

**Protected by PDF Anti-Copy Free**  
**KOMPARASI ALGORITMA NAIVE BAYES DAN**  
*(Upgrade to Pro Version to Remove the Watermark)*  
**SUPPORT VECTOR MACHINE (SVM) DALAM MENGENALISIS**  
**ELEKTABILITAS MOCHAMMAD RIDWAN KAMIL PADA PILKADA**  
**DKI 2024 MELALUI MEDIA SOSIAL YOUTUBE**



**SKRIPSI**

**Diajukan Sebagai Syarat Untuk  
Menyelesaikan Program Strata satu(S1)  
Pada Program Studi Informatika**

**Disusun Oleh:**

**DANIEL ADLESS ERICH**

**NIM: 2102020154**

**PROGRAM STUDI INFORMATIKA  
FAKULTAS ILMU TEKNIK  
UNIVERSITAS BINA INSAN  
2025**

**KOMPARASI ALGORITMA NAIVE BAYES DAN  
SUPPORT VECTOR MACHINE (SVM) DALAM MENGANALISIS  
ELEKTABILITAS MOCHAMMAD RIDWAN KAMIL PADA PILKADA  
DKI 2024 MELALUI DATA MEDIA SOSIAL YOUTUBE**

Oleh:

**DANIEL ADLESS ERICH**

**NIM: 2102020154**

**Lubuklinggau, Januari 2025**

**Pembimbing I**

**Pembimbing II**

**Lukman Sunardi, M. Kom**

**Nelly Khairani Daulay, M. Kom**

**Mengetahui,**

**Dekan Fakultas Ilmu Teknik**

**Universitas Bina Insan**

**Dr. Rudi Kurniawan, S.T., M. Kom**

**HALAMAN PERSETUJUAN KOMISI PENGUJI SKRIPSI**  
**Protected by PDF Anti-Copy Free**

**(Upgrade to Pro Version to Remove the Watermark)**

Pada hari Kamis tanggal 23 bulan Januari Tahun 2025 telah dilaksanakan Ujian Skripsi oleh program studi Informatika Universitas Bina Insan Lubuklinggau



Nama : Daniel Adless Erich

NIM : 2102020154

Jurusan : Informatika

Judul : *Komparasi Algoritma Naïve Bayes dan Support Vector Machine (SVM) Dalam Menganalisis Elektabilitas Mochammad Ridwan Kamil Pada Pilkada DKI Jakarta 2024 Melalui data Media Sosial YouTube*

**Komisi Penguji**

1. Ketua : Lukman Sunardi, M. Kom (. ..)

2. Sekretaris : Nelly Khairani Daulay, M. Kom ( ..)

3. Anggota : Dr. Muhamad Akbar, S.T., M.IT (... ..)

Mengesahkan,  
Ketua Program Studi Informatika  
Universitas Bina Insan  
Lubuklinggau

**Budi Santoso, M. Kom**

## HALAMAN MOTTO DAN PERSEMBAHAN

Protected by PDF Anti-Copy Free

MOTTO : (Upgrade to Pro Version to Remove the Watermark)

*"In The Name of Lord Jesus Christ, the Savior and King"*



"Tuhan tidak lalai menepati janji-Nya, sekalipun ada orang yang menganggapnya sebagai kelalaian, tetapi Ia sabar terhadap kamu, karena Ia menghendaki supaya jangan ada yang binasa, melainkan supaya semua orang berbalik dan bertobat."

( 2 Petrus 3:9)

- ❖ Perlakukanlah orang lain seperti kamu ingin diperlakukan, sebagaimana Tuhan mengajarkan kita untuk mengasihi sesama
- ❖ Kegagalan bukanlah akhir, andalkan Tuhan dan lakukan yang terbaik, sebab Dia memberikan kekuatan baru setiap hari
- ❖ Percayalah kepada rencana Tuhan, karena segala sesuatu indah pada waktunya

### *Persembahan kepada :*

- ❖ *"Untuk kedua orang tuaku Suhunan Sinaga dan Zozor Hutajulu yang tiada henti memberikan cinta, dukungan, dan doa yang selalu menguatkan aku. Terima kasih atas segala motivasi, semangat, dan pengorbanan yang tak terhingga untukku."*
- ❖ *"Terima kasih kepada teman-teman seperjuangan, sahabat, dan teman-teman Angkatan 2021 atas motivasi, dukungan, dan kebersamaannya yang menjadi sumber kekuatan dalam perjalanan ini."*
- ❖ *"Dan untuk orang tersayang Maraynelle Alicia yang selalu memberikan cinta tanpa syarat, terima kasih atas dukungan dan kehadiranmu yang selalu membuatku merasa kuat."*

