

PENGGUNAAN ALGORITMA GENETIKA UNTUK SELEKSI FITUR DALAM ANALISIS KEPRIHDADIAN BERDASARKAN CAPTION TWEET

M. Irsyad Hasbadi¹, Nelly Indriani Widiastuti, S.Si., M.T.²

^{1,2} Universitas Komputer Indonesia

Jl. Dipati Ukur No. 112-116, Lebak Gede, Kecamatan Coblong, Kota Bandung, Jawa Barat

E-mail : irsyadhasbadi@gmail.com¹, nelly.indriani@email.unikom.ac.id²

Abstrak

Penelitian ini mengimplementasikan metode seleksi fitur berbasis Algoritma Genetika (GA) untuk klasifikasi kepribadian artis atau tokoh publik Indonesia berdasarkan *caption tweet* dengan menggunakan pendekatan *Myers – Briggs Type Indikator* (MBTI). GA merupakan algoritma berbasis evolusi biologis yang bekerja melalui proses seleksi *crossover* dan mutasi untuk mencari kombinasi fitur terbaik. Dalam penelitian ini, GA akan digunakan untuk menyeleksi fitur (kata) pada *caption* untuk proses klasifikasi. Pengujian akan dilakukan dengan 3 skenario jumlah individu dalam populasi (10, 15, dan 20 Individu) dengan kriteria pemberhentian 50 generasi yang akan dilakukan pada empat label kepribadian : *Introvert – Ekstrovert* (IE), *Intuition – Sensing* (NS), *Thinking – Feeling* (TF), dan *Judging – Perceiving* (JP). Hasil penelitian menunjukkan bahwa performa terbaik diperoleh oleh label IE dengan performa akurasi 0,65 pada populasi 20 individu, sementara label lain cenderung memiliki performa dibawah 0,60. Hasil yang didapat menunjukkan bahwa performa seleksi fitur dipengaruhi oleh variasi jumlah populasi, tetapi kualitas data dan keseimbangan antar label juga berpengaruh. Dengan demikian GA berpotensi menjadi seleksi fitur untuk klasifikasi kepribadian berbasis teks, meskipun diperlukan optimasi lebih lanjut untuk menghadapi karakteristik data tweet yang dinamis dan tidak terstruktur.

Kata kunci : Algoritma Genetika, Seleksi Fitur, MBTI, Klasifikasi Teks

Abstract

This study implements a feature selection method based on Genetic Algorithm (GA) for personality classification of Indonesian public figures or artists using caption tweets with the Myers–Briggs Type Indicator (MBTI) approach. GA is an evolutionary-based algorithm that operates through the processes of selection, crossover, and mutation to search for the optimal feature combination. In this research, GA is applied to select the most relevant features (words) from captions for the classification process. The evaluation was conducted under three scenarios of population sizes (10, 15, and 20 individuals) with a termination criterion of 50 generations, across four MBTI personality dimensions: Introvert–Extrovert (IE), Intuition–Sensing (NS), Thinking–Feeling (TF), and Judging–Perceiving (JP). The results show that the best performance was achieved in the IE dimension with an accuracy of 0.65 using a population of 20 individuals, while the other dimensions tended to perform below 0.60. These findings indicate that the performance of feature selection is influenced not only by the variation in population size but also by the quality of data and class balance. Thus, GA has the potential to be applied as a feature selection method for text-based personality classification, although further optimization is required to address the dynamic and unstructured nature of tweet data.

Keywords : Genetic Algorithm, Feature Selection, MBTI, Text Classification.

1. PENDAHULUAN

Kepribadian merupakan karakter unik yang menjadi pembeda antar individu satu dengan individu lain. Salah satu metode yang cukup populer untuk mengklasifikasikan kepribadian adalah *Myers – Briggs Type Indikator* (MBTI). Metode ini dikembangkan berdasarkan teori Carl Jung oleh Katharine Cook Briggs dan Isabel Briggs Myers pada saat masa Perang Dunia II [1] [2]. MBTI mengklasifikasikan individu kedalam empat dimensi kepribadian, yaitu *Introvert – Ekstrovert* (IE), *Intuition – Sensing* (NS), *Thinking – Feeling* (TF), dan *Judging – Perceiving* (JP) [3].

Seiring perkembangan zaman dan teknologi, media sosial Twitter (X) menjadi sumber data yang kaya untuk memahami perilaku dan kepribadian seseorang. Dengan lebih 600 juta pengguna aktif dan 250 juta

tweet setiap harinya, Twitter (X) menyediakan data teks yang sangat beragam. Namun, karakteristik pada *caption* tweet seperti penggunaan bahasa slang, campuran antara bahasa formal dan informal, serta emotikon dan *hashtag* menjadi tantangan dalam proses analisis kepribadian berbasis *caption* tweet, khususnya pada tahap seleksi fitur [4] [5].

