

**KLASIFIKASI JUDUL BERITA CLICKBAIT BERBAHASA
INDONESIA MENGGUNAKAN ALGORITMA GRADIENT
BOOSTING TREE**

SKRIPSI

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi Informatika



disusun oleh

MUHAMMAD ZHAFAR AL FATHI

21.11.4374

Kepada

**FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA**

2025

**KLASIFIKASI JUDUL BERITA CLICKBAIT BERBAHASA
INDONESIA MENGGUNAKAN ALGORITMA GRADIENT
BOOSTING TREE**

SKRIPSI

untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi Informatika



disusun oleh

MUHAMMAD ZHAFAR AL FATHI

21.11.4374

Kepada

**FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA**

2025

HALAMAN PERSETUJUAN

SKRIPSI

**KLASIFIKASI JUDUL BERITA CLICKBAIT BERBAHASA
INDONESIA MENGGUNAKAN ALGORITMA GRADIENT
BOOSTING TREE**

yang disusun dan diajukan oleh

Muhammad Zhafar Al Fathi

21.11.4374

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 25 Agustus 2025

Dosen Pembimbing,


Nuri Cahyono M.Kom
NIK. 190302278

HALAMAN PENGESAHAN
SKRIPSI
KLASIFIKASI JUDUL BERITA CLICKBAIT BERBAHASA
INDONESIA MENGGUNAKAN ALGORITMA GRADIENT
BOOSTING TREE

yang disusun dan diajukan oleh

Muhammad Zhafar Al Fathi

21.11.4374

Telah dipertahankan di depan Dewan Penguji
pada tanggal 25 Agustus 2025

Susunan Dewan Penguji

Nama Penguji

Majid Rahardi, S.Kom., M.Eng.
NIK. 190302393

Ade Pujiyanto, M.Kom.
NIK. 190302494

Nuri Cahyono, M.Kom.
NIK. 190302278

Tanda Tangan



Skripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer Tanggal
25 Agustus 2025

DEKAN FAKULTAS ILMU KOMPUTER



Prof. Dr. Kusriani, M.Kom.
NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

Nama mahasiswa : Muhammad Zhafar Al Fathi
NIM : 21.11.4374

Menyatakan bahwa Skripsi dengan judul berikut:

Klasifikasi Judul Berita Clickbait Berbahasa Indonesia Menggunakan Algoritma Gradient Boosting Tree

Dosen Pembimbing : Nuri Cahyono, M.Kom

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya.
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Dosen Pembimbing.
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini.
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta.
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi.

Yogyakarta, 25 Agustus 2025

Yang Menyatakan,



Muhammad Zhafar Al Fathi

KATA PENGANTAR

Alhamdulillah, segala puji dan syukur penulis panjatkan ke hadirat Allah SWT, Tuhan semesta alam, atas segala rahmat, taufik, dan hidayah-Nya yang tiada terhingga, sehingga penulis dapat menyelesaikan skripsi yang berjudul “Klasifikasi Judul Berita Clickbait Berbahasa Indonesia Menggunakan Algoritma Gradient Boosting Tree” ini. Skripsi ini adalah sebuah capaian yang tidak mungkin penulis lewati tanpa bantuan dan doa dari banyak pihak. Rasa syukur yang mendalam penulis ucapkan kepada:

1. Kedua orang tua tercinta, atas segala kasih sayang, dukungan, doa, dan pengorbanan, baik moril maupun materil, yang senantiasa menjadi sumber kekuatan dan inspirasi bagi penulis.
2. Bapak Nuri Cahyono M.Kom selaku dosen pembimbing yang telah memberikan arahan selama pengerjaan skripsi ini. Terima kasih atas kesabarannya selama membimbing penulis dalam mengerjakan skripsi ini.
3. Tim dosen penguji yang telah memberikan evaluasi, masukan, dan saran yang sangat berharga terhadap penyempurnaan skripsi ini.
4. Teman-teman mahasiswa yang senantiasa kebersamai penulis selama masa perkuliahan. Terima kasih telah menjadi teman yang selalu *support* dan menyemangati penulis semasa menjalankan kuliah maupun saat pengerjaan skripsi ini.

Karya ini hanyalah sebuah ikhtiar yang masih jauh dari kata sempurna. Oleh sebab itu, penulis sangat mengharapkan kritik dan saran yang membangun. Semoga hasil penelitian ini dapat memberikan manfaat bagi pengembangan ilmu pengetahuan maupun bagi pihak-pihak yang membutuhkan.

