

**KLASIFIKASI JENIS KELAMIN PENGGUNA X
BERDASARKAN CUITAN MENGGUNAKAN ALGORITMA
LOGISTIC REGRESSION, RANDOM FOREST, DAN
SUPPORT VECTOR MACHINE**

SKRIPSI

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



disusun oleh
INES PUJI LESTARI
22.11.4884

Kepada
FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2026

SKRIPSI

untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi Informatika



disusun oleh

INES PUJILESTARI

22.11.4884

Kepada

**FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2026**

**KLASIFIKASI JENIS KELAMIN PENGGUNA X
BERDASARKAN CUITAN MENGGUNAKAN ALGORITMA
LOGISTIC REGRESSION, RANDOM FOREST, DAN
SUPPORT VECTOR MACHINE
HALAMAN PERSETUJUAN**

SKRIPSI

KLASIFIKASI JENIS KELAMIN PENGGUNA X BERDASARKAN
CUITAN MENGGUNAKAN ALGORITMA LOGISTIC REGRESSION,
RANDOM FOREST, DAN SUPPORT VECTOR MACHINE

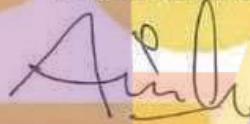
yang disusun dan diajukan oleh

Ines Puji Lestari

22.11.4884

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 24 Februari 2026

Dosen Pembimbing,



I Made Artha Agastva, S.T., M.Eng., Ph.D.
NIK. 190302352

HALAMAN PENGESAHAN

SKRIPSI

KLASIFIKASI JENIS KELAMIN PENGGUNA X BERDASARKAN
CUITAN MENGGUNAKAN ALGORITMA LOGISTIC REGRESSION,
RANDOM FOREST, DAN SUPPORT VECTOR MACHINE

yang disusun dan diajukan oleh

Ines Puji Lestari

22.11.4884

Telah dipertahankan di depan Dewan Penguji
pada tanggal 24 Februari 2026

Susunan Dewan Penguji

Nama Penguji

Tanda Tangan

Mulia Sulistiyono, S.Kom., M.Kom.
NIK. 190302248

Afriq Aminuddin, S.Kom., M.Eng., Ph.D.
NIK. 190302351

I Made Artha Agastya, S.T., M.Eng., Ph.D.
NIK. 190302352



Skripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 24 Februari 2026

DEKAN FAKULTAS ILMU KOMPUTER



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

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

Nama mahasiswa : Ines Puji Lestari
NIM : 22.11.4884

Menyatakan bahwa Skripsi dengan judul berikut:

**KLASIFIKASI JENIS KELAMIN PENGGUNA X BERDASARKAN
CUITAN MENGGUNAKAN ALGORITMA LOGISTIC REGRESSION,
RANDOM FOREST, DAN SUPPORT VECTOR MACHINE**

Dosen Pembimbing : I Made Artha Agastya, S.T., M.Eng., Ph.D.

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, 24 Februari 2026

Yang Menyatakan,



Ines Puji Lestari

HALAMAN PERSEMBAHAN

Dengan rasa syukur yang mendalam karena telah diberi kesempatan untuk menyelesaikan skripsi ini. Sebagai wujud terima kasih atas motivasi yang diberikan, penulis mempersembahkan skripsi ini kepada:

1. Untuk Ayah, Ibu, Nanang, dan Doah, yang selalu memberi doa, dukungan, dan ruang bagi saya untuk terus belajar dan bertumbuh hingga sampai di titik ini.
2. Untuk Bapak I Made Artha Agastya, S.T., M.Eng., Ph.D., selaku dosen pembimbing yang telah dengan sabar memberikan arahan, bimbingan, dan kepercayaan selama proses penyusunan skripsi ini.
3. Untuk pasangan saya, yang selalu memberi dukungan dengan caranya sendiri dan tetap ada dalam proses ini sampai selesai.
4. Untuk sahabat-sahabat saya, yang tetap memberi dukungan di tengah kesibukan dan perjuangan masing-masing.