Seleksi fitur menjadi salah satu tahap penting dalam proses klasifikasi teks karena mempengaruhi pola klasifikasi dan performa akurasi model [6]. Metode seleksi fitur berbasis *wrapped* seringkali digunakan karena menghasilkan kinerja dan performa yang baik dalam mengevaluasi subset fitur untuk proses klasifikasi. Meskipun memiliki waktu komputasi yang cukup tinggi, pendekatan *wrapped*, mampu menghasilkan performa akurasi yang baik [7]. Penelitian terdahulu menunjukkan bahwa penerapan *wrapped* dengan Algoritma Genetika (GA) dapat meningkatkan performa akurasi klasifikasi secara signifikan hingga 99% dalam mengoptimalkan Extreme *Learning Machine* (ELM) [7].

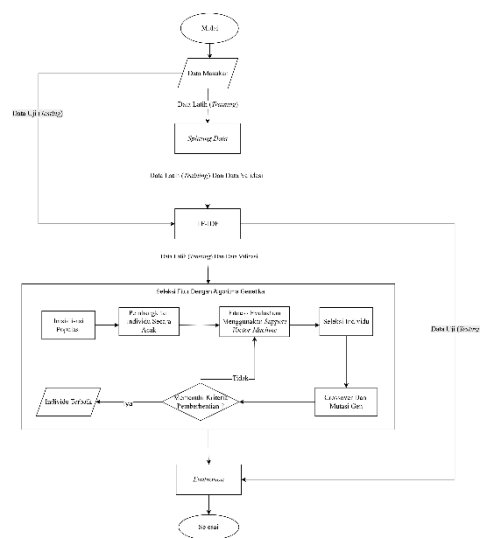
Oleh karena itu, penelitian ini akan menggunakan pendekatan *wrapped* dengan Algoritma Genetika (GA) sebagai teknik seleksi fitur. GA sendiri merupakan algoritma yang berbasis pada teori evolusi yang mampu melakukan seleksi fitur untuk algoritma induk. Beberapa penelitian sebelumnya menunjukkan bahwa penggabungan GA dengan algoritma klasifikasi terbukti berhasil meningkatkan performa akurasi secara signifikan. Misalnya, performa akurasi model klasifikasi dengan *Support Vector Machine* (SVM) sebesar 69,32% meningkat menjadi 97,97% setelah dilakukan seleksi fitur dengan GA. Hal ini menunjukkan bahwa SVM merupakan algoritma klasifikasi yang cocok jika dipadukan dengan GA dalam konteks teks mining dengan data berdimensi tinggi [6].

Berdasarkan kondisi yang sudah dijelaskan sebelumnya, penelitian ini bertujuan untuk menganalisis kepribadian artis atau tokoh publik Indonesia melalui *caption* dari Twitter (X) dengan menggunakan pendekatan MBTI. Untuk mengatasi tantangan dalam seleksi fitur pada data teks yang tidak terstruktur, penelitian ini akan menerapkan pendekatan *wrapped* dengan GA sebagai metode seleksi fitur serta SVM sebagai algoritma klasifikasi.

2. METODOLOGI

2.1 Alur Penelitian

Alur penelitian merupakan gambaran proses awal hingga akhir yang dilakukan didalam penelitian. Tahap awal pada penelitian ini yaitu menganalisis data masukan. Yang dimana data yang digunakan dalam penelitian ini sudah dilakukan proses *preprocessing* dan *split* (pemisahan) antara data latih (*training*) dan data uji (*testing*) yaitu 80% data latih dan 20% data uji . Setelah dilakukan analisis, data latih akan dibagi kembali menjadi dua, yaitu 80% data uji dan 20% data validasi. Proses selanjutnya adalah ekstraksi fitur dengan TF-IDF untuk memberi bobot pada kata. Kemudian dilakukan seleksi fitur dengan Algoritma Genetika yang mencakup beberapa tahapan yaitu inisialisasi populasi, pembangkitan individu, *fitness evaluation*, seleksi individu, *crossover* dan mutasi untuk memperoleh individu dengan susunan gen (fitur) terbaik. Hasil dari seleksi fitur selanjutnya akan diuji menggunakan *Support Vector Machine* (SVM) pada data uji untuk mengevaluasi performa akurasi model. Seluruh tahapan penelitian akan digambarkan pada diagram alur sistem dalam gambar 1.



Gambar 1. Diagram Alur Model

2.2 Data Masukan

Data yang digunakan dalam penelitian ini merupakan data yang sudah siap digunakan yang telah melalui proses pengolahan sebelumnya. Data mentah awalnya dikumpulkan melalui proses *web scraping* dari platform media sosial Twitter (X) tokoh publik atau artis Indonesia, kemudian dilakukan proses pelabelan menggunakan pendekatan MBTI. Kemudian data dilakukan *preprocessing* dan *split* data untuk memisahkan 80% data latih (*training*) dan 20% data uji (*testing*). Berikut Adalah hasil *split* untuk data latih dan data uji pada tabel 1.

