribusi masing-masing atribut terhadap model klasifikasi. Model ini digunakan sebagai baseline pada skenario penggunaan fitur penuh.
- d. Skenario Variasi Fitur
Empat skenario variasi fitur dibangun berdasarkan nilai *feature importance*. Skenario 1 menggunakan seluruh fitur hasil pra-pemrosesan. Skenario 2 melakukan reduksi fitur noise dengan menghapus atribut yang memiliki nilai *feature importance* di bawah 0,01. Skenario 3 menggunakan 15 fitur utama dengan kontribusi tertinggi. Skenario 4 melakukan penghapusan lima fitur dominan dengan nilai *feature importance* tertinggi untuk menguji ketahanan model terhadap perubahan komposisi atribut.
- e. Evaluasi Model dan Optimasi Model
Setiap model digunakan untuk memprediksi data uji, kemudian performa dievaluasi berdasarkan metrik *accuracy*, *precision*, *recall*, *f1-score*, dan *confusion matrix*. Selain pengujian performa, waktu pelatihan dan waktu prediksi diukur untuk melihat efisiensi model pada masing-masing skenario. Setelah tahap evaluasi baseline dilakukan, penelitian dilanjutkan dengan tahap optimasi model menggunakan penanganan *outlier*, *feature engineering*, *hyperparameter tuning*, dan *threshold tuning* untuk meningkatkan performa

klasifikasi popularitas berita digital. Optimasi dilakukan pada algoritma Random Forest untuk memperoleh keseimbangan performa *accuracy*, *precision*, *recall*, dan *f1-score* pada setiap skenario variasi fitur [16].

4. HASIL DAN PEMBAHASAN

Penelitian ini menghasilkan empat skenario pengujian utama untuk menganalisis pengaruh variasi jumlah fitur terhadap performa algoritma Random Forest dalam memprediksi popularitas berita digital. Setiap skenario diuji menggunakan dataset yang sama dengan komposisi atribut yang berbeda, kemudian dievaluasi berdasarkan metrik akurasi, *precision*, *recall*, *f1-score*, serta efisiensi waktu komputasi. Selain itu, analisis juga dilakukan terhadap perubahan perilaku model melalui *confusion matrix* dan tren degradasi akurasi pada setiap skenario.

Hasil eksperimen dibagi menjadi empat bagian utama, yaitu analisis performa model baseline menggunakan seluruh fitur, analisis reduksi fitur *noise*, analisis penggunaan fitur utama, serta analisis sensitivitas melalui penghapusan fitur dominan. Seluruh hasil tersebut selanjutnya dibandingkan untuk menentukan konfigurasi fitur yang paling optimal.

4.1. Pengumpulan Data

Dataset yang digunakan dalam penelitian ini berasal dari dataset *Online News Popularity* yang bersumber dari *UCI Machine Learning Repository*. Dataset tersebut memuat data artikel berita daring dari portal berita Mashable beserta berbagai atribut numerik yang merepresentasikan karakteristik artikel. Dataset digunakan untuk melakukan prediksi tingkat popularitas berita berdasarkan jumlah *shares* yang diperoleh setelah artikel dipublikasikan.

Dataset awal memiliki jumlah data sebanyak 39.644 baris dengan 61 atribut, yang terdiri dari atribut identitas artikel, atribut statistik teks, atribut kata kunci (*keywords*), atribut sentimen, atribut performa referensi, serta atribut target berupa jumlah *shares*. Pada penelitian ini, atribut *shares* digunakan sebagai dasar pembentukan kelas popularitas berita.

Tahap pengumpulan dataset diperoleh melalui proses unduh dari repository UCI, kemudian data dibaca menggunakan bahasa pemrograman Python melalui library *pandas*. Setelah data berhasil dimuat, dilakukan pengecekan struktur data untuk memastikan jumlah atribut dan kelengkapan dataset sebelum memasuki tahap pra-pemrosesan data.

Tabel 1. Sampel Dataset Online News Popularity

N o	n_tokens_title	n_tokens_content	num_images	shares
1	12	219	1	593
2	9	255	1	711
3	9	211	1	1500
4	9	531	1	1200
5	13	1072	20	505

Berdasarkan Tabel 1, dapat dilihat bahwa dataset *Online News Popularity* memiliki berbagai atribut numerik yang merepresentasikan karakteristik artikel berita digital, seperti jumlah kata pada judul artikel, jumlah kata pada isi artikel, jumlah gambar, serta jumlah *shares* yang diperoleh artikel. Atribut-atribut tersebut digunakan sebagai variabel masukan dalam proses prediksi popularitas berita digital. Keterangan:

- n_tokens_title* : jumlah kata pada judul artikel.
- n_tokens_content* : jumlah kata pada isi artikel.
- num_imgs* : jumlah gambar pada artikel.
- shares* : jumlah penyebaran artikel.

