

ANALISIS SENTIMEN PADA MEDIA SOSIAL MENGGUNAKAN WORD2VEC DAN GATED RECURRENT UNIT (GRU) DENGAN OPTIMASI GENETIC ALGORITHM

Syafa Fahreza

NIM : 1301204241

PROGRAM STUDI SARJANA INFORMATIKA

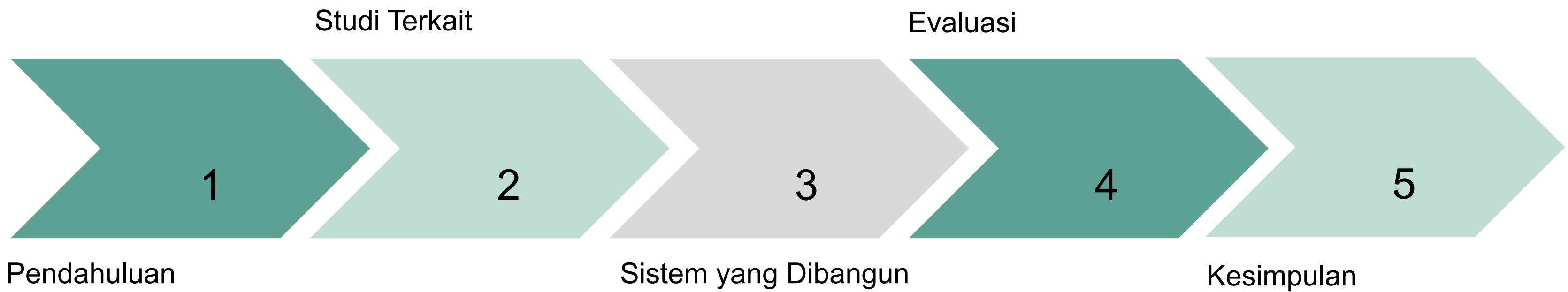
Fakultas Informatika

2024

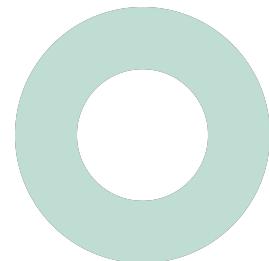


UNIVERSITAS
TELKOM
BANDUNG

Daftar Isi

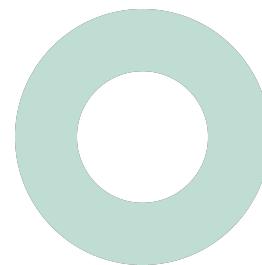


Latar Belakang



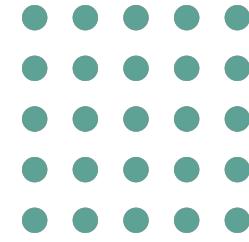
- Kemajuan teknologi informasi mengubah peran media social
- Sentimen analisis penting untuk dilakukan
- Salah satu Teknik deep learning adalah GRU
- Penelitian ini akan mengabungkan GRU dengan Word2Vec sebagai ekspansi fitur dan Genetic Algorithm sebagai optimasi.

Latar Belakang

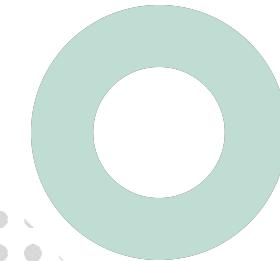


Kemajuan teknologi informasi telah mengubah peran media sosial dari sekedar tempat penyimpanan informasi menjadi sebuah platform untuk menyampaikan pendapat dan aspirasi. Salah satu platform media sosial yang banyak digunakan oleh masyarakat adalah Twitter. Dengan jumlah pengguna sebanyak 19,5 juta, Indonesia menempati urutan kelima di dunia untuk pengguna Twitter.

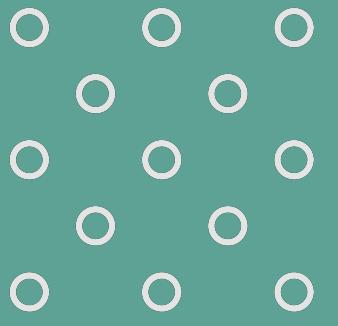
Analisis sentimen di media sosial diperlukan karena memungkinkan pemahaman yang lebih baik tentang opini dan preferensi pengguna, memantau persepsi publik, mengidentifikasi tren dan peluang, dan menyediakan layanan pelanggan.