Tabel 1. Distribusi Data Latih Dan Data Uji

No	Data Uji Atau Data Latih (<i>Train</i>)	Jumlah Data
1.	Data Uji (<i>Testing</i>)	2300
2.	Data Latih (<i>Training</i>)	9350

2.3 Splitting Data

Proses *split* data dilakukan untuk membagi 80% data latih yang sudah disiapkan sebelumnya menjadi 2, yaitu 80% data latih dan 20% data validasi. Data validasi digunakan untuk mengevaluasi performa model selama proses pelatihan berlangsung. Adapun hasil dari *split* data latih dan data validasi pada tabel 2 sampai tabel 3.

Tabel 2. Distribusi Data Latih Dan Data Validasi

No	Data Uji Atau Data Latih (<i>Train</i>)	Jumlah Data
1.	Data Latih (<i>Training</i>)	7480
2.	Data Validasi	1870

Tabel 3. Distribusi Data Latih Dan Data Validasi Pada Setiap Label

No	Data Validasi Atau Data Latih (<i>Train</i>)	Label	Jumlah Data
1.	Distribusi Data Latih	I (<i>Introvert</i>)	2640
		E (<i>Ekstrovert</i>)	4840
	Distribusi Data Validasi	I (<i>Introvert</i>)	600
		E (<i>Ekstrovert</i>)	1210
2.	Distribusi Data Latih	N (<i>Intuition</i>)	3979
		S (<i>Sensing</i>)	3501
	Distribusi Data Validasi	N (<i>Intuition</i>)	971
		S (<i>Sensing</i>)	899
3.	Distribusi Data Latih	T (<i>Thinking</i>)	2715
		F (<i>Feeling</i>)	4765
	Distribusi Data Validasi	T (<i>Thinking</i>)	635
		F (<i>Feeling</i>)	1235
4.	Distribusi Data Latih	J (<i>Judging</i>)	3322
		P (<i>Perceiving</i>)	4158
	Distribusi Data Validasi	J (<i>Judging</i>)	828

		P (Perceiving)	1042
--	--	----------------	------

2.4 Term Frequency – Inverse Document Frequency (TF-IDF)

Term Frequency – Inverse Document Frequency (TF-IDF) merupakan metode yang digunakan untuk memberikan pembobotan pada setiap kata berdasarkan kemunculan dalam suatu dokumen [8]. Kata yang sering muncul pada suatu dokumen (tweet) tetapi jarang muncul pada dokumen (tweet) lain akan memiliki bobot yang tertinggi sehingga dianggap lebih penting. Didapatkan persamaan untuk mencari nilai TF – IDF pada persamaan (1) sampai (4).

$$tf - idf(t, d) = tf(t, d) * idf(t) \quad (1)$$

$$tf(t, d) = \frac{\text{jumlah } t \text{ didalam sebuah } d}{\text{jumlah kata dalam } d} \quad (2)$$

$$df(t) = \text{jumlah dokumen yang memuat term } t \quad (3)$$

$$idf(t) = \log\left(\frac{N}{df(t)}\right) \quad (4)$$

2.5 Seleksi Fitur Dengan Algoritma Genetika

Seleksi fitur dengan Algoritma Genetika digunakan untuk memilih subet (kumpulan fitur) terbaik yang nantinya akan dipakai didalam proses klasifikasi. Tujuannya Adalah mengurangi kompleksitas yang tidak relevan sehingga bisa memaksimalkan performa akurasi model. Dalam penelitian ini, evaluasi fitur akan dilakukan menggunakan *Support Vector Machine* (SVM) sehingga hanya fitur yang paling relevan yang akan di pertahankan [4]. Tahapan dari Algoritma Genetika meliputi :

A. Inisialisasi Populasi

Individu dibangkitkan secara acak yang masing – masing individu merepresentasikan fitur (kata) yang disebut dengan gen. Setiap gen memiliki bobot yang disebut dengan sebagai nilai *allele* [9]. Nilai *allele* diambil dari hasil ekstrasi fitur dengan TF-IDF yang sebelumnya sudah dilakukan. Sementara sekumpulan individu akan disebut sebagai populasi. Pada penelitian ini, percobaan akan dilakukan dengan menggunakan jumlah populasi yang berbeda yaitu 10,15 dan 20 dengan batasan *max_features* = 1000 untuk panjang masing – masing individu [10].