KATA PENGANTAR

Puji syukur kehadiran Tuhan Yang Maha Esa atas rahmat dan karunia-Nya, sehingga penulis dapat menyelesaikan skripsi ini dengan judul “Klasifikasi Jenis Kelamin Pengguna X Berdasarkan Cuitan Menggunakan Algoritma Logistic Regression, Random Forest, dan Support Vector Machine” sebagai salah satu syarat dalam menyelesaikan studi program sarjana jurusan Informatika.

Dalam penyusunan skripsi ini, penulis menyadari bahwa pencapaian ini tidak terlepas dari dukungan, bimbingan, serta bantuan dari berbagai pihak. Oleh karena itu, penulis ingin menyampaikan rasa terima kasih yang sebesar-besarnya kepada:

1. Keluarga kecil dan kedua orang tua, yang selalu memberikan doa, kasih sayang, serta dukungan moral dan material selama masa perkuliahan hingga selesainya skripsi ini.
2. Bapak Prof. Dr. M. Suyanto, M.M., selaku Rektor Universitas AMIKOM Yogyakarta.
3. Ibu Prof. Dr. Kusrini, S.Kom., M.Kom., selaku Dekan Fakultas Ilmu Komputer Universitas AMIKOM Yogyakarta.
4. Bapak I Made Artha Agastya, S.T., M.Eng., Ph.D., selaku dosen pembimbing yang selalu memberi ilmu dan arahannya yang sangat berharga dalam penyusunan skripsi ini.
5. Orang-orang terdekat, yang selalu memberikan dukungan, semangat, serta motivasi dalam menyelesaikan skripsi ini.