Latar Belakang



01. **Gated Recurrent Unit (GRU)**
Beberapa teknik deep learning telah dibuat untuk analisis sentimen, dan salah satu contohnya adalah Gated Recurrent Unit (GRU). Salah satu jenis desain jaringan syaraf tiruan yang disebut GRU digunakan dalam Natural Language Processing (NLP) untuk menstimulasikan urutan data, seperti teks atau data bahasa manusia.
02. **Word2Vec**
Penelitian ini menggunakan Word2Vec sebagai metode ekspansi fitur dan Genetic Algorithm sebagai optimasi fitur. Word2Vec sendiri digunakan untuk mengubah data menjadi vektor dalam bentuk angka, sedangkan Genetic Algorithm digunakan untuk menyelesaikan masalah optimasi. Genetic Algorithm memiliki keunggulan dalam mengatasi tantangan komputasi yang luas secara efektif



Tujuan Penelitian

Kontribusi utama dari penelitian ini adalah mengoptimalkan model Gated Recurrent Unit (GRU) menggunakan Genetic Algorithm dan mengombinasikan ekspansi fitur dengan Word2Vec untuk sentiment analysis pada topik pemilihan presiden 2024 di Indonesia.



Studi Terkait

01.

Studi Satu

Peneliti lain membandingkan algoritma LSTM, GRU, Bi-GRU, dan Bi-LSTM dengan menggunakan dataset yang diambil dari ulasan amazon. Hasil yang didapatkan secara keseluruhan GRU lebih unggul.

02.

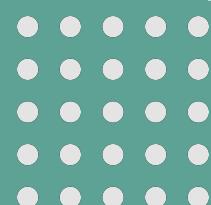
Studi Dua

Peneliti lain menggunakan teknik Gated Recurrent Unit (GRU) untuk memprediksi harga emas dengan menggunakan Mean Square Error (MSE) sebagai metrik untuk mengukur keakuratan prediksi. Hasilnya menunjukkan tingkat kesalahan MSE sebesar 0,111, nilai RMSE sebesar 0,334, dan R-squared sebesar 0,5

03.

Studi Tiga

Penelitian lain melakukan analisis sentimen pada dataset SS-Tweet dengan menggunakan enam metode klasifikasi yang menggunakan fitur TF-IDF dan N-Gram. Jika dibandingkan dengan N-Gram, temuan penelitian secara konsisten menunjukkan bahwa fitur TF-IDF berkinerja lebih baik (3-4%).



Studi Terkait

04.

Studi Empat

Kush dkk melakukan penelitian dengan menggunakan dataset ulasan film dengan bahasa Hindi untuk analisis sentimen. Penelitian tersebut membandingkan 8 algoritma model yang berbeda. Dari semua algoritma yang diuji, kombinasi GA-GRU mendapatkan akurasi terbaik sebesar 88,2%.

05.