4.2. Pra-pemrosesan Data

Tahapan pra-pemrosesan bertujuan menyiapkan dataset sebelum memasuki proses pelatihan model machine learning. Proses ini bertujuan untuk membersihkan data, mengurangi noise, serta meningkatkan kualitas atribut agar model dapat bekerja lebih optimal. Pada penelitian ini, tahapan pra-pemrosesan meliputi *data cleaning*, *outlier handling*, *feature engineering*, dan pembentukan label kelas popularitas.

Tahap pertama adalah *data cleaning* dengan menghapus atribut url dan *timedelta* karena kedua atribut tersebut tidak memberikan kontribusi langsung terhadap proses klasifikasi popularitas berita. Setelah proses pembersihan data dilakukan, jumlah atribut berkurang dari 61 atribut menjadi 59 atribut.

Tahap berikutnya adalah penanganan *outlier* pada atribut *shares*. Dataset *Online News Popularity* memiliki distribusi nilai *shares* yang sangat tidak merata dengan beberapa artikel memiliki jumlah *shares* yang sangat tinggi. Kondisi tersebut dapat menyebabkan model menjadi bias terhadap data ekstrem. Oleh karena itu, dilakukan penghapusan data pada nilai *shares* di atas persentil ke-99 untuk mengurangi noise akibat artikel viral ekstrem. Setelah

proses ini dilakukan, jumlah data berubah dari 39.644 data menjadi 39.247 data.

Selanjutnya dilakukan proses *feature engineering* untuk menghasilkan atribut baru yang diharapkan dapat memperkuat informasi pada dataset. Beberapa fitur baru yang ditambahkan meliputi rasio *keyword*, rasio *self-reference*, *token density*, dan rasio jumlah kata pada judul terhadap isi artikel. Setelah proses *feature engineering*, jumlah atribut meningkat menjadi 64 atribut.

Tahap terakhir adalah pembentukan label kelas popularitas berita menggunakan metode kuartil ketiga (*Q3 threshold*). Artikel dengan jumlah *shares* di atas nilai kuartil ketiga dikategorikan sebagai berita populer, sedangkan artikel di bawah nilai tersebut dikategorikan sebagai berita tidak populer.

Tabel 2. Tahapan Pra-pemrosesan Data

Tahap	Hasil
Dataset awal	39.644 data, 61 atribut
Penghapusan atribut url dan <i>timedelta</i>	39.644 data, 59 atribut
Outlier handling (persentil 99)	39.247 data, 59 atribut
Feature engineering	39.247 data, 64 atribut
Labeling popularity (Q3)	Kelas populer dan tidak populer

Berdasarkan Tabel 2, proses pra-pemrosesan dilakukan secara bertahap mulai dari pembersihan atribut, penanganan *outlier*, hingga proses *feature engineering*. Tahapan tersebut menghasilkan perubahan jumlah atribut dan data yang digunakan dalam proses pelatihan model klasifikasi. Selain itu, proses labeling dilakukan menggunakan metode kuartil ketiga (*Q3 threshold*) untuk membentuk kelas popularitas berita.

Tabel 3. Data Setelah Pra-pemrosesan

<i>n_tokens_title</i>	<i>token_density</i>	<i>kw_ratio_max_avg</i>	<i>self_reference_ratio</i>	<i>popularity</i>
12	0,54	1,12	0,08	Tidak Populer
9	0,47	0,95	0,05	Tidak Populer
13	0,62	1,43	0,11	Populer
11	0,51	1,08	0,07	Tidak Populer
15	0,69	1,58	0,14	Populer

Berdasarkan Tabel 3, dapat dilihat bahwa dataset hasil pra-pemrosesan telah memiliki atribut tambahan hasil *feature engineering*, seperti *token_density* dan *kw_ratio_max_avg*. Dataset tersebut juga telah memiliki label *popularity* yang digunakan sebagai target klasifikasi pada tahap pelatihan model klasifikasi.