Yogyakarta, 24 Februari 2026

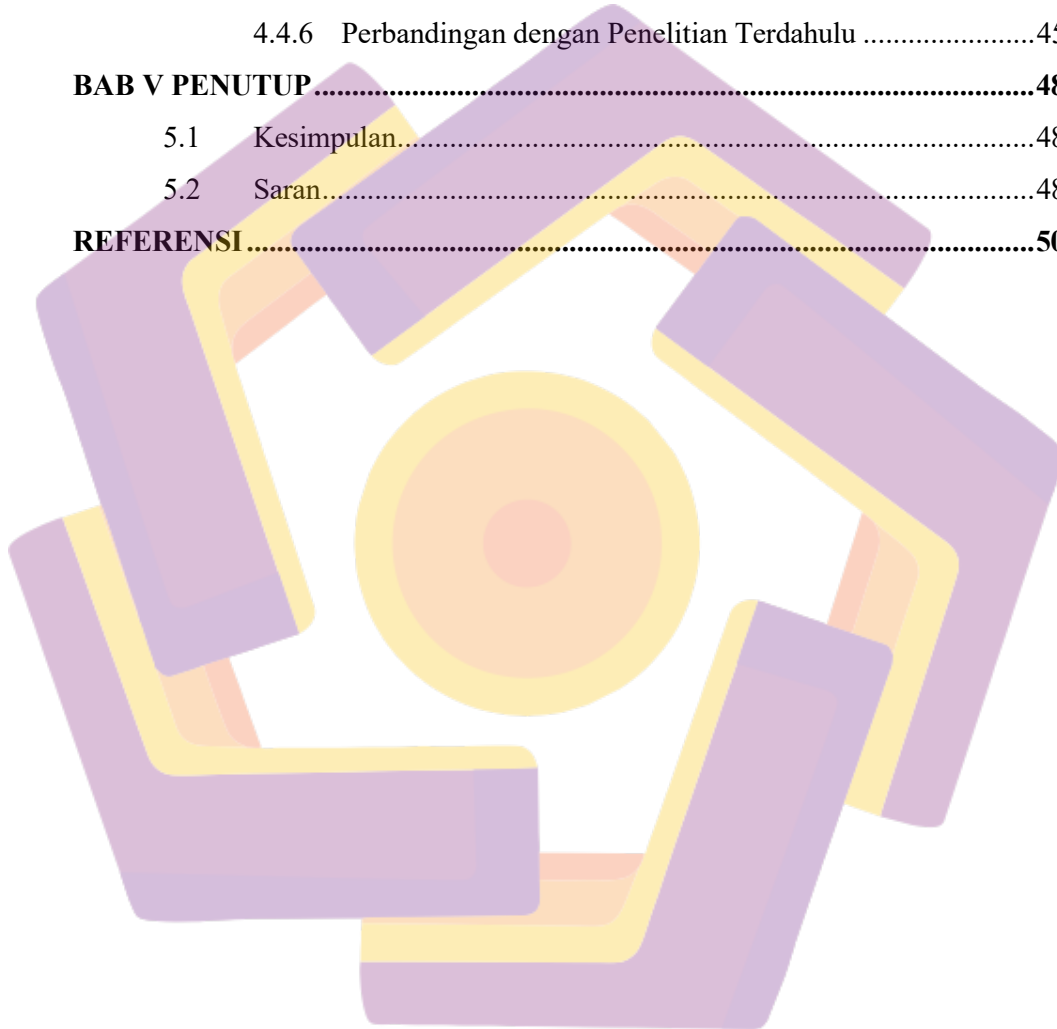
Penulis

DAFTAR ISI

HALAMAN PERSETUJUAN	ii
HALAMAN PENGESAHAN	iii
HALAMAN PERNYATAAN KEASLIAN SKRIPSI	iv
HALAMAN JUDUL	v
HALAMAN PERSEMBAHAN	vi
KATA PENGANTAR	vii
DAFTAR ISI	viii
DAFTAR TABEL	xi
DAFTAR GAMBAR	xii
DAFTAR LAMBANG DAN SINGKATAN	xiii
DAFTAR ISTILAH	xiv
INTISARI	xv
ABSTRACT	xvi
BAB I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	2
1.3 Batasan Masalah.....	2
1.4 Tujuan Penelitian.....	3
1.5 Manfaat Penelitian.....	3
1.5.1 Manfaat Teoritis	3
1.5.2 Manfaat Praktis.....	3
1.6 Sistematika Penulisan	4
BAB II TINJAUAN PUSTAKA	6
2.1 Studi Literatur	6
2.2 Dasar Teori.....	13
2.2.1 Data Mining.....	13
2.2.2 Teks Mining.....	13
2.2.3 Natural Language Processing (NLP)	13
2.2.4 Klasifikasi Teks	13