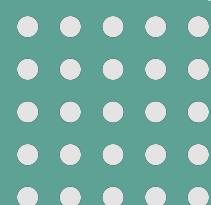
Studi Lima

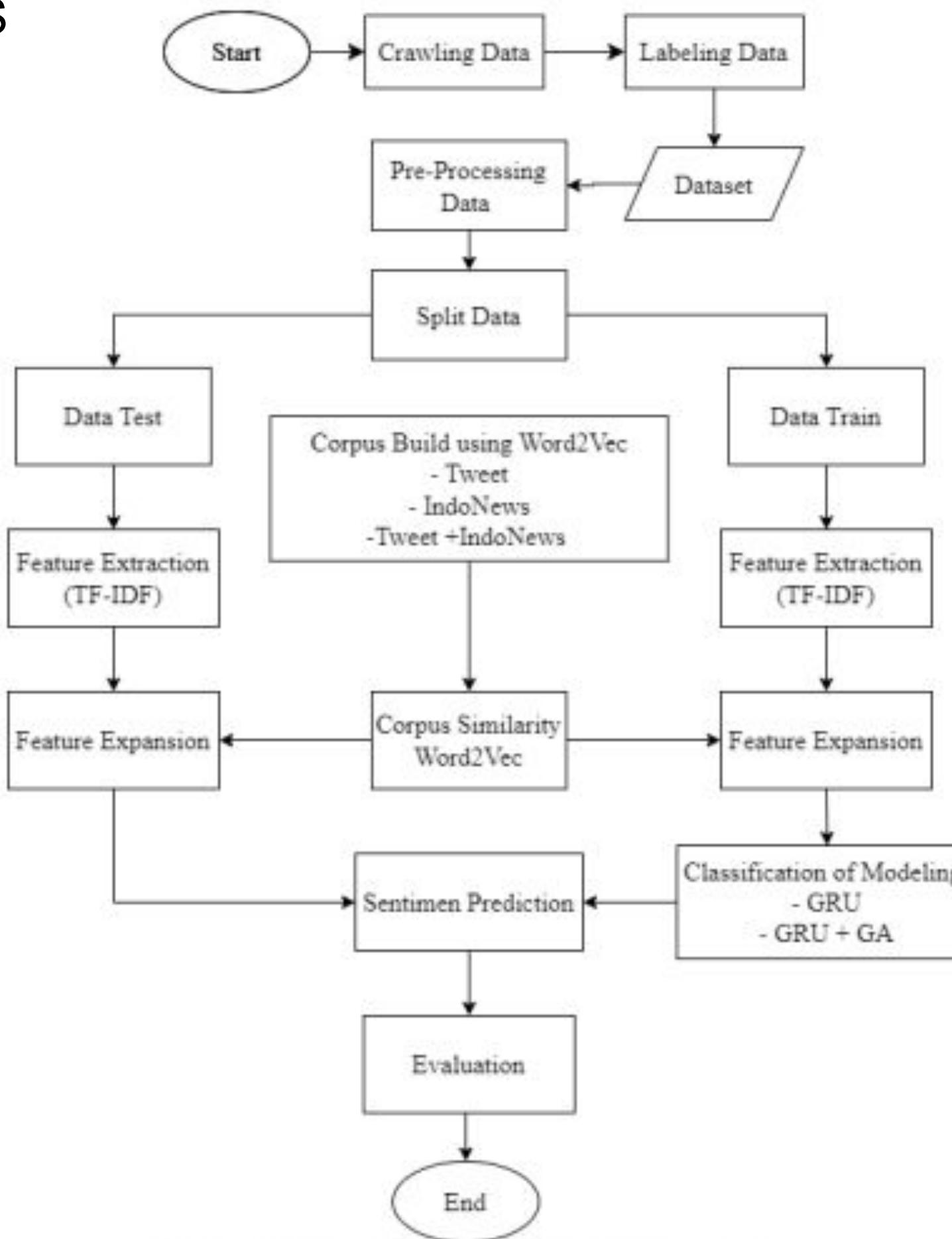
Peneliti lain untuk menilai dan membandingkan kinerja metode GRU dengan metode LSTM untuk melihat potensi peningkatan kinerja. Hasil akurasi yang dicapai oleh metode GRU pada dataset Facebook e-commerce cina mencapai 87%.

06.

Studi Enam

penelitian dari Farhan Wahyu Kurniawan dan Warih Maharan menunjukkan bahwa perbedaan dalam arsitektur model Word2Vec memiliki dampak pada hasil klasifikasi. Temuan penelitian menunjukkan bahwa penerapan model skip-gram dengan dimensi 100 memberikan hasil klasifikasi yang optimal.





Sistem Yang Dibangun

Sistem ini bertujuan untuk membangun model analisis sentimen yang dirancang untuk menganalisis opini pengguna.

Sistem Yang Dibangun

Crawling Data

Mengumpulkan informasi dari sumber tertentu dikenal sebagai "crawling". Hasil dari proses crawling secara otomatis disimpan dalam format Comma Separated Value (CSV) menggunakan bahasa pemrograman Python.

Keyword	Amount	Ratio (%)
Anies Baswedan	10,434	27.90
Ganjar Pranowo	8,027	21.47
Capres	7,296	19.51
Calon Presiden	6,972	18.65
Prabowo Subianto	4,662	12.47
Total	37,391	100

Pelabelan Data

Dalam proses pengembangan dataset untuk sistem klasifikasi. Label akan diberikan dalam 2 kelas yaitu positif dan negatif.