Yogyakarta, 22 Agustus 2025

Penulis

DAFTAR ISI

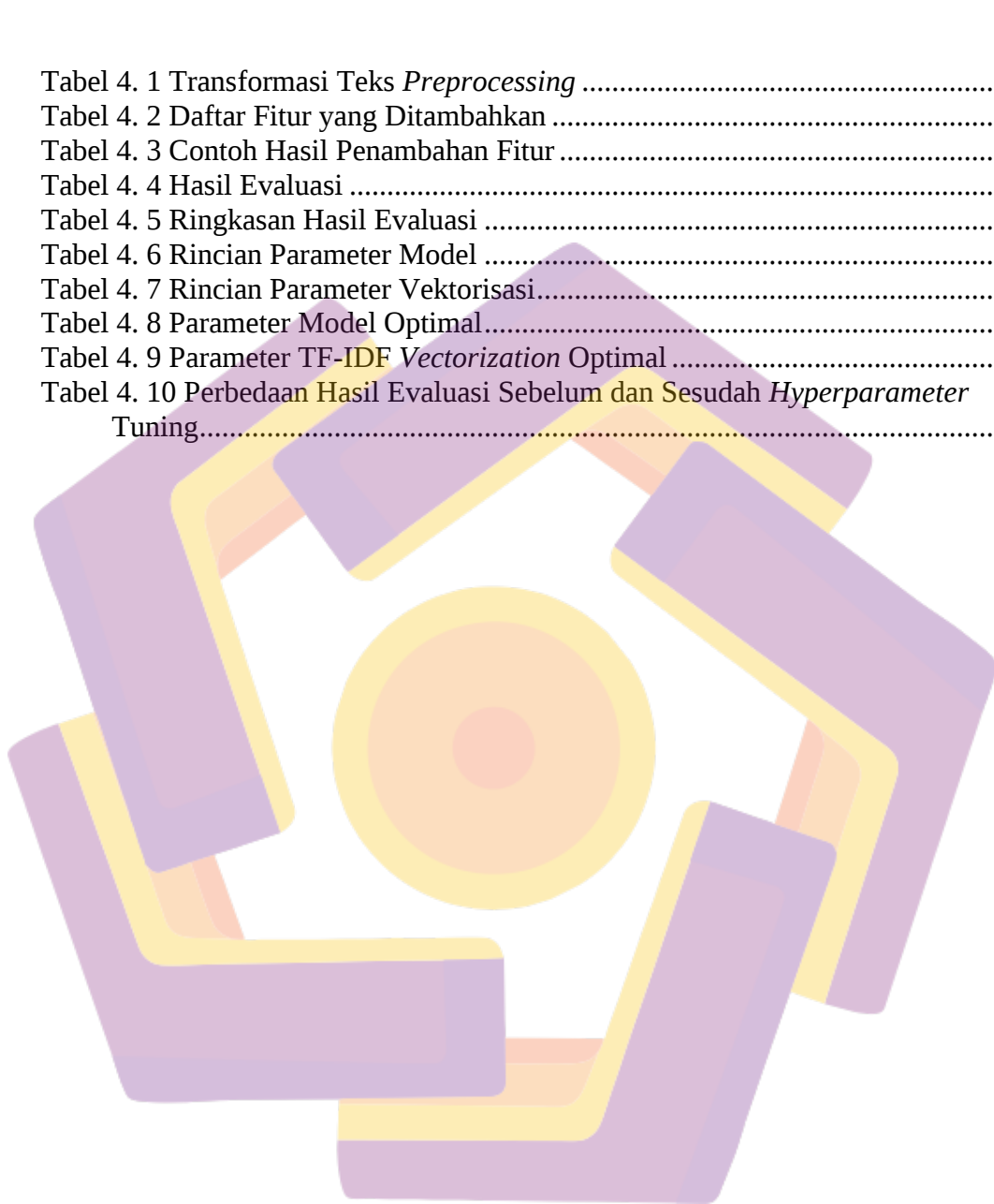
HALAMAN JUDUL	I
HALAMAN PERSETUJUAN	II
HALAMAN PENGESAHAN.....	III
HALAMAN PERNYATAAN KEASLIAN SKRIPSI.....	IV
KATA PENGANTAR.....	V
DAFTAR ISI.....	VI
DAFTAR TABEL	IX
DAFTAR GAMBAR.....	X
DAFTAR LAMBANG DAN SINGKATAN	XI
DAFTAR ISTILAH	XIII
INTISARI	XXIII
ABSTRACT	XXIV
BAB I PENDAHULUAN.....	1
1.1 LATAR BELAKANG.....	1
1.2 RUMUSAN MASALAH	3
1.3 BATASAN MASALAH	4
1.4 TUJUAN PENELITIAN	4
1.5 MANFAAT PENELITIAN	4
1.6 SISTEMATIKA PENULISAN	4
BAB II TINJAUAN PUSTAKA	6
2.1 STUDI LITERATUR.....	6
2.2 DASAR TEORI.....	11
2.2.1 <i>Clickbait</i>	11
2.2.2 <i>Machine Learning</i>	12
2.2.3 Model	12
2.2.4 <i>Ensemble Learning</i>	13

2.2.5	Boosting	14
2.2.6	Preprocessing	14
2.2.7	Word Embedding	15
2.2.8	Gradient Boosting	15
2.2.9	TF-IDF Vectorizer	17
2.2.10	Feature Engineering	17
2.2.11	Hyperparameter Tuning	18
2.2.12	Akurasi	19
2.2.13	Precision	19
2.2.14	Recall	20
2.2.15	F1-Score	20
2.2.16	Optuna	21
2.2.17	Inference	21
2.2.18	Transformer	22
2.2.19	Quantization	22
2.2.20	Data Augmentation	23
2.2.21	Pipeline	24
BAB III METODE PENELITIAN		26
3.1	OBJEK PENELITIAN	26
3.2	ALUR PENELITIAN	26
3.2.1	Pengumpulan Data	27
3.2.2	Preprocessing	28
3.2.3	Exploratory Data Analysis	28
3.2.4	Feature Engineering	29
3.2.5	Pembuatan Model	29
3.2.6	Evaluasi Model	30
3.2.7	Optimalisasi Model	31
3.2.8	Implementasi Model	31
3.3	ALAT DAN BAHAN	32
3.3.1	Alat	32

3.3.2 Bahan	34
BAB IV HASIL DAN PEMBAHASAN	36
4.1 PENGUMPULAN DATA	36
4.2 PREPROCESSING	36
4.3 EXPLORATORY DATA ANALYSIS	40
4.4 PENAMBAHAN FITUR.....	47
4.5 PEMBUATAN MODEL	49
4.6 HASIL EVALUASI.....	50
4.7 OPTIMALISASI MODEL	51
4.8 IMPLEMENTASI MODEL	56
BAB V PENUTUP	58
5.1 KESIMPULAN.....	58
5.2 SARAN	59
REFERENSI.....	60
LAMPIRAN.....	65