B. Fitness Evaluation

Setiap individu didalam populasi, akan dievaluasi berdasarkan performa akurasi model *Support Vector Machine*. Setelah dilakukan analisa pada distribusi data, hasil menunjukkan bahwa data tidak bisa menggunakan pendekatan linear. Maka pendekatan kernel yang digunakan akan menggunakan pendekatan *Radial Basis Function* (RBF). Maka didapati persamaan yang akan digunakan Adalah persamaan (5) sampai persamaan (9) [11] [12].

$$f(x) = \text{sign}\left(\sum_{i=1}^n \alpha_i y_i K(x_i, x) + b\right) \quad (5)$$

Yang dimana :

$f(x)$ = fungsi keputusan dari *Support Vector Machine* (SVM)

sign = Fungsi sign yaitu : $\text{sign}(f(x)) = \begin{cases} +1 & \text{jika } f(x) > 0 \\ -1 & \text{jika } f(x) < 0 \end{cases}$

$$\|x - y\|^2 \quad (6)$$

Yang dimana :

$\|x - y\|^2$ = Merupakan jarak euclidean kuadrat antara dua vektor

$$k\gamma(x, y) = \exp(-\gamma \|x - y\|^2) \quad (7)$$

Yang dimana :

$\|x - y\|^2$ = Merupakan jarak euclidean kuadrat antara dua vektor

γ = Adalah parameter positif yang mengatur lebar kernel

$$\max_{\alpha} L(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j x_i^T x_j \quad (8)$$

Dengan syarat:

1. $\sum_{i=1}^n \alpha_i y_i = 0$
2. $\alpha_i \geq 0$, untuk setiap α_i
3. $w = \sum \alpha_i y_i \bar{x}_i$ dan $b = y_k - \bar{w}^T \bar{x}_k$ untuk setiap \bar{x}_k sedemikian hingga $\alpha_k \neq 0$
4. Untuk setiap α_i yang tidak nol, mengidentifikasi \bar{x} adalah *support vector*.

$$b = y_k - \sum_{j=i}^n \alpha_i y_j K(x_j, x_k) \quad (9)$$

Dimana :

y_i = Label data

α_i = *Lagrange Multiplier* dari persamaan (2.10)

Dengan α_i pada titik i dimana $0 < \alpha_i <$

Untuk mencari nilai akurasi maka dibutuhkan persamaan (10) sampai persamaan (13)

$$acc = \frac{(TP + TN)}{TP + TN + FP + FN} \quad (10)$$

$$Precision = \frac{TP}{TP + FP} \quad (11)$$

$$Recall = \frac{TP}{TP + FN} \quad (12)$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (13)$$

Yang dimana :

TP = Prediksi model positif yang mengidentifikasi kelas positif

TN = Prediksi model negatif yang mengidentifikasi kelas negatif

FP = Prediksi model salah memprediksi kelas negatif menjadi kelas positif

FN = Prediksi model salah memprediksi kelas positif menjadi kelas negatif

C. Seleksi Individu

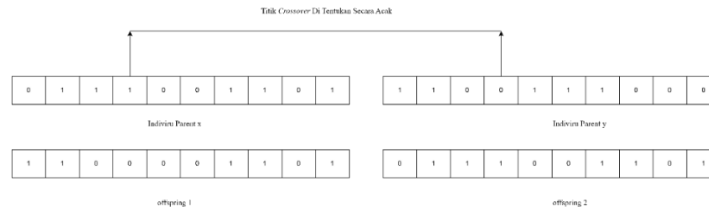
Setelah nilai *fitness* dihitung pada pembangkitan individu, tahap selanjutnya adalah seleksi individu untuk memilih individu terbaik yang nantinya akan disebut dengan *parents* yang akan digunakan dalam proses pembentukan generasi baru [13]. Individu dengan nilai *fitness* tertinggi memiliki peluang lebih besar untuk dipilih [14].

Metode yang digunakan dalam proses seleksi individu adalah *tournament selection*. Individu dalam populasi akan bersaing satu sama lain sampai ukuran individu pada populasi terpenuhi, sehingga jumlah *parents* dalam populasi sesuai dengan jumlah populasi yang sudah ditentukan. Individu dengan nilai *fitness* tertinggalah yang akan dianggap kemudian akan disebut dengan *parents*. *Parents* yang dihasilkan akan kemudian dilanjutkan kedalam tahap evolusi yaitu *crossover* dan mutasi [15].

D. Crossover

Proses rekombinasi atau *crossover* dilakukan dengan cara menggabungkan gen – gen dari *parents* sehingga akan dihasilkan keturunan yang akan disebut dengan (*offspring*) [14], dengan hasil kombinasi sifat genetik dari kedua *parent* yang digabungkan [16].

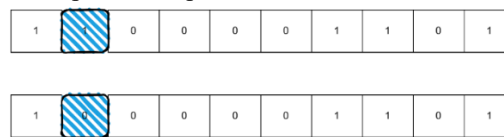
Metode yang digunakan dalam penelitian ini Adalah *one – point crossover* dengan probabilitas sebesar 0,9. Metode ini akan mengambil sebuah titik potong secara acak dari urutan gen pada pasangan *parents* untuk dilakukan pertukaran gen dengan tujuan untuk memperluas ruang eksplorasi dalam pencarian solusi. Hasil dari pertukaran gen antar *parents* akan menghasilkan individu baru yang disebut dengan *child* [11]. Dalam proses *crossover* nilai *fitness* tidak akan langsung dihitung kembali, melainkan setelah melalui proses mutase [12]. Adapun proses *crossover* akan disimulasikan dalam bentuk gambar 2.