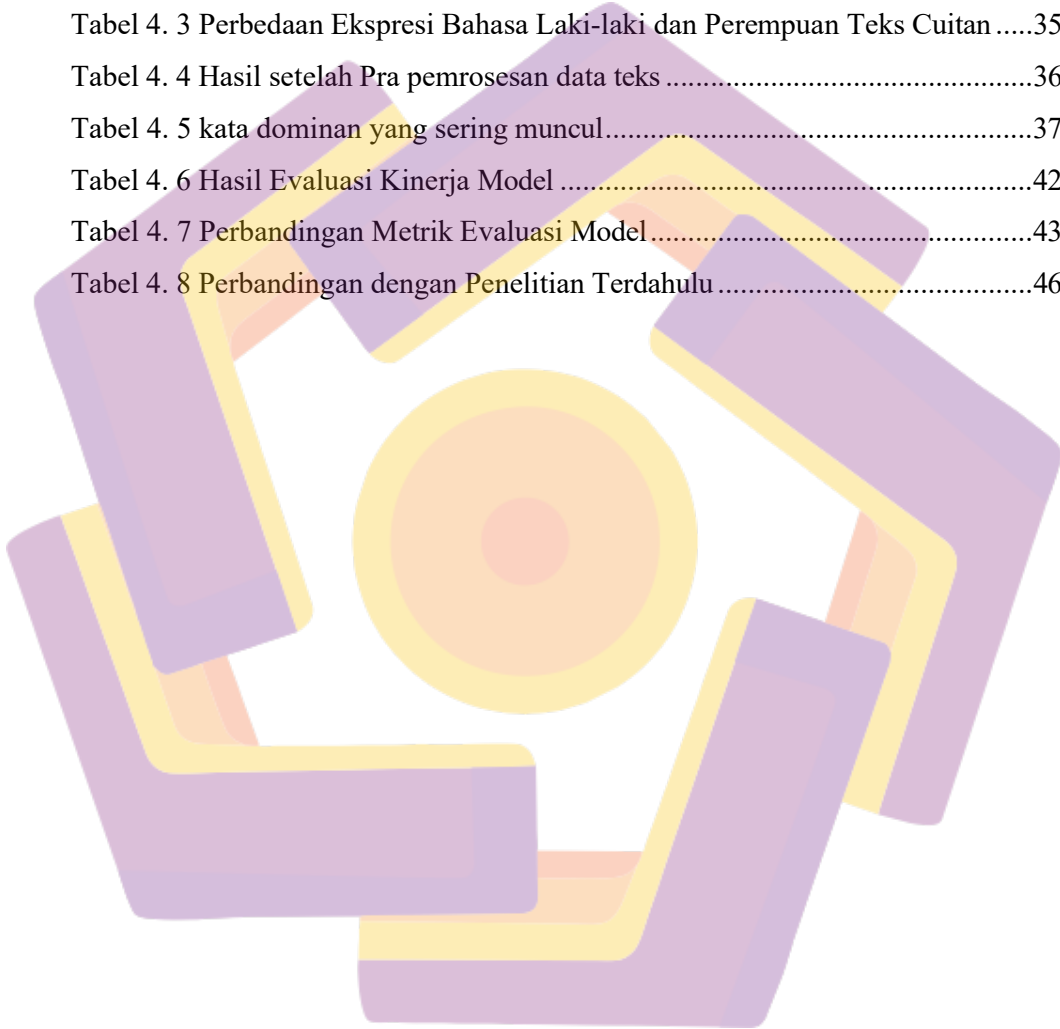
2.2.5	Media Sosial X	14
2.2.6	Pra-pemrosesan Data Teks	14
2.2.7	Pembobotan TF-IDF.....	15
2.2.7.4	Logistic Regression	17
2.2.7.5	Random Forest.....	18
2.3	Mekanisme Random Forest dalam Proses Klasifikasi	18
2.3.1	Support Vector Machine (SVM)	19
2.4	Ilustrasi Hyperplane dan Support Vector pada Support Vector Machine.....	20
2.4.1	Analisis Linguistik Gender.....	20
2.4.2	Evaluasi Model.....	21
2.4.3	K-Fold Cross Validation.....	23
BAB III METODE PENELITIAN.....		25
3.1	Objek Penelitian	25
3.2	Alur Penelitian.....	25
3.2.1	Sumber Dataset.....	25
3.2.2	Penyaringan dan Seleksi Data Awal.....	25
3.2.3	Pelabelan data	26
3.2.4	Pra pemrosesan teks.....	26
3.2.5	Ekstraksi Fitur (TF-IDF)	26
3.2.6	Pemodelan dan Validasi Model.....	27
3.2.7	Evaluasi Kinerja Model	27
3.3	Alat dan Bahan	28
3.3.1	Data Penelitian.....	28
3.3.2	Alat dan Instrumen	28
BAB IV HASIL DAN PEMBAHASAN.....		30
4.1	Pengumpulan data Persiapan Data	30
4.1.1	Acuan Linguistik untuk Interpretasi Pola Bahasa	33
4.2	Hasil Pra-Pemrosesan Data Teks	35
4.2.1	Analisis Kata Dominan dan Khas Berdasarkan Gender	36
4.3	Ekstraksi Fitur TF-IDF.....	41

4.4	Hasil dan Evaluasi	42
4.4.1	Perbandingan Metrik Evaluasi Model	43
4.4.2	Logistic Regression	44
4.4.3	Random Forest.....	44
4.4.4	Support Vector Machine.....	44
4.4.5	Analisis perbandingan antar algoritma	45
4.4.6	Perbandingan dengan Penelitian Terdahulu	45
BAB V PENUTUP		48
5.1	Kesimpulan.....	48
5.2	Saran.....	48
REFERENSI		50



DAFTAR TABEL

Tabel 2. 1 Keaslian Penelitian.....	8
Tabel 4. 1 Contoh Proses Verifikasi Label Dataset.....	31
Tabel 4. 2 Karakteristik Pola Bahasa Laki-laki dan Perempuan [24]	34
Tabel 4. 3 Perbedaan Ekspresi Bahasa Laki-laki dan Perempuan Teks Cuitan	35
Tabel 4. 4 Hasil setelah Pra pemrosesan data teks	36
Tabel 4. 5 kata dominan yang sering muncul.....	37
Tabel 4. 6 Hasil Evaluasi Kinerja Model	42
Tabel 4. 7 Perbandingan Metrik Evaluasi Model.....	43
Tabel 4. 8 Perbandingan dengan Penelitian Terdahulu	46