DAFTAR TABEL

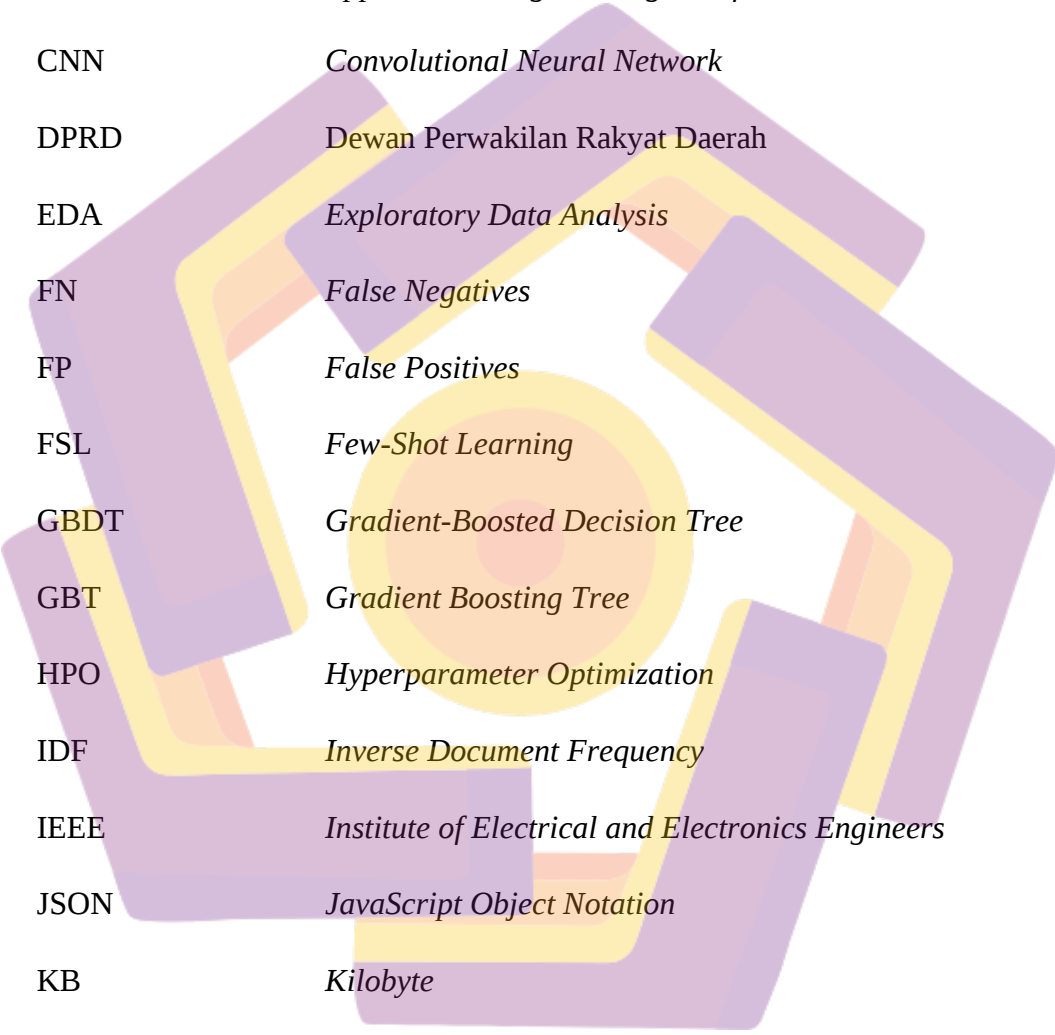
Tabel 4. 1 Transformasi Teks <i>Preprocessing</i>	40
Tabel 4. 2 Daftar Fitur yang Ditambahkan	47
Tabel 4. 3 Contoh Hasil Penambahan Fitur	48
Tabel 4. 4 Hasil Evaluasi	51
Tabel 4. 5 Ringkasan Hasil Evaluasi	51
Tabel 4. 6 Rincian Parameter Model	52
Tabel 4. 7 Rincian Parameter Vektorisasi.....	53
Tabel 4. 8 Parameter Model Optimal.....	53
Tabel 4. 9 Parameter TF-IDF <i>Vectorization</i> Optimal	54
Tabel 4. 10 Perbedaan Hasil Evaluasi Sebelum dan Sesudah <i>Hyperparameter</i> Tuning.....	55