4.3. Pembangunan Model Random Forest

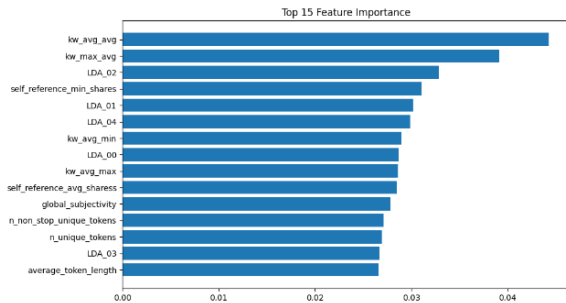
Tahap ini berfokus pada proses pembentukan model klasifikasi menggunakan algoritma Random Forest sebagai model baseline penelitian. Random Forest adalah algoritma berbasis decision tree yang

membentuk banyak pohon keputusan dari beberapa subset data untuk menghasilkan keputusan klasifikasi akhir melalui mekanisme voting menghasilkan keputusan final berdasarkan suara keputusan mayoritas yang dihasilkan oleh seluruh pohon klasifikasi.

Model Random Forest dibangun menggunakan data latih sebesar 80% dari total dataset hasil pra-pemrosesan. Pada penelitian ini, parameter dasar yang digunakan yaitu *n_estimators* = 100 dengan *random state* 42. Setelah proses pelatihan model selesai dilakukan, selanjutnya dihitung nilai *feature importance* untuk mengetahui tingkat kontribusi

atribut terhadap proses klasifikasi popularitas berita digital.

Nilai *feature importance* digunakan sebagai dasar dalam membentuk skenario variasi fitur pada tahap pengujian berikutnya. Semakin tinggi nilai *feature importance*, maka semakin besar kontribusi atribut tersebut terhadap keputusan klasifikasi model Random Forest.



Gambar 2. Grafik Feature Importance

Berdasarkan Gambar 2, beberapa atribut menunjukkan kontribusi yang lebih dominan dibanding atribut lainnya dalam proses klasifikasi popularitas berita digital. Atribut dengan nilai *feature importance* tertinggi digunakan sebagai acuan dalam

proses seleksi fitur dan pembentukan skenario variasi fitur pada tahap pengujian model. Selain itu, hasil *feature importance* juga menunjukkan bahwa Random Forest mampu mengidentifikasi atribut-atribut yang paling berpengaruh terhadap tingkat popularitas berita.

4.4. Skenario Variasi Fitur

Tahap ini mencakup pengujian model Random Forest dengan empat skenario variasi fitur yang dibentuk berdasarkan nilai *feature importance*. Pengujian dilakukan untuk mengevaluasi dampak perubahan komposisi atribut terhadap kinerja model klasifikasi popularitas berita digital. Tahap ini juga digunakan untuk mengamati tingkat ketahanan model Random Forest ketika jumlah dan susunan fitur mengalami perubahan.

Skenario pertama menggunakan seluruh fitur hasil pra-pemrosesan sebagai model baseline. Skenario kedua melakukan reduksi atribut noise dengan menghapus fitur yang memiliki nilai *feature importance* rendah. Skenario ketiga menggunakan 15 fitur utama dengan kontribusi tertinggi, sedangkan skenario keempat menghapus lima fitur dominan untuk menguji ketahanan model terhadap hilangnya atribut dengan kontribusi terbesar.

Tabel 4. Hasil Pengujian Variasi Fitur Random Forest

Skenario Pengujian	Jumlah Fitur	Akurasi (%)	Avg Training (detik)	Avg Prediksi (detik)
Skenario 1 - Fitur Penuh	58	66,34	17,5550	0,1733
Skenario 2 - Reduksi Noise	41	65,71	20,2194	0,1670
Skenario 3 – 15 Fitur Utama	15	64,52	14,3196	0,1588
Skenario 4 - Hapus Fitur Dominan	53	65,36	15,0027	0,1500

Berdasarkan Tabel 4, Skenario 1 menggunakan seluruh atribut menghasilkan akurasi tertinggi sebesar 66,34% sehingga digunakan sebagai baseline model Random Forest. Sementara itu, Skenario 2 dan Skenario 4 masih mampu mempertahankan performa klasifikasi di atas 65% meskipun dilakukan pengurangan atribut pada proses pelatihan model.