Gambar 2. Gambaran Proses *Crossover*

E. Mutasi

Setelah melalui proses *crossover*, individu baru (*child*) yang dihasilkan akan mengalami tahap mutasi. Mutasi dilakukan dengan cara mengubah gen secara acak dengan tingkat probabilitas yang kecil, sehingga memungkinkan munculnya variasi dalam populasi [11]. Probabilitas mutasi biasanya berada pada rentang 0,01 hingga 0,1 per-gen. Jika probabilitas terlalu tinggi, maka Algoritma Genetika dapat menjadi terlalu acak sehingga berisiko gagal menemukan solusi terbaik [11]. Pada penelitian ini, probabilitas mutasi ditetapkan sebesar 0,1 atau 10%, yang berarti apabila terdapat 30 gen pada satu individu, maka kemungkinan gen yang dimutasi dibatasi maksimal 3 gen. Ilustrasi proses mutasi dapat dilihat pada Gambar 2.3.



Gambar 3. Gambaran Proses Mutasi

Kemudian nilai *fitness* akan kembali dihitung dengan cara yang sama dengan menghitung *fitness* sebelumnya

F. Kriteria Pemberhentian

Kriteria pemberhentian dalam Algoritma Genetika (GA) diperlukan untuk menentukan kapan proses evolusi harus dihentikan. Pemberhentian dapat ditentukan berdasarkan beberapa faktor, seperti tercapainya keseimbangan antara kualitas solusi dan sumber daya komputasi yang digunakan. Secara umum, kriteria pemberhentian mencakup penetapan jumlah generasi maksimum [11]. Pada penelitian ini, kriteria pemberhentian ditetapkan berdasarkan jumlah generasi maksimum yaitu sebanyak 50 generasi. Dengan demikian, proses GA akan dihentikan ketika seluruh generasi yang direncanakan telah dijalankan.

3 HASIL DAN PEMBAHASAN

3.1 HASIL

Pada penelitian ini dilakukan 3 skenario pengujian seleksi fitur dengan 3 variasi jumlah populasi pada proses Algoritma Genetika (GA) yaitu 10,15 dan 20. Selain itu pengujian tanpa seleksi fitur dan pengujian metode *filtering* dengan menggunakan metode *Chi – Square* juga coba. Hal ini bertujuan untuk melihat pengaruh perubahan jumlah populasi, dan pengaruh GA terhadap kualitas seleksi fitur pada performa akurasi pada model klasifikasi. Setiap skenario dianalisis berdasarkan performa akurasi pada masing-masing label kepribadian.

A. Hasil Pengujian Skenario 1 (Populasi Sebanyak 10 Individu)

Pengujian skenario 1 akan menggunakan jumlah individu pada populasi sebanyak 10. Hasil pengujian skenario 1 akan disajikan dalam bentuk tabel yaitu pada tabel 4.

Tabel 4. Hasil Pengujian Skenario 1

Hasil Pengujian 10 Pop					
Label	Kelas	Precision	Recall	F1-Score	Accuracy

IE	E (0)	0.68	0.84	0.75	0.63
IE	I (1)	0.37	0.20	0.26	0.63
NS	N (0)	0.60	0.66	0.63	0.54
NS	S (1)	0.44	0.39	0.42	0.54
TF	F (0)	0.61	0.83	0.70	0.57
TF	T (1)	0.41	0.19	0.26	0.57
JP	J (0)	0.70	0.18	0.28	0.55
JP	P (1)	0.53	0.92	0.67	0.55

B. Hasil Pengujian Skenario 2 (Populasi Sebanyak 15 Individu)

Kemudian pengujian skenario 2 akan menggunakan jumlah individu pada populasi sebanyak 15. Hasil pengujian skenario 2 akan disajikan dalam bentuk tabel yaitu pada tabel 5.

Tabel 5. Hasil Pengujian Skenario 2

Hasil Pengujian 15 Pop					
Label	Kelas	<i>Precision</i>	<i>Recall</i>	<i>F1-Score</i>	<i>Accuracy</i>
IE	E (0)	0.68	0.82	0.74	0.61
IE	I (1)	0.33	0.19	0.24	0.61
NS	N (0)	0.61	0.63	0.62	0.54
NS	S (1)	0.45	0.42	0.43	0.54
TF	F (0)	0.62	0.83	0.71	0.58
TF	T (1)	0.43	0.19	0.26	0.58
JP	J (0)	0.61	0.34	0.43	0.55
JP	P (1)	0.54	0.78	0.64	0.55

C. Hasil Pengujian Skenario 3 (Populasi Sebanyak 20 Individu)

Selanjutnya pengujian skenario 3 akan menggunakan jumlah individu pada populasi sebanyak 20. Hasil pengujian skenario 3 akan disajikan dalam bentuk tabel yaitu pada tabel 6.