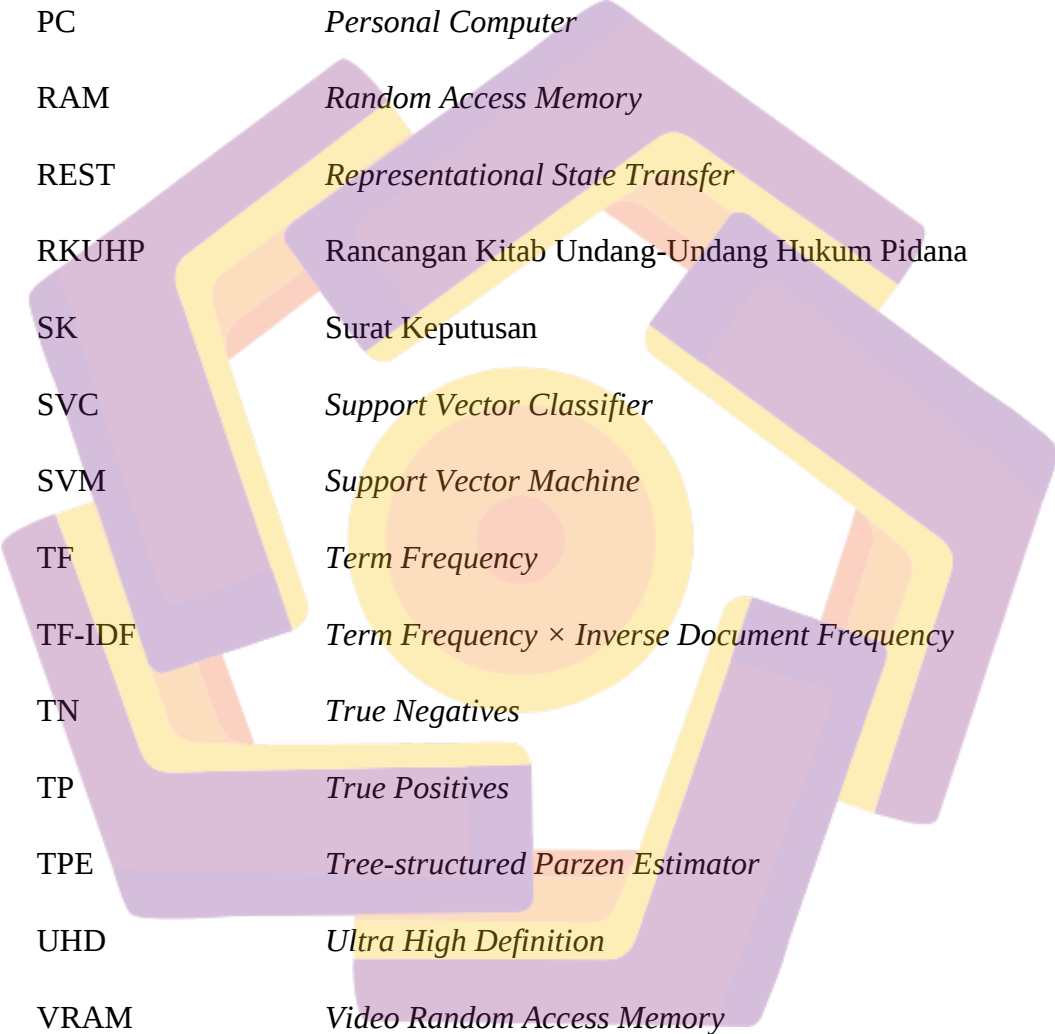
DAFTAR GAMBAR

Gambar 2. 1 Ilustrasi Gradient Boosting	16
Gambar 3. 1 Diagram Alur Penelitian	27
Gambar 4. 1 Distribusi data Sebelum Balancing	37
Gambar 4. 2 Distribusi data setelah <i>balancing</i>	39
Gambar 4. 3 Analisis Penggunaan Simbol	41
Gambar 4. 4 Analisis Jumlah Digit pada Judul.....	42
Gambar 4. 5 Analisis <i>Wordcloud</i>	43
Gambar 4. 6 Analisis Distribusi Kata	44
Gambar 4. 7 Rasio Penggunaan Kata	45
Gambar 4. 8 Analisis <i>Bigram</i>	46
Gambar 4. 9 Kode Program Perancangan <i>Transformer</i> Kustom	50
Gambar 4. 10 Visualisasi Perkembangan <i>Hyperparameter</i>	54
Gambar 4. 11 Antarmuka Percakapan Dengan Bot <i>Discord</i>	56

DAFTAR LAMBANG DAN SINGKATAN

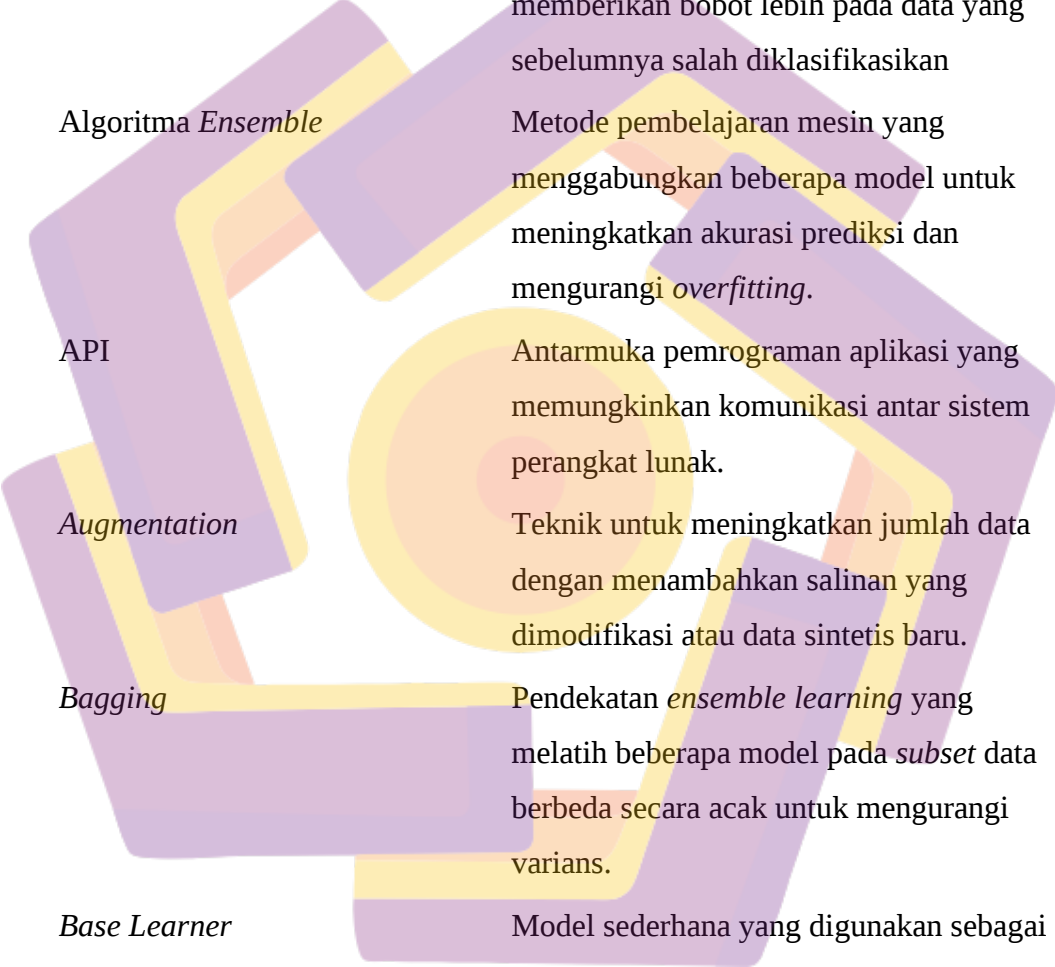


AI	<i>Artificial Intelligence</i>
AMD	<i>Advanced Micro Devices</i>
API	<i>Application Programming Interface</i>
CNN	<i>Convolutional Neural Network</i>
DPRD	<i>Dewan Perwakilan Rakyat Daerah</i>
EDA	<i>Exploratory Data Analysis</i>
FN	<i>False Negatives</i>
FP	<i>False Positives</i>
FSL	<i>Few-Shot Learning</i>
GBDT	<i>Gradient-Boosted Decision Tree</i>
GBT	<i>Gradient Boosting Tree</i>
HPO	<i>Hyperparameter Optimization</i>
IDF	<i>Inverse Document Frequency</i>
IEEE	<i>Institute of Electrical and Electronics Engineers</i>
JSON	<i>JavaScript Object Notation</i>
KB	<i>Kilobyte</i>
KPK	<i>Komisi Pemberantasan Korupsi</i>
<i>LGBMClassifier</i>	<i>LightGBM Classifier</i>
LLM	<i>Large Language Model</i>
LSTM	<i>Long Short-Term Memory</i>
MB	<i>Megabyte</i>



ML	<i>Machine Learning</i>
MSI	<i>Micro-Star International</i>
NLP	<i>Natural Language Processing</i>
NLTK	<i>Natural Language Toolkit</i>
PC	<i>Personal Computer</i>
RAM	<i>Random Access Memory</i>
REST	<i>Representational State Transfer</i>
RKUHP	Rancangan Kitab Undang-Undang Hukum Pidana
SK	Surat Keputusan
SVC	<i>Support Vector Classifier</i>
SVM	<i>Support Vector Machine</i>
TF	<i>Term Frequency</i>
TF-IDF	<i>Term Frequency $\times$ Inverse Document Frequency</i>
TN	<i>True Negatives</i>
TP	<i>True Positives</i>
TPE	<i>Tree-structured Parzen Estimator</i>
UHD	<i>Ultra High Definition</i>
VRAM	<i>Video Random Access Memory</i>
VS Code	<i>Visual Studio Code</i>
X	Variabel Independen
XGBoost	<i>Extreme Gradient Boosting</i>
y	Variabel Dependen