Pada Skenario 3, penggunaan 15 fitur utama menyebabkan akurasi mengalami penurunan menjadi 64,52%. Hal tersebut menunjukkan bahwa meskipun fitur utama memiliki kontribusi terbesar terhadap model, beberapa atribut lain tetap memberikan informasi tambahan yang membantu proses klasifikasi popularitas berita digital.

Selain performa akurasi, hasil pengujian juga menunjukkan bahwa pengurangan jumlah fitur mampu menurunkan waktu pelatihan dan waktu prediksi model. Skenario 3 menghasilkan waktu pelatihan tercepat dibandingkan skenario lainnya karena jumlah atribut yang digunakan lebih sedikit. Dengan demikian, variasi fitur tidak hanya memengaruhi performa klasifikasi, tetapi juga memengaruhi efisiensi komputasi model Random Forest.

4.5. Evaluasi dan Optimasi Model

Selain metrik akurasi, hasil klasifikasi pada setiap skenario juga dievaluasi menggunakan

precision, recall, f1-score, dan confusion matrix guna menganalisis distribusi prediksi pada masing-masing kelas. Evaluasi ini dilakukan untuk mengetahui kemampuan model dalam mengenali kelas berita populer dan tidak populer secara lebih detail.

Berdasarkan hasil pengujian baseline Random Forest pada empat skenario variasi fitur, performa klasifikasi model masih berada pada kisaran 64%–66%. Skenario penggunaan seluruh fitur menghasilkan performa terbaik dibandingkan skenario lainnya dengan nilai akurasi sebesar 66,34%. Sementara itu, pengurangan jumlah fitur pada Skenario 3 menyebabkan performa model mengalami penurunan karena sebagian informasi penting yang mendukung proses klasifikasi ikut berkurang.

Hasil evaluasi baseline membuktikan bahwa Random Forest cukup stabil dalam menghadapi perubahan komposisi atribut masukan. Akan tetapi, performa klasifikasi model masih belum optimal karena nilai metrik evaluasi secara keseluruhan masih berada di bawah 70%. Oleh karena itu, dilakukan tahap optimasi model untuk meningkatkan performa klasifikasi popularitas berita digital.

Tahap optimasi dilakukan melalui beberapa proses, yaitu penanganan *outlier*, *feature engineering*, *hyperparameter tuning*, dan *threshold tuning* pada model Random Forest. Penanganan *outlier* dilakukan

menggunakan batas persentil ke-99 pada atribut *shares* untuk mengurangi pengaruh data ekstrem. Selain itu, dilakukan penambahan atribut turunan seperti *token density*, *keyword ratio*, dan *self-reference ratio* untuk memperkuat informasi prediktor pada dataset.

Pada tahap optimasi, parameter Random Forest disesuaikan menggunakan nilai *n_estimators* = 300,

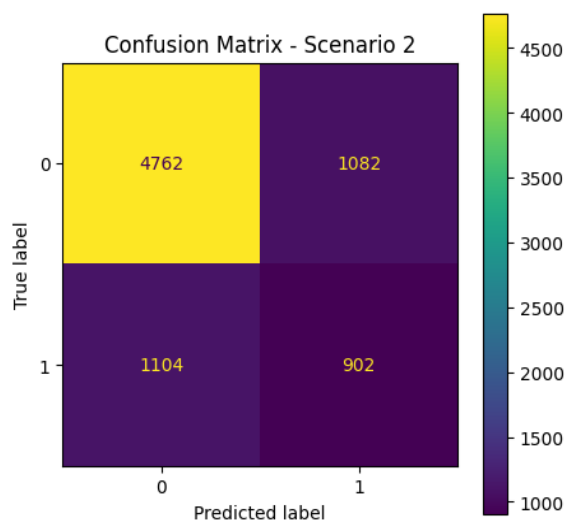
max_depth = 20, *min_samples_split* = 5, *min_samples_leaf* = 2, dan *class_weight* = *balanced*. Selain itu, dilakukan penyesuaian *threshold* probabilitas klasifikasi sebesar 0,40 untuk memperoleh keseimbangan performa antara *accuracy*, *precision*, *recall*, dan *f1-score*.