Tabel 6. Hasil Pengujian Skenario 3

Hasil Pengujian 20 pop					
Label	Kelas	<i>Precision</i>	<i>Recall</i>	<i>F1-Score</i>	<i>Accuracy</i>
IE	E (0)	0.68	0.92	0.78	0.65
IE	I (1)	0.37	0.09	0.15	0.65
NS	N (0)	0.60	0.80	0.69	0.57
NS	S (1)	0.47	0.24	0.32	0.57
TF	F (0)	0.61	0.86	0.71	0.58
TF	T (1)	0.40	0.15	0.22	0.58
JP	J (0)	0.57	0.33	0.42	0.54
JP	P (1)	0.53	0.75	0.62	0.54

D. Pengujian Tanpa Seleksi Fitur

Pengujian tanpa seleksi dengan menggunakan *Support Vektor Machine* (SVM) dengan pendekatan kernel yang sama yaitu kernel RBF dan *max_features* = 1000 pada TF – IDF. Hasil pengujian tanpa seleksi fitur akan disajikan dalam bentuk tabel, pada tabel 7.

Tabel 7. Hasil Pengujian Tanpa Seleksi Fitur
Hasil Pengujian Tanpa Seleksi Fitur Dengan SVM

Label	Kelas	<i>Precision</i>	<i>Recall</i>	<i>F1-Score</i>	<i>Accuracy</i>
IE	E (0)	0,67	0,80	0,73	0,60
IE	I (1)	0,33	0,20	0,25	0,60
NS	N (0)	0,60	0,64	0,62	0,54
NS	S (1)	0,44	0,41	0,42	0,54
TF	F (0)	0,62	0,77	0,69	0,57
TF	T (1)	0,43	0,27	0,34	0,57
JP	J (0)	0,56	0,72	0,63	0,57
JP	P (1)	0,61	0,42	0,50	0,57

E. Pengujian Dengan Metode *Chi - Square*

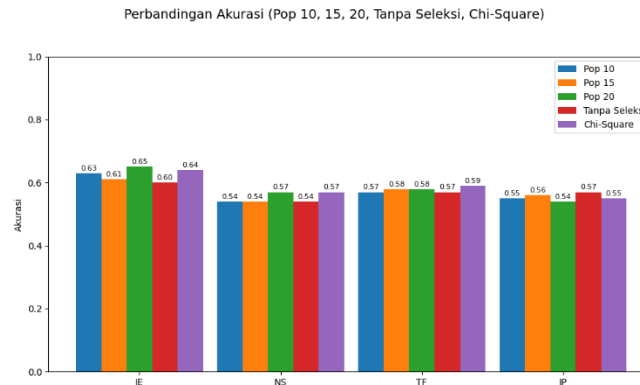
Sebagai pembanding, penelitian ini juga menerapkan metode seleksi fitur lain menggunakan pendekatan *filtering* berbasis *Chi-Square*. Proses seleksi dilakukan pada representasi teks dengan TF-IDF menggunakan parameter *max_features* = 1000, sehingga diperoleh 1000. Pengujian dilakukan dengan data latih dan data uji yang sama, maka didapati hasil pengujian sebagaimana ditunjukkan pada tabel 8.

Tabel 8. Hasil Pengujian *Chi - Square*
Hasil Pengujian *Chi - Square*

Label	Kelas	<i>Precision</i>	<i>Recall</i>	<i>F1-Score</i>	<i>Accuracy</i>
IE	E (0)	0,68	0,88	0,77	0,64
IE	I (1)	0,38	0,16	0,22	0,64
NS	N (0)	0,62	0,74	0,68	0,57
NS	S (1)	0,49	0,36	0,42	0,57
TF	F (0)	0,62	0,82	0,71	0,59
TF	T (1)	0,46	0,23	0,31	0,59
JP	J (0)	0,60	0,32	0,42	0,55
JP	P (1)	0,54	0,79	0,64	0,55

3.2 PEMBAHASAN

Pengujian pada seluruh skenario menunjukkan adanya variasi performa akurasi antar label kepribadian MBTI. Adapun hasil dari keseluruhan pengujian akan ditampilkan dalam bentuk diagram batang pada gambar 4.



Gambar 4. Diagram Batang Hasil Pengujian

Seperti yang dapat dilihat pada gambar 4, Secara umum, label IE konsisten menunjukkan performa akurasi terbaik pada ketiga skenario (0,61; 0,61; 0,65). Peningkatan ini mengindikasikan bahwa individu terbaik hasil seleksi GA cukup mampu membedakan karakteristik kepribadian Ekstrovert (E) dan Introvert (I). Hal tersebut dimungkinkan karena kosakata pada caption tweet masing-masing kelas cenderung lebih spesifik dan berbeda, sehingga fitur yang dipilih GA menjadi lebih representatif.