DAFTAR ISTILAH



<i>Accuracy</i>	Metrik yang mengukur proporsi prediksi yang benar dari seluruh prediksi yang dibuat oleh model
<i>AdaBoost</i>	Algoritma <i>boosting</i> yang secara iteratif memberikan bobot lebih pada data yang sebelumnya salah diklasifikasikan
Algoritma <i>Ensemble</i>	Metode pembelajaran mesin yang menggabungkan beberapa model untuk meningkatkan akurasi prediksi dan mengurangi <i>overfitting</i> .
API	Antarmuka pemrograman aplikasi yang memungkinkan komunikasi antar sistem perangkat lunak.
<i>Augmentation</i>	Teknik untuk meningkatkan jumlah data dengan menambahkan salinan yang dimodifikasi atau data sintesis baru.
<i>Bagging</i>	Pendekatan <i>ensemble learning</i> yang melatih beberapa model pada <i>subset</i> data berbeda secara acak untuk mengurangi <i>varians</i> .
<i>Base Learner</i>	Model sederhana yang digunakan sebagai komponen dasar dalam <i>ensemble learning</i> .
Bias	Kesalahan sistematis yang terjadi ketika model terlalu sederhana untuk menangkap pola yang mendasari dalam data.
<i>Bigram</i>	Pasangan dua kata yang muncul secara berurutan dalam sebuah teks.

Boosting

Metode *ensemble learning* yang melatih model secara berurutan dengan fokus pada kesalahan model sebelumnya.

ChatGPT

Model bahasa besar (LLM) dari *OpenAI* yang digunakan untuk berbagai tugas pemrosesan bahasa alami.

Clickbait

Judul atau konten teks yang dirancang untuk menciptakan *information gap* guna memancing rasa penasaran dan mendorong klik.

Clickworthy

Konten yang memiliki daya tarik tinggi untuk diklik, meskipun tidak selalu informatif.

CNN

Jenis jaringan saraf yang efektif untuk pemrosesan data dengan struktur *grid* seperti gambar atau teks.

colsample_bytree

Proporsi fitur yang digunakan oleh setiap pohon.

Cross-Validation

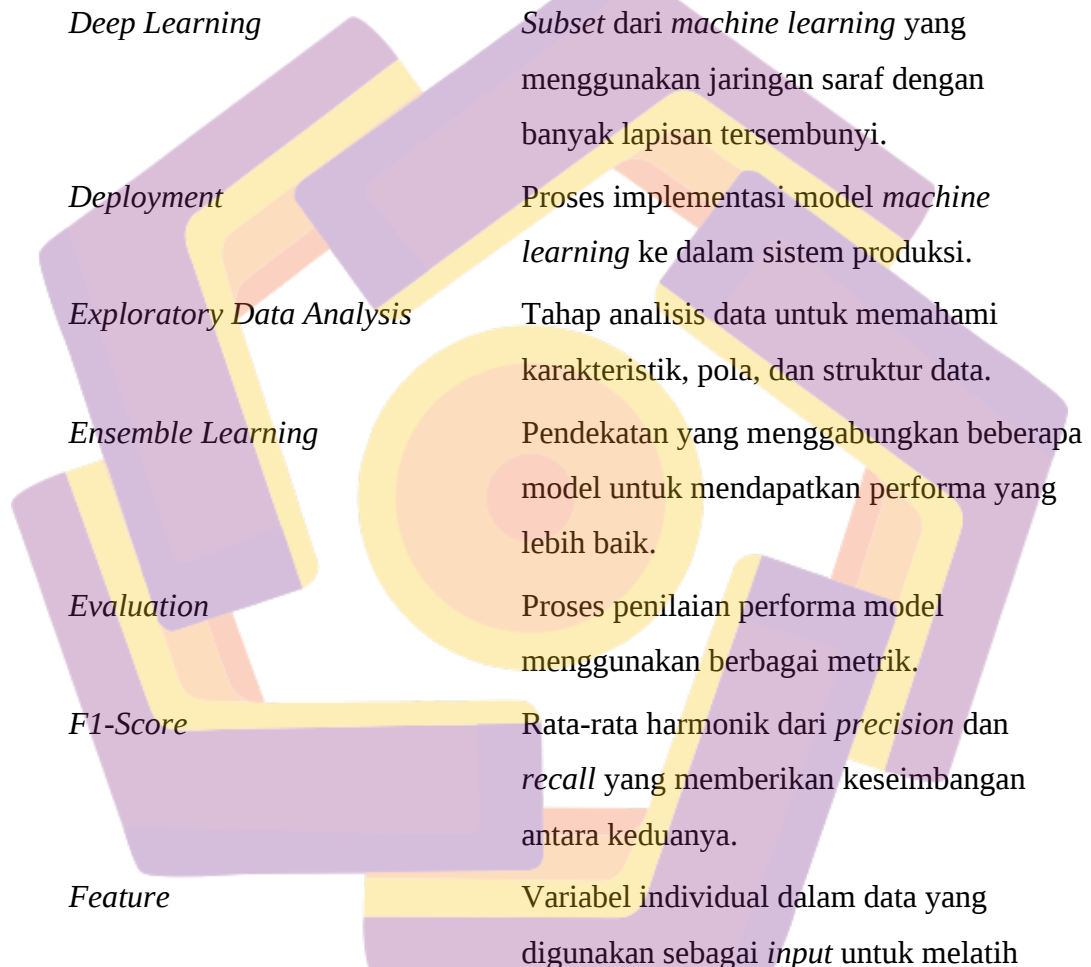
Teknik evaluasi model yang membagi data menjadi beberapa *subset* untuk pengujian yang lebih *robust*.