Tabel 5. Hasil Optimasi Random Forest

Skenario	Accuracy	Precision	Recall	F1-Score
Skenario 1 - Fitur Penuh	72,06%	72,01%	72,06%	72,04%
Skenario 2 - Reduksi Noise	72,15%	72,05%	72,15%	72,10%
Skenario 3 - 15 Fitur Utama	65,58%	71,60%	65,58%	67,47%
Skenario 4 - Hapus Fitur Dominan	71,76%	71,35%	71,76%	71,54%

Berdasarkan Tabel 5, hasil optimasi Random Forest menunjukkan peningkatan performa klasifikasi yang signifikan dibandingkan model baseline. Skenario 2 menghasilkan performa terbaik dengan nilai akurasi sebesar 72,15%, precision sebesar 72,05%, recall sebesar 72,15%, dan f1-score sebesar 72,10%. Hasil tersebut menunjukkan bahwa penghapusan atribut dengan kontribusi rendah mampu meningkatkan performa model karena mengurangi noise pada proses pembelajaran Random Forest.

Skenario 1 dan Skenario 4 juga menunjukkan performa klasifikasi yang relatif stabil dengan nilai metrik di atas 71%. Hal tersebut menunjukkan bahwa algoritma Random Forest masih mampu mempertahankan kestabilan terhadap penghapusan fitur dominan karena model masih mampu memanfaatkan atribut lain sebagai sumber informasi klasifikasi.



Gambar 3. Confusion Matrix Hasil Optimasi Random Forest

Sementara itu, Skenario 3 mengalami penurunan performa dibandingkan skenario lainnya dengan nilai akurasi sebesar 65,58%. Hasil tersebut menunjukkan bahwa penggunaan hanya 15 fitur utama belum mampu merepresentasikan keseluruhan informasi klasifikasi secara optimal. Dengan demikian, meskipun fitur utama memiliki kontribusi terbesar

terhadap model, atribut tambahan lainnya tetap memberikan pengaruh terhadap performa klasifikasi popularitas berita digital.

Berdasarkan Gambar 3, model Random Forest hasil optimasi mampu melakukan klasifikasi pada kedua kelas dengan distribusi prediksi yang lebih seimbang dibandingkan model baseline. Jumlah prediksi benar pada kelas populer maupun tidak populer mengalami peningkatan setelah dilakukan proses optimasi model. Selain itu, jumlah kesalahan klasifikasi juga mengalami penurunan sehingga berdampak pada peningkatan nilai *accuracy*, *precision*, *recall*, dan *f1-score* secara keseluruhan.

Hasil confusion matrix menunjukkan bahwa proses optimasi berhasil meningkatkan kemampuan model dalam mengenali pola popularitas berita digital tanpa menyebabkan ketimpangan prediksi pada salah satu kelas.

Secara keseluruhan, hasil optimasi menunjukkan bahwa kombinasi penanganan *outlier*, *feature engineering*, *hyperparameter tuning*, dan *threshold tuning* mampu meningkatkan performa Random Forest secara signifikan dibandingkan model baseline. Selain itu, hasil penelitian juga menunjukkan bahwa variasi fitur memberikan pengaruh terhadap performa klasifikasi model Random Forest pada dataset *Online News Popularity*.

5. KESIMPULAN DAN SARAN

Hasil penelitian menunjukkan bahwa algoritma Random Forest mampu digunakan untuk klasifikasi popularitas berita digital pada dataset *Online News Popularity*. Pengujian menggunakan empat skenario variasi fitur membuktikan bahwa Random Forest tetap memiliki kinerja yang stabil terhadap perubahan komposisi atribut masukan. Skenario penggunaan seluruh fitur menghasilkan akurasi tertinggi sebesar 66,34% pada model baseline, sedangkan skenario reduksi fitur dan penghapusan fitur dominan masih mampu mempertahankan performa klasifikasi pada kisaran 65%.

Hasil evaluasi baseline menggunakan *precision*, *recall*, dan *f1-score* menunjukkan bahwa hasil klasifikasi pada keempat skenario masih relatif stabil dengan nilai metrik berada pada kisaran 0,64–0,66.

Hal tersebut menunjukkan bahwa pengurangan atribut tidak secara langsung menyebabkan penurunan performa model secara drastis, sehingga Random Forest tetap mampu memanfaatkan atribut lain dalam proses klasifikasi popularitas berita digital.

Selain pengujian baseline Random Forest, penelitian ini juga melakukan tahap optimasi model melalui penanganan outlier, feature engineering, hyperparameter tuning, dan threshold tuning. Hasil optimasi menunjukkan peningkatan performa klasifikasi secara signifikan dibandingkan model baseline. Skenario reduksi fitur noise menghasilkan performa terbaik memperoleh accuracy 72,15%, precision 72,05%, recall 72,15%, dan f1-score 72,10%.