Label	Amount	Ratio (%)
Positive	21,866	58.48
Negative	15,525	41.52
Total	37,391	100

Sistem Yang Dibangun



Preprocessing Data

Preprocessing data adalah proses dimana data mentah dibersihkan sehingga data siap digunakan untuk analisis lebih lanjut. Langkah-langkah dalam preprocessing adalah sebagai berikut :

- Pembersihan Data
- Case Folding
- Tokenisasi
- Penghapusan Stopword

Ekstraksi Fitur

Ekstraksi fitur memainkan peran penting dalam pemrosesan dokumen di mesin pencari karena memiliki dampak yang signifikan terhadap keberhasilan proses text mining.

$$tf_{dt} = 0,5 + 0,5 \times \frac{tf}{\max(tf)} \quad (1)$$

$$idf_t = \log\left(\frac{D}{df_t}\right) \quad (2)$$

$$w_{dt} = tf_{dt} \times idf_{dt} \quad (3)$$

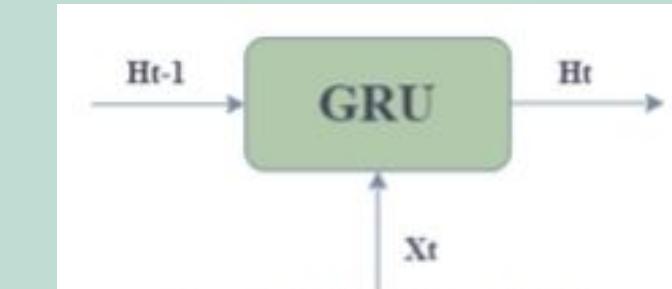
Sistem Yang Dibangun

Word2Vec

Word2Vec merupakan salah satu metode word embedding yang dikembangkan oleh Mikolov. Tujuan dari Word2Vec adalah merepresentasikan kata ke dalam sebuah vektor dengan panjang N, sehingga dapat digunakan untuk mengidentifikasi hubungan atau korelasi antar kata.

Gated Recurrent Unit (GRU)

GRU (Gated Recurrent Unit) merupakan salah satu jenis arsitektur jaringan syaraf tiruan (RNN) yang digunakan dalam bidang pembelajaran mesin dan pemrosesan bahasa alami. GRU dikembangkan sebagai jaringan yang lebih sederhana dan lebih efisien dibandingkan LSTM (Long Short-Term Memory)

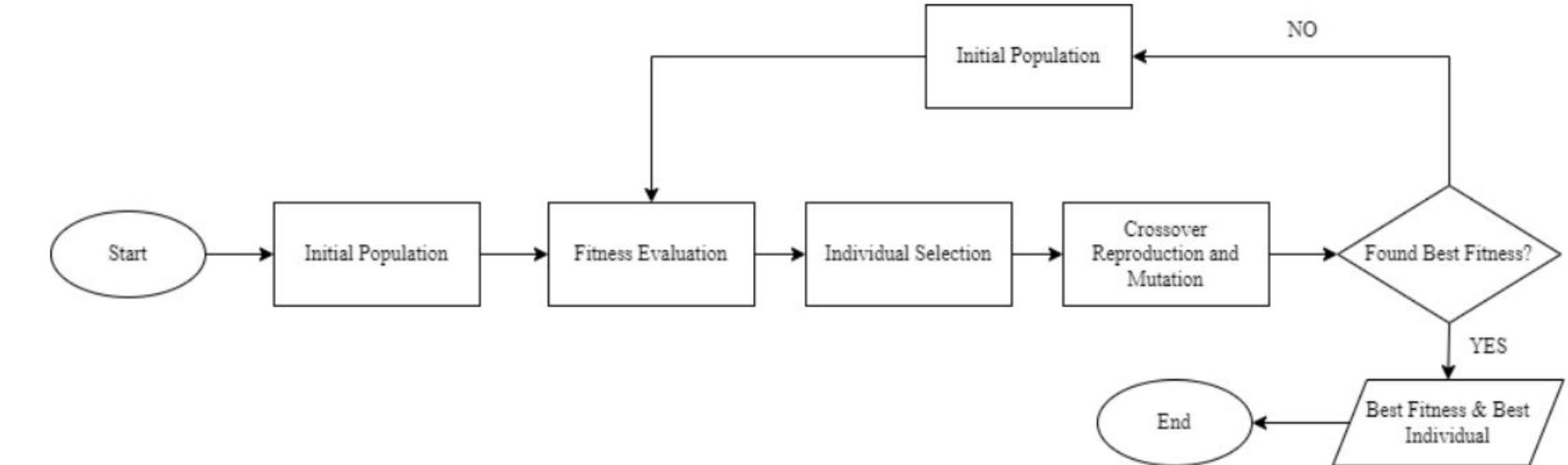


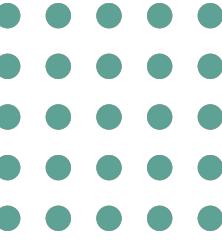


Sistem Yang Dibangun

Genetic Algorithm

Genetic Algorithm atau Algoritma Genetika merupakan salah satu metode yang digunakan untuk menyelesaikan masalah yang berkaitan dengan optimasi. Genetic Algorithm pertama kali dikembangkan oleh John Holland dengan mengadopsi teori dasar yang dijelaskan oleh Charles Darwin.