Custom Transformer

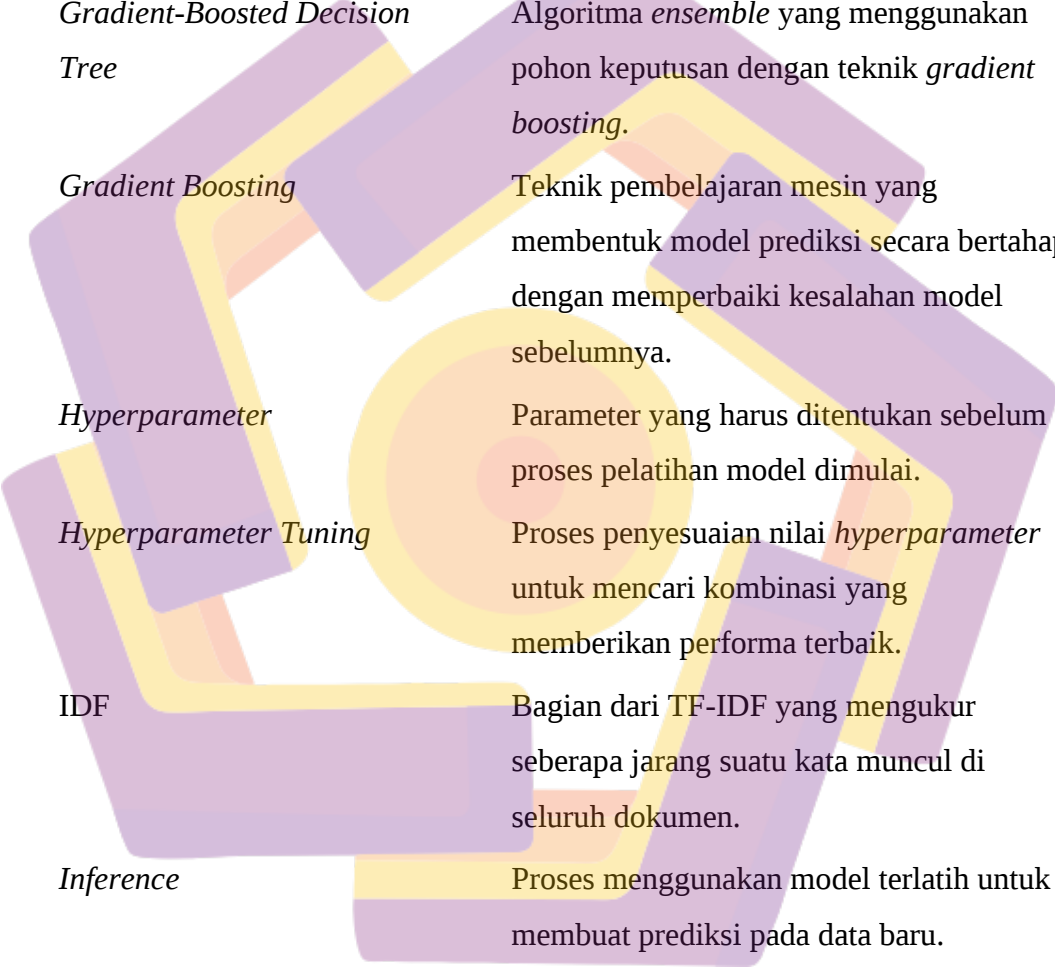
Objek yang dirancang khusus untuk melakukan transformasi data sesuai kebutuhan spesifik.

Data Augmentation

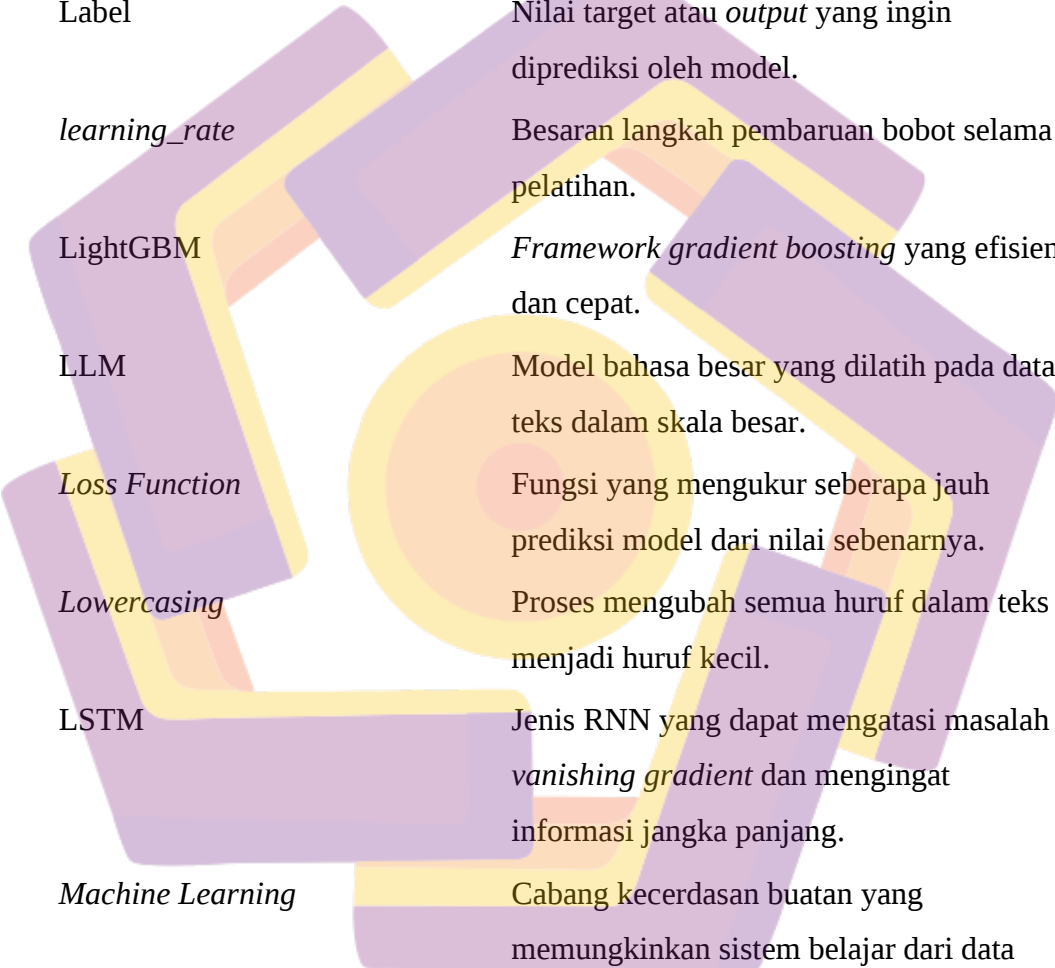
Metode untuk meningkatkan keragaman data pelatihan melalui modifikasi atau pembuatan data baru.