Secara keseluruhan, penelitian menunjukkan bahwa proses pra-pemrosesan data, variasi fitur, serta strategi optimasi Random Forest memiliki pengaruh terhadap performa klasifikasi popularitas berita digital. Selain itu, penelitian juga menunjukkan bahwa penghapusan fitur dengan kontribusi rendah mampu meningkatkan performa model tanpa mengurangi kestabilan klasifikasi secara signifikan.

DAFTAR PUSTAKA

- [1] Y. Zhang and K. Lin, "Predicting and Evaluating the Online News Popularity based on Random Forest," *J. Phys. Conf. Ser.*, vol. 1994, no. 1, 2021.
- [2] M. J. Tsai and Y. Q. Wu, "Predicting online news popularity based on machine learning," *Computers and Electrical Engineering*, vol. 104, 2022.
- [3] T. Agustina, M. Masrizal, and I. Irmayanti, "Performance Analysis of Random Forest Algorithm for Network Anomaly Detection using Feature Selection," *Sinkron*, vol. 8, no. 2, 2024.
- [4] A. R. Syulistyo and V. M. Agustin, "Predicting News Article Popularity with Multi Layer Perceptron Algorithm," *Journal of Applied Intelligent System*, vol. 7, no. 2, 2022.
- [5] M. F. Kurniawan and D. A. Megawaty, "Comparison of Logistic Regression, Random Forest, Support Vector Machine and K-Nearest Neighbor Algorithms in Prediction," *Journal of Applied Informatics and Computing*, vol. 9, no. 5, 2025.
- [6] N. Rodicky, S. Soim, and S. Zefi, "Seleksi Fitur dengan Genetika Algorithm untuk Optimasi Random Forest," *Transmisi: Jurnal Ilmiah Teknik Elektro*, vol. 27, no. 3, pp. 185–190, 2025.
- [7] N. E. Azarin, R. A. Saputra, and Subardin, "Prediksi Popularitas Novel Berbasis Fitur-Fitur Teks Menggunakan Metode Random Forest," *JIMI*, vol. 9, no. 1, pp. 57–62, 2024.
- [8] Wibowo et al., "Impact of Feature Selection on Decision Tree and Random Forest for Classifying Student Study Success," *BAREKENG: Journal of Mathematics and Its Applications*, vol. 19, no. 3, 2025.
- [9] A. A. Syahputra and R. E. Saputro, "Application of the XGBoost Model with Hyperparameter Tuning for Industry Classification for Job Applicants," *Sinkron*, vol. 8, no. 3, pp. 1920–1931, 2024.
- [10] J. Ding, J. Du, and H. Wang, "A novel two-stage feature selection method based on random forest and improved genetic algorithm for enhancing classification in machine learning," *Sci. Rep.*, vol. 15, 2025.
- [11] D. B. Saputra, V. Atina, and F. E. Nastiti, "Penerapan Model CRISP-DM pada Prediksi Nasabah Kredit Menggunakan Algoritma Random Forest," *Idealis: Indonesia Journal Information System*, vol. 7, no. 2, pp. 240–247, 2024.
- [12] A. Farida, V. Atina, and D. Suwandi, "Mathematical Modeling and Integration of Machine Learning-Based Prediction System on E-Learning Platform to Improve Students' Academic Performance," *JTAM (Jurnal Teori dan Aplikasi Matematika)*, vol. 9, no. 3, 2025.
- [13] M. Imani, A. Beikmohammadi, and H. R. Arabnia, "Comprehensive Analysis of Random Forest and XGBoost Performance Under Varying Imbalance Levels," *Technologies (Basel)*, vol. 13, no. 3, 2025.
- [14] A. T. P. Subandono and D. Ariatmanto, "Optimalisasi Seleksi Fitur dalam Analisis Sentimen Menggunakan Information Gain dan Chi-Square," *Sistemasi: Jurnal Sistem Informasi*, 2025.
- [15] B. M. et al. Kouassi, "Top-K Feature Selection for IoT Intrusion Detection," *Future Internet*, vol. 17, no. 11, 2025.
- [16] J. L. et al. Leevy, "Threshold optimization and random undersampling for imbalanced classification," *J. Big Data*, vol. 10, 2023.