Sistem Yang Dibangun

Evaluasi

GRU (Gated Recurrent Unit) merupakan salah satu jenis arsitektur jaringan syaraf tiruan (RNN) yang digunakan dalam bidang pembelajaran mesin dan pemrosesan bahasa alami. GRU dikembangkan sebagai jaringan yang lebih sederhana dan lebih efisien dibandingkan LSTM (Long Short-Term Memory)

Confusion Matrix		Actual Value	
		Positive	Negative
Predicted Value	Positive	TP	FP
	Negative	FN	TN

Berikut adalah beberapa perhitungan yang digunakan dalam perhitungan confusion matrix :

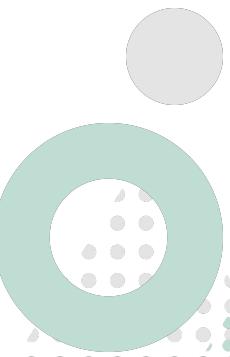
- Akurasi
- Precision
- Recall
- F1-Score

Evaluasi

Hasil Pengujian

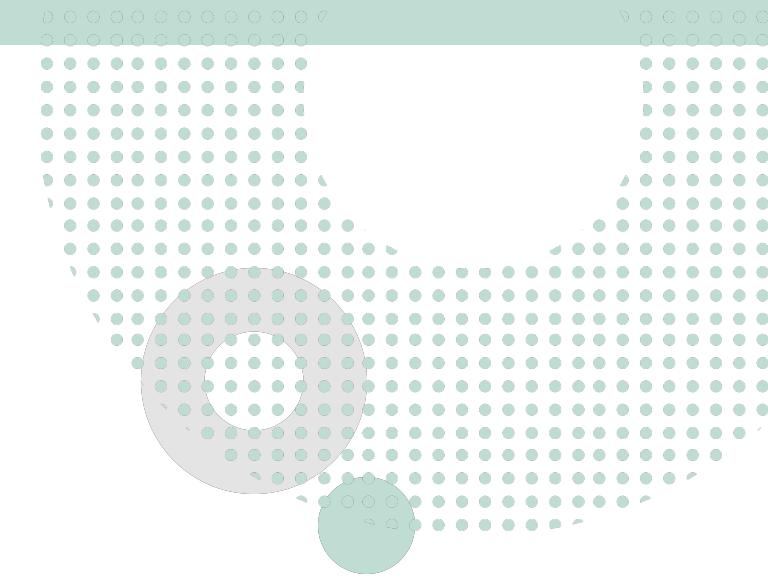
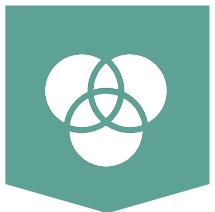


Penelitian ini memiliki empat skenario pengujian. Pada skenario pertama, rasio data latih dan data uji dibandingkan guna menentukan baseline untuk model GRU. Skenario kedua adalah menggunakan ekstraksi fitur TF-IDF pada baseline. Pada skenario ketiga, Word2Vec digunakan untuk ekspansi fitur. Untuk skenario keempat menggunakan Genetic Algorithm sebagai optimasi fitur. Rata-rata dari lima hasil pengujian akan dijadikan nilai akurasi dan nilai F1 pada setiap skenario.



Tujuannya adalah untuk mengetahui apakah ada peningkatan akurasi dan F1-score pada setiap tahap pengujian.





Skenario 1:

Pengujian Best Split Size

Pada skenario pertama, penulis membandingkan akurasi dan skor F1 dari tiga pembagian data: 70:30, 80:20, dan 90:10. Proses ini terdiri dari melatih model pada data training dan mengukur performa pada data testing.