<i>Dataset</i>	Kumpulan data yang digunakan untuk melatih dan menguji model <i>machine learning</i> .
<i>Decision Tree</i>	Algoritma <i>machine learning</i> yang menggunakan struktur pohon untuk membuat keputusan klasifikasi atau regresi.
<i>Deep Learning</i>	Subset dari <i>machine learning</i> yang menggunakan jaringan saraf dengan banyak lapisan tersembunyi.
<i>Deployment</i>	Proses implementasi model <i>machine learning</i> ke dalam sistem produksi.
<i>Exploratory Data Analysis</i>	Tahap analisis data untuk memahami karakteristik, pola, dan struktur data.
<i>Ensemble Learning</i>	Pendekatan yang menggabungkan beberapa model untuk mendapatkan performa yang lebih baik.
<i>Evaluation</i>	Proses penilaian performa model menggunakan berbagai metrik.
<i>F1-Score</i>	Rata-rata harmonik dari <i>precision</i> dan <i>recall</i> yang memberikan keseimbangan antara keduanya.
<i>Feature</i>	Variabel individual dalam data yang digunakan sebagai <i>input</i> untuk melatih model <i>machine learning</i> .
<i>Feature Engineering</i>	Proses memilih, mentransformasi, atau menciptakan fitur-fitur baru dari data mentah.
Fitur Simbolik	Karakteristik atau atribut data yang bersifat kategorikal atau nominal, bukan numerik.



Fitur Teksual	Atribut yang diekstrak dari teks, seperti kata kunci, pola bahasa, atau struktur kalimat.
Flask	<i>Framework</i> web <i>Python</i> yang ringan untuk membangun aplikasi web dan API.
Fold	<i>Subset</i> data dalam teknik <i>cross-validation</i> .
Gradient-Boosted Decision Tree	Algoritma <i>ensemble</i> yang menggunakan pohon keputusan dengan teknik <i>gradient boosting</i> .
Gradient Boosting	Teknik pembelajaran mesin yang membentuk model prediksi secara bertahap dengan memperbaiki kesalahan model sebelumnya.
Hyperparameter	Parameter yang harus ditentukan sebelum proses pelatihan model dimulai.
Hyperparameter Tuning	Proses penyesuaian nilai <i>hyperparameter</i> untuk mencari kombinasi yang memberikan performa terbaik.
IDF	Bagian dari TF-IDF yang mengukur seberapa jarang suatu kata muncul di seluruh dokumen.
Inference	Proses menggunakan model terlatih untuk membuat prediksi pada data baru.
Information Gap	Celah antara apa yang diketahui pembaca dan apa yang ingin mereka ketahui; sering dimanfaatkan dalam clickbait.
K-Fold Cross-Validation	Teknik validasi yang membagi data menjadi <i>k subset</i> untuk evaluasi model yang lebih <i>robust</i> .



Klasifikasi	Tugas pembelajaran mesin untuk mengkategorikan data ke dalam kelas-kelas tertentu.
Kurasi Algoritmik	Sistem otomatis yang mengatur dan memprioritaskan konten berdasarkan algoritma tertentu.
Label	Nilai target atau <i>output</i> yang ingin diprediksi oleh model.
<i>learning_rate</i>	Besaran langkah pembaruan bobot selama pelatihan.
LightGBM	<i>Framework gradient boosting</i> yang efisien dan cepat.
LLM	Model bahasa besar yang dilatih pada data teks dalam skala besar.
<i>Loss Function</i>	Fungsi yang mengukur seberapa jauh prediksi model dari nilai sebenarnya.
<i>Lowercasing</i>	Proses mengubah semua huruf dalam teks menjadi huruf kecil.
LSTM	Jenis RNN yang dapat mengatasi masalah <i>vanishing gradient</i> dan mengingat informasi jangka panjang.
<i>Machine Learning</i>	Cabang kecerdasan buatan yang memungkinkan sistem belajar dari data tanpa diprogram secara eksplisit.
<i>max_depth</i>	Kedalaman maksimum dari setiap pohon dalam model.
Media Arus Utama	Media massa tradisional yang memiliki jangkauan luas dan pengaruh besar dalam masyarakat.

min_child_samples

Jumlah minimum sampel yang dibutuhkan dalam satu *leaf node*.

min_split_gain

Nilai minimum gain yang dibutuhkan untuk melakukan pemisahan (*split*).

Model

Representasi matematis atau komputasional yang dibangun melalui proses pelatihan untuk mempelajari pola dari data.

n_estimators

Jumlah total pohon dalam *ensemble* model.

n_jobs

Jumlah *thread* yang digunakan untuk proses paralel.

N-gram

Urutan n item berturut-turut dari teks, seperti kata atau karakter.

NLP

Bidang yang menggabungkan linguistik dan *machine learning* untuk memproses bahasa manusia.

NLTK

Library Python untuk pemrosesan bahasa alami.

Noising

Teknik data *augmentation* yang menambahkan gangguan acak tanpa merusak makna.

num_leaves

Jumlah maksimum daun dalam satu pohon keputusan.

Optuna

Framework open-source untuk optimasi *hyperparameter* yang menggunakan algoritma Bayesian optimization.

Overfitting

Kondisi ketika model terlalu menyesuaikan diri dengan data pelatihan sehingga performa pada data baru menurun.

Oversampling

Teknik untuk mengatasi ketidakseimbangan kelas dengan menambah jumlah sampel pada kelas minoritas.

Paraphrasing

Teknik data *augmentation* yang menghasilkan parafrase dari data asli dengan makna yang sama.

Pipeline

Rangkaian tahapan pemrosesan data dan pelatihan model yang terstruktur dan otomatis.

Pola Linguistik

Struktur atau keteraturan dalam penggunaan bahasa yang dapat diidentifikasi dan dianalisis.

Pragmatik

Cabang linguistik yang mempelajari penggunaan bahasa dalam konteks komunikasi nyata.

Precision

Metrik yang mengukur seberapa akurat prediksi positif yang dibuat oleh model.

Preprocessing

Tahap pembersihan dan persiapan data sebelum digunakan dalam pelatihan model.

Prinsip Relevansi

Teori komunikasi bahwa komunikator berusaha memberikan informasi yang relevan dengan usaha pemrosesan minimal.

PySastrawi

Library Python untuk *stemming* bahasa Indonesia.

Quantization

Teknik kompresi yang mengurangi ukuran model dengan mengubah representasi numerik ke presisi yang lebih rendah.

Random Forest

Algoritma *ensemble* yang menggunakan banyak *decision tree* dengan voting mayoritas.

random_state

Nilai acak yang ditetapkan untuk memastikan hasil yang dapat direproduksi.