DAFTAR GAMBAR

Gambar 2. 1 Kurva Sigmoid pada Logistic Regression 8.....	17
Gambar 2. 2 Pohon Keputusan.....	18
Gambar 2. 3 Ilustrasi Hyperplane.....	20
Gambar 2. 4 Ilustrasi Proses K-Fold Cross Validation	23
Gambar 3. 1 Alur Penelitian.....	27
Gambar 4. 1 Distribusi Dataset Berdasarkan Jenis Kelamin Pengguna	33
Gambar 4. 2 world cloud.....	38
Gambar 4. 3 Kata paling khas Male	39
Gambar 4. 4 Kata Paling Khas Female	40
Gambar 4. 5 Hasil ekstraksi fitur TF-IDF	41

DAFTAR LAMBANG DAN SINGKATAN

x : Data input berupa teks cuitan

y : Hasil klasifikasi

w : Bobot fitur pada model klasifikasi

b : Bias pada model klasifikasi

$P(y|x)$: Probabilitas kelas terhadap data input

tf : Frekuensi kemunculan kata dalam dokumen

idf : Inverse document frequency

d : Dokumen

N : Jumlah dokumen dalam dataset

$h_t(x)$: Prediksi dari model atau pohon ke- t

t : Indeks model atau pohon

TP : True Positive

TN : True Negative

FP : False Positive

FN : False Negative

μ : Nilai rata-rata

σ : Standar deviasi

DAFTAR ISTILAH

Cuitan : Teks yang diunggah pengguna pada media sosial X.

Dataset : Kumpulan data cuitan yang digunakan dalam penelitian.

Label : Kategori kelas yang diberikan pada data, yaitu male dan female.

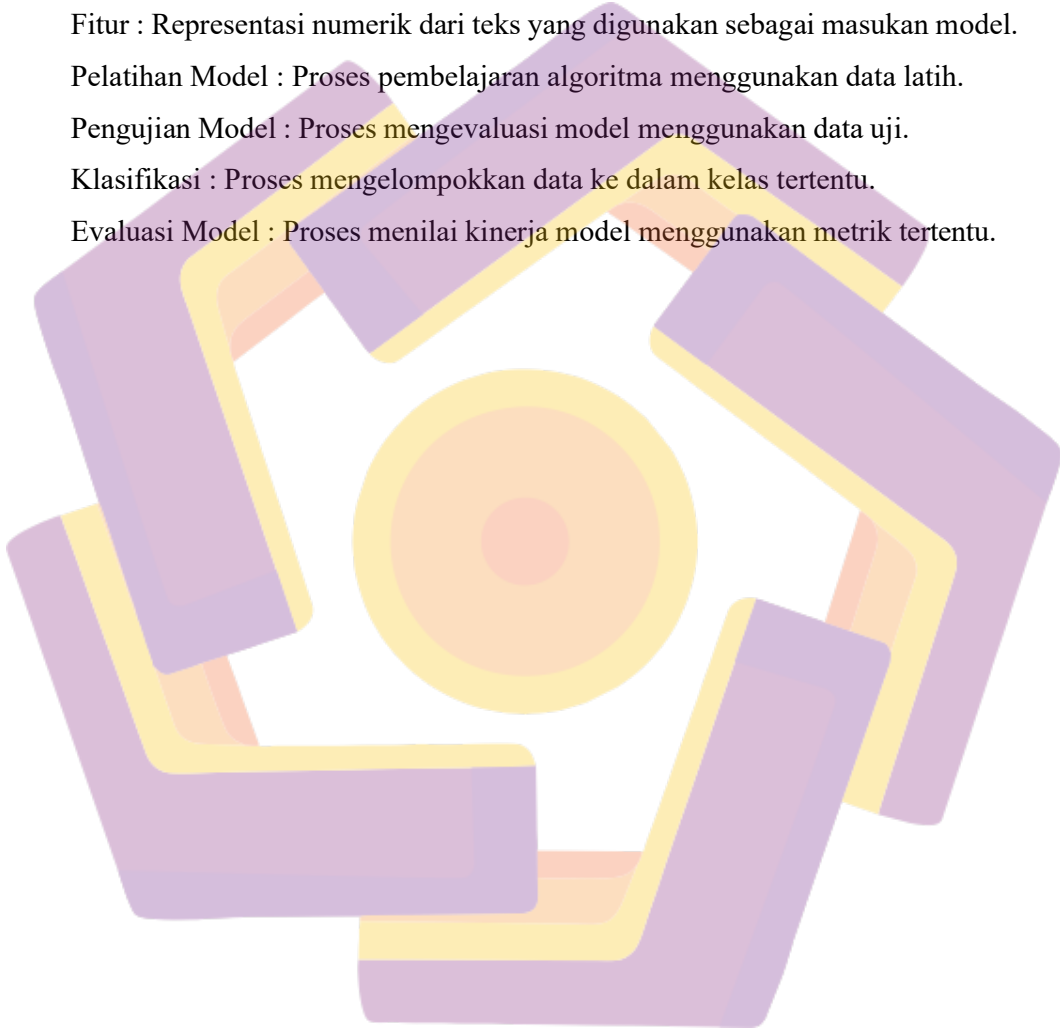
Fitur : Representasi numerik dari teks yang digunakan sebagai masukan model.