Untuk label NS, hasil pengujian memperlihatkan adanya peningkatan bertahap (0,54; 0,54; 0,57). Peningkatan terbesar terjadi pada recall kelas N di skenario 3 (0,80), meskipun recall pada kelas S justru menurun signifikan. Hal ini menunjukkan bahwa perbaikan fitur lebih optimal untuk kelas mayoritas, sementara kelas minoritas masih belum terakomodasi dengan baik.

Pada label TF, performa akurasi relatif stabil (0,57–0,58). Pola yang terlihat adalah recall tinggi pada kelas F (hingga 0,86 pada populasi dengan 20 individu), namun recall sangat rendah pada kelas T (<0,20) terutama saat jumlah individu lebih kecil. Konsistensi pola ini mengindikasikan bahwa fitur terpilih melalui GA cenderung lebih mewakili satu sisi kelas (F) dibandingkan T, sehingga distribusi performa antar kelas masih timpang.

Sementara itu, label JP memperlihatkan sedikit peningkatan performa akurasi pada skenario 1 dan 2 (0,55 – 0,56), namun menurun pada skenario 3 (0,54). Meskipun kelas P secara konsisten memiliki recall tinggi (0,78–0,92), ketidakseimbangan performa antar kelas berdampak pada rendahnya akurasi keseluruhan.

Jika dibandingkan dengan hasil pengujian tanpa seleksi fitur, terlihat bahwa GA memberi variasi positif terutama pada label IE dan sebagian NS. Akurasi awal tanpa seleksi fitur adalah 0,60 (IE), 0,54 (NS), 0,57 (TF), dan 0,57 (JP). Setelah penerapan GA, label IE meningkat hingga 0,65 pada skenario 3, NS mengalami peningkatan performa sedikit, menjadi 0,57, TF meningkat tipis ke 0,58, sementara JP justru menurun menjadi 0,57–0,54. Hal ini menegaskan bahwa GA lebih efektif untuk label IE dan sebagian NS, meskipun peningkatan akurasi secara keseluruhan relatif kecil dan tidak merata.

Sebagai perbandingan, pengujian juga dilakukan dengan Chi-Square. Pada label IE, GA unggul dengan akurasi tertinggi 0,65 dibanding Chi-Square (0,64) maupun tanpa seleksi fitur (0,60). Untuk label NS, performa GA (0,54–0,57) relatif setara dengan Chi-Square. Pada label TF, Chi-Square lebih baik dengan akurasi 0,59 dibanding GA (0,58). Sedangkan pada label JP, GA mencapai 0,56 pada skenario 2, hampir sama dengan Chi-Square (0,55), namun masih lebih rendah dari hasil tanpa seleksi fitur (0,57). Hasil ini menunjukkan bahwa GA dan Chi-Square memiliki keunggulan masing-masing pada label tertentu, dengan peningkatan akurasi yang tidak terlalu signifikan secara keseluruhan.

Temuan ini sejalan dengan penelitian terdahulu [6]. Yang dimana, seleksi fitur berbasis GA terbukti mampu meningkatkan akurasi klasifikasi teks melalui pemilihan fitur relevan pada pengujian dengan SVM. Demikian pula pada jurnal [16], GA terbukti dapat menghasilkan performa klasifikasi yang cukup baik, khususnya pada beberapa label kepribadian. Hasil penelitian ini menunjukkan bahwa peran GA memberikan kontribusi terhadap seleksi fitur dalam klasifikasi kepribadian, terutama pada dimensi IE dan sebagian NS. Namun, metode *filtering* seperti *Chi-Square* tetap mampu bersaing, bahkan lebih unggul pada beberapa label seperti TF. Dengan demikian, pemanfaatan GA pada penelitian ini memperlihatkan adanya kontribusi GA dalam proses seleksi fitur, meskipun peningkatan performanya masih terbatas dan tidak merata di seluruh dimensi kepribadian MBTI.

4 PENUTUP

Dari hasil pengujian, dapat disimpulkan bahwa penggunaan Algoritma Genetika (GA) tetap berperan dalam seleksi fitur pada klasifikasi kepribadian MBTI berbasis *caption* tweet, meskipun peningkatan

akurasi yang diperoleh relatif kecil dan tidak jauh berbeda dengan metode filtering dengan menggunakan pendekatan *Chi – Square*. Hal ini dipengaruhi oleh karakteristik data yang dinamis, kosakata dalam setiap *caption* yang acak, serta distribusi fitur yang kurang seimbang. Sehingga baik GA maupun *Chi – Square* menunjukkan performa akurasi yang hampir setara. Berbeda dengan penelitian terdahulu yang menggunakan data terstruktur atau analisis sentimen, pemetaan kepribadian dari *caption* tweet menghadirkan tantangan tersendiri [4]. Sehingga perbedaan antar performa tidak terlalu signifikan. Namun demikian, GA tetap berkontribusi dengan menunjukkan performa akurasi yang lebih baik pada beberapa label dibandingkan dengan metode lainnya. Untuk penelitian selanjutnya, disarankan mencoba pendekatan *filtering* seperti *Chi – Square* untuk mengurangi waktu komputasi [7], melakukan optimasi parameter GA seperti meningkatkan jumlah individu pada populasi maupun probabilitas pada proses evolusi untuk mencari parameter yang lebih efisien. Pengujian juga dapat diperluas dengan menggabungkan prediksi semua dimensi label MBTI menjadi satu label kepribadian penuh per akunnya. Lalu dilakukan perbandingan dengan label aslinya untuk memperoleh gambaran akurasi keseluruhan.