Real-time

Sistem yang dapat memproses dan merespons *input* secara langsung tanpa penundaan signifikan.

Recall

Metrik yang mengukur kemampuan model untuk menemukan semua instance positif yang sebenarnya ada.

reg_alpha

Parameter regulasi L1 untuk mengurangi kompleksitas model.

reg_lambda

Parameter regulasi L2 untuk mencegah *overfitting*.

Regular Expression

Pola pencarian yang digunakan untuk mencocokkan dan memanipulasi string.

Retraining

Proses pelatihan ulang model dengan parameter atau data yang telah diperbarui.

Sampling

Teknik data *augmentation* yang menghasilkan data baru berdasarkan distribusi data yang telah dipelajari.

Scikit-learn

Library machine learning untuk *Python* yang menyediakan berbagai algoritma dan *tools*.

Stemming

Proses mengubah kata ke bentuk dasarnya dengan menghilangkan imbuhan.

Stop Words

Kata-kata umum yang biasanya diabaikan dalam analisis teks karena tidak memberikan informasi bermakna.

subsample

Proporsi sampel data pelatihan yang digunakan pada setiap pohon.

subsample_freq

Frekuensi *subsampling* untuk memperbarui pohon.

SVM

Algoritma *machine learning* yang mencari *hyperplane* optimal untuk memisahkan kelas.

TF

Frekuensi kemunculan suatu kata dalam dokumen dibagi dengan frekuensi kata paling sering.

TF-IDF

Metode untuk mengukur pentingnya suatu kata dalam dokumen relatif terhadap koleksi dokumen.

Tokenisasi

Proses memecah teks menjadi unit-unit yang lebih kecil seperti kata atau frasa.

Transformer

Objek yang dapat mempelajari parameter dari data pelatihan dan mengubah data tersebut.

Tree-based Ensemble

Metode *ensemble* yang menggunakan pohon keputusan sebagai model dasar.

Tuning

Proses penyesuaian parameter untuk mengoptimalkan performa model.

Varians

Ukuran seberapa besar model berubah ketika dilatih pada dataset yang berbeda.

Vectorizer

Alat yang mengubah teks menjadi representasi numerik yang dapat diproses oleh algoritma *machine learning*.

verbose

Tingkat keluaran log selama pelatihan model.

Word Embedding

Representasi vektor berdimensi-n dari kata yang menangkap makna semantik.

Word2Vec

Teknik *word embedding* yang mempelajari representasi vektor kata berdasarkan konteks penggunaannya.

Wordcloud

Visualisasi yang menampilkan kata-kata dengan ukuran yang proporsional dengan frekuensi kemunculannya.



INTISARI

Perkembangan *clickbait* di media digital Indonesia menimbulkan permasalahan serius dalam ekosistem jurnalistik, dimana praktik ini menurunkan kualitas informasi dan kepercayaan publik terhadap media. Judul-judul *clickbait* yang menggunakan bahasa hiperbolik dan emosional untuk memancing klik demi pendapatan iklan digital telah menyebar luas di berbagai platform berita daring, bahkan pada media arus utama yang sebelumnya kredibel. Fenomena ini diperparah oleh algoritma media sosial yang mendorong konten dengan tingkat interaksi tinggi, sehingga media cenderung memproduksi konten "*clickworthy*" alih-alih informatif.

Penelitian ini mengembangkan model *machine learning* untuk klasifikasi otomatis judul berita *clickbait* berbahasa Indonesia menggunakan algoritma *Gradient Boosting Tree* dengan kombinasi fitur tekstual dan simbolik. Metodologi meliputi pengumpulan dataset CLICK-ID dari Kaggle yang berisi 15.000 judul berita, *preprocessing* dengan teknik data *augmentation* berbasis *paraphrasing* menggunakan ChatGPT untuk menyeimbangkan distribusi kelas, ekstraksi 19 fitur simbolik seperti tanda baca dan kata-kata sensasional, serta implementasi pipeline dengan TF-IDF Vectorizer dan LightGBM classifier. Optimalisasi dilakukan melalui *hyperparameter tuning* menggunakan framework Optuna dengan 500 trial.

Model yang dikembangkan mencapai akurasi 83,23% dengan peningkatan performa sebesar 2,43% setelah optimalisasi, serta berhasil mereduksi ukuran model sebesar 35,15% menjadi 4.682 KB. Implementasi *real-time* dilakukan melalui REST API berbasis Flask dan bot Discord bernama "Hamzah Bot" yang dapat diakses publik. Hasil penelitian ini berkontribusi dalam meningkatkan literasi digital masyarakat dan membantu platform media dalam melakukan kurasi konten yang lebih selektif. Penelitian lebih lanjut dapat dikembangkan dengan eksplorasi teknik model *quantization* dan penggunaan dataset yang lebih terkini untuk meningkatkan akurasi dan efisiensi model.

Kata kunci: *clickbait, gradient boosting, feature engineering, data augmentation*

ABSTRACT

The rise of clickbait in Indonesia's digital media landscape has created serious issues within the journalistic ecosystem, where this practice undermines the quality of information and erodes public trust in the media. Clickbait headlines, often crafted using hyperbolic and emotional language to attract clicks for digital ad revenue, have become widespread across various online news platforms—even among previously credible mainstream outlets. This phenomenon is further exacerbated by social media algorithms that prioritize high-engagement content, prompting media organizations to produce “clickworthy” rather than informative material.

This study develops a machine learning model for the automatic classification of Indonesian-language clickbait headlines using the Gradient Boosting Tree algorithm, incorporating both textual and symbolic features. The methodology involves collecting the CLICK-ID dataset from Kaggle, which contains 15,000 news headlines; performing preprocessing through paraphrasing-based data augmentation using ChatGPT to balance class distribution; extracting 19 symbolic features such as punctuation marks and sensational words; and implementing a pipeline with a TF-IDF Vectorizer and LightGBM classifier. Model optimization was conducted through hyperparameter tuning using the Optuna framework across 500 trials.

The developed model achieved an accuracy of 83.23%, with a performance improvement of 2.43% after optimization, and successfully reduced model size by 35.15% to 4,682 KB. Real-time deployment was implemented via a Flask-based REST API and a publicly accessible Discord bot named “Hamzah Bot.” This research contributes to enhancing public digital literacy and supports media platforms in curating content more selectively. Future studies may explore model quantization techniques and the use of more recent datasets to further improve accuracy and model efficiency.

Keyword: clickbait, gradient boosting, feature engineering, data augmentation