Tabel 4. Hasil Skenario 1

Split Size	Accuracy (%)	F1-Score (%)
90:10	81.97	81.47
80:20	81.50	81.00
70:30	80.66	79.94

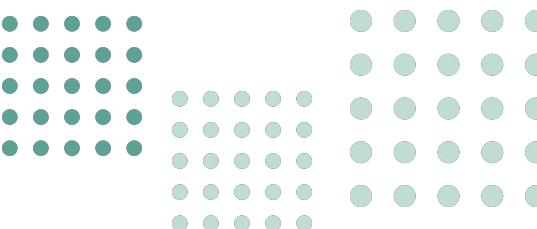
Skenario 2:

Pengujian Max Feature

Tujuan dari skenario ini adalah untuk menentukan hubungan antara fitur maksimum dan kinerja model berdasarkan akurasi dan F1-score. Fitur maksimum yang akan dibandingkan adalah 1.000, 2.000, 5.000, 8.000, dan

Tabel 5. Hasil Skenario 2

Max Feature	Accuracy (%)	F1-Score (%)
1000	81.71 (-0.26)	81.31 (-0.16)
2000	82.66 (+0.69)	82.36 (+0.89)
5000	82.96 (+0.99)	82.57 (+1,1)
8000	82.95 (+0.98)	82.56 (+1.09)
12000	82.90 (+0.93)	82.41 (0.94)





Skenario 3:

Pengujian Top Similarity

Skenario pengujian 3 dilakukan untuk meningkatkan hasil terbaik dari skenario 2 dengan menerapkan ekspansi fitur Word2Vec.

Tabel 6. Hasil Skenario 3

Top-N	Twitter		IndoNews		Twitter+IndoNews	
	Accuracy (%)	F1-Score (%)	Accuracy (%)	F1-Score (%)	Accuracy (%)	F1-Score (%)
1	82.13 (+0.16)	81.68 (+0.21)	83.00 (+1.03)	82.70 (+1.23)	82.76 (+0.79)	82.38 (+0.91)
5	81.50 (-0.47)	81.03 (-0.44)	82.70 (+0.73)	82.26 (+0.79)	82.47 (+0.5)	82.06 (+0.59)
10	79.95 (-2.02)	79.57 (-1.9)	82.14 (+0.17)	81.67 (+0.2)	81.94 (-0.03)	81.55 (+0.08)



Skenario 4:

Penerapan Genetic Algorithm

Pada skenario keempat ini adalah penggunaan Genetic Algorithm sebagai optimasi fitur. Dalam setiap generasi Genetic Algorithm (GA), berbagai tahapan seperti crossover, mutasi, dan evaluasi fitness dilakukan.

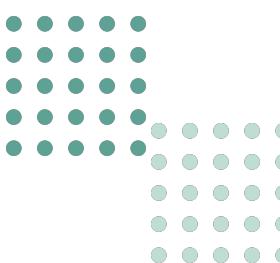
Kompleksitas waktu dari GA dipengaruhi oleh karakteristik fungsi fitness yang digunakan.

Tabel 7. Parameter Evolusi GA

Evolutionary Parameter	Value
Cpb	0.5
Mutpb	0.2
Ngen	5

Tabel 8. Parameter Genetik GA

Genetic Parameter	Value
Mate	Indpb = 0.5
Mutate	Mu = 0, Sigma = 1, indpb = 0.2
Select	Tournsize = 3



Analisis Hasil Pengujian

Hasil pengujian dalam penelitian ini memberikan pemahaman yang mendalam mengenai efektivitas model GRU dalam melakukan analisis sentimen pada data Twitter dengan topik pemilihan presiden di Indonesia tahun 2024.



Gambar 4 Hasil Pengujian Semua Skenario



Kesimpulan

Pada penelitian ini, analisis sentimen dilakukan dengan menggunakan dataset dari media sosial Twitter dengan fokus pada topik pemilihan presiden di Indonesia.

Dataset yang digunakan terdiri dari 37.391 yang akan diberi label sebagai negatif dan positif, yang dilakukan melalui proses pelabelan secara manual. Hasil pengujian menunjukkan bahwa akurasi tertinggi yang dicapai adalah 83,39%, yang menunjukkan peningkatan sebesar 1,42% dibandingkan dengan baseline. Performa ini dicapai dengan mengombinasikan TF-IDF dengan 5.000 fitur maksimum, menerapkan Word2Vec dengan 142.545 korpus dari IndoNews pada top 1 similarity, dan menerapkan Algoritma Genetika untuk optimasi fitur.



Terima Kasih