DAFTAR PUSTAKA

- [1] K. A. De Jong, An analysis of the behavior of a class of genetic adaptive systems, Ph.D. dissertation, University of Michigan, Ann Arbor, MI, USA, 1975.
- [2] A. T. Ramly, "Genealogical Critique of the MBTI (Myers Briggs Type Indicator)," 2011. [Online].
- [3] E. Susanto and M. Mudaim, "Pengembangan inventori MBTI sebagai alternatif instrumen pengukuran tipe kepribadian," *Indonesian Journal of Educational Counseling*, vol. 1, no. 1, pp. 41–52, 2017.
- [4] D. A. Putri, "Algoritma Support Vector Machine Berbasis Algoritma Genetika Untuk Analisis Sentimen Pada Twitter," Konferensi Nasional Ilmu Pengetahuan Dan Teknologi (KNIT), pp. 1–8, 2015.
- [5] R. N. Harahap, K. Muslim, and P. Korespondensi, "Peningkatan Akurasi Pada Prediksi Kepribadian MBTI Pengguna Twitter Menggunakan Augmentasi Data", doi: 10.25126/jtiik.202073622.
- [6] D. A. Putri, "Penerapan Algoritma Support Vector Machine Berbasis Algoritma Genetika Untuk Analisis Sentimen Pada Twitter," *Jurnal Teknik Informatika Stmik Antar Bangsa*, vol. 1, no. 01, 2015.
- [7] E. M. Maseno and Z. Wang, "Hybrid wrapper feature selection method based on genetic algorithm and extreme learning machine for intrusion detection," *J Big Data*, vol. 11, no. 1, Dec. 2024, doi: 10.1186/s40537-024-00887-9.
- [8] I. Widaningrum, D. Mustikasari, R. Arifin, S. L. Tsaqila, and D. Fatmawati, "Algoritma Term Frequency-Inverse Document Frequency (TF-IDF) dan K-Means Clustering Untuk Menentukan Kategori Dokumen," 2022.
- [9] N. I. Widiastuti, "Algoritma genetik pada masalah tata letak mesin dengan pengkodean kromosom untuk ukuran mesin yang berbeda-beda," **J. Computech & Bisnis**, vol. 5, no. 2, pp. 81–88, 2011.
- [10] Rizky Fatih Syahputra and Yahfizham Yahfizham, "Menganalisis Konsep Dasar Algoritma Genetika," *Bhinneka: Jurnal Bintang Pendidikan dan Bahasa*, vol. 2, no. 1, pp. 120–132, Dec. 2023, doi: 10.59024/bhinneka.v2i1.643.
- [11] J. Brownlee, Ph.D., "Genetic Algorithm," *Algorithm Afternoon*, 2024. [Online].
- [12] A. S. Nugroho, A. B. Witarto, and D. Handoko, "Support Vector Machine-Teori dan Aplikasinya dalam Bioinformatika 1," 2003. [Online]. Available: <http://asnugroho.net>
- [13] Y. Sari, M. Alkaff, E. S. Wijaya, S. Soraya, and D. P. Kartikasari, "Optimasi Penjadwalan Mata Kuliah Menggunakan Metode Algoritma Genetika Dengan Teknik Tournament Selection," vol. 6, no. 1, pp. 85–92, 2019, doi: 10.25126/jtiik.201961262.
- [14] K. E. Dewi, "Perbandingan Metode Newton Raphson dengan Algoritma Genetika untuk Prediksi Saham," *Jurnal Ilmiah Komputer dan Informatika (KOMPUTA)*, vol. 1, no. 2, pp. 9–15, Oct. 2012, ISSN: 2089-9033.
- [15] A. Yehezkiel Mauko, A. Faggidae, dan Yulianto Triwahyuadi Polly, P. Studi Ilmu Komputer, U. Nusa Cendana, and J. Adi Sucipto Kupang, "Analisis Elitisme Pada Algoritma Genetika Menggunakan Pengkodean Ordinal Representation Dalam Travelling Salesman Problem," *J-ICON*, vol. 10, no. 2, pp. 216–222, 2022, doi: 10.35508/jicon.v10i2.8473.
- [16] F. Novianti dan N. Ulinuha, "Seleksi Fitur Algoritma Genetika dalam Klasifikasi Data Rekam Medis PCOS Menggunakan SVM," *Paradigma - Jurnal Komputer dan Informatika*, vol. 26, no. 1, hlm. 67–74, Jan. 2024.