Pelatihan Model : Proses pembelajaran algoritma menggunakan data latih.

Pengujian Model : Proses mengevaluasi model menggunakan data uji.

Klasifikasi : Proses mengelompokkan data ke dalam kelas tertentu.

Evaluasi Model : Proses menilai kinerja model menggunakan metrik tertentu.



INTISARI

Penelitian ini dilatarbelakangi oleh banyaknya data teks yang dihasilkan pengguna media sosial X, sementara informasi demografis seperti jenis kelamin tidak selalu ditampilkan secara terbuka. Kondisi ini menyebabkan identifikasi karakteristik pengguna menjadi sulit dilakukan secara langsung. Oleh karena itu, penelitian ini bertujuan untuk mengklasifikasikan jenis kelamin pengguna berdasarkan teks cuitan menggunakan pendekatan Natural Language Processing dan machine learning. Dataset diperoleh dari Kaggle dengan jumlah awal sekitar 20.000 data, kemudian dilakukan penyaringan dan pelabelan ulang hingga diperoleh 1.026 data yang terdiri dari 456 data perempuan dan 570 data laki-laki. Tahap penelitian meliputi pra-pemrosesan teks, ekstraksi fitur menggunakan TF-IDF, serta pemodelan menggunakan algoritma Logistic Regression, Random Forest, dan Support Vector Machine. Evaluasi dilakukan menggunakan K-Fold Cross Validation dengan metrik accuracy, precision, recall, dan F1-score. Hasil penelitian menunjukkan adanya perbedaan pola bahasa antara pengguna laki-laki dan perempuan yang dapat dimanfaatkan dalam proses klasifikasi. Algoritma Support Vector Machine menghasilkan performa terbaik dengan akurasi rata-rata 89,76%. Penelitian ini diharapkan dapat dimanfaatkan dalam pengembangan sistem analisis pengguna media sosial, pemasaran digital, serta penelitian lanjutan di bidang analisis teks.

Kata kunci: klasifikasi gender, teks media sosial, TF-IDF, machine learning, Support Vector Machine

ABSTRACT

This research is motivated by the large amount of textual data generated by users on the social media platform X, while demographic information such as gender is not always displayed openly. This condition makes it difficult to directly identify user characteristics. Therefore, this study aims to classify user gender based on tweet text using Natural Language Processing and machine learning approaches. The dataset was obtained from Kaggle with an initial size of around 20,000 entries, then filtered and relabeled to produce 1,026 samples consisting of 456 female and 570 male data. The research stages include text preprocessing, feature extraction using TF-IDF, and classification modeling using Logistic Regression, Random Forest, and Support Vector Machine algorithms. Model evaluation was carried out using K-Fold Cross Validation with accuracy, precision, recall, and F1-score metrics. The results show that there are differences in language patterns between male and female users that can be utilized in the classification process. Among the tested models, Support Vector Machine achieved the best performance with an average accuracy of 89.76%. This research is expected to be useful for social media user analysis systems, digital marketing applications, and further research in text analytics.

Keyword: *gender classification, social media text, TF-IDF, machine learning, Support Vector Machine*