**Protected by PDF Anti-Copy Free**  
(Upgrade to Pro Version to Remove the Watermark)

**HALAMAN PERNYATAAN**

Saya yang bertanda tangan dibawah



Nama Mahasiswa : Daniel Adless Erich

Nim : 2102020154

Program Studi : Informatika

Menyatakan dengan sesungguhnya bahwa penelitian dan penulisan skripsi yang saya susun sebagai persyaratan untuk memperoleh gelar sarjana (S-1) Universitas Bina Insan merupakan hasil kerja saya sendiri dan tidak menyuruh orang lain yang mengerjakannya. Ada bagian tertentu dalam penulisan skripsi ini yang saya kutip dari hasil karya orang lain dan telah saya tuliskan sumbernya secara jelas sesuai dengan norma, kaidah dan etika penulisan ilmiah.

Jika dikemudian hari ternyata terbukti bahwa penelitian dan tugas akhir ini bukan hasil kerja saya sendiri atau plagiat dalam bagian-bagian tertentu, maka saya bersedia dikenakan sanksi sesuai dengan peraturan perundangan yang berlaku.

**Lubuklinggau, Januari 2025**

**Penulis,**

**DANIEL ADLESS ERICH**

**21.02.02.0154**

**Protected by PDF Anti-Copy Free**  
**(Upgrade to Pro Version to Remove the Watermark)**



*This research aims to analyze public sentiment against potential political leaders, especially Mochammad Ridwan Kamil, in the 2024 DKI Regional Election using YouTube social media data. Data was collected via web scraping from August to September 2024, with a total of 10,843 entries. The preprocessing process includes data cleaning, casefolding, tokenization, normalization, stopword removal, and stemming. Sentiments are analyzed using the Naive Bayes and Support Vector Machine (SVM) algorithms. The results of the analysis showed that 50.35% of public opinion was negative, 25.56% neutral, and 24.10% positive. The Naive Bayes algorithm with Gaussian models produced the highest accuracy of 94.49%, higher than that of SVMs with 83%. This finding shows that the dominant negative sentiment can affect Ridwan Kamil's electability in the 2024 DKI Regional Head Election. This study highlights the importance of managing public perception on social media as part of political strategy.*

**Keywords:** *Sentiment Analysis, Naive Bayes, Vector Machine Support, Social Media, 2024 DKI Regional Election.*

# Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

## ABSTRAK



Penelitian ini bertujuan untuk menganalisis sentimen publik terhadap calon pemimpin politik, khususnya Mochammad Ridwan Kamil, dalam Pilkada DKI 2024 menggunakan data media sosial YouTube. Data dikumpulkan melalui *web scraping* pada periode Agustus hingga September 2024, dengan total 10.843 entri. Proses *preprocessing* mencakup data *cleaning*, *casefolding*, tokenisasi, normalisasi, *stopword removal*, dan *stemming*. Sentimen dianalisis menggunakan algoritma *Naive Bayes* dan *Support Vector Machine (SVM)*. Hasil analisis menunjukkan bahwa 50,35% opini masyarakat bersifat negatif, 25,56% netral, dan 24,10% positif. Algoritma *Naive Bayes* dengan model *Gaussian* menghasilkan akurasi tertinggi sebesar 94,49%, lebih tinggi dibandingkan dengan *SVM* yang memperoleh 83%. Temuan ini menunjukkan bahwa sentimen negatif yang dominan dapat mempengaruhi elektabilitas Ridwan Kamil dalam Pilkada DKI 2024. Penelitian ini menyoroti pentingnya pengelolaan persepsi publik di media sosial sebagai bagian dari strategi politik.

**Kata kunci:** *Analisis Sentimen, Naive Bayes, Support Vector Machine, Media Sosial, Pilkada DKI 2024.*

# Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

## KATA PENGANTAR

Segala kemuliaan hanya bagi Tuhan, penulis mengucapkan terima kasih atas segala berkat dan penyertaan-Nya yang melimpah, sehingga penulis dapat menyelesaikan Skripsi ini dengan baik dan tepat waktu. Proposal ini merupakan salah satu syarat untuk menyelesaikan Pendidikan Strata Satu (S1) di Program Studi Informatika Fakultas Ilmu Teknik Universitas Bina Insan Lubuklinggau. Semoga melalui karya ini, segala berkat dan kasih dari Tuhan dapat senantiasa mengalir dalam kehidupan penulis dan orang-orang yang berpartisipasi dalam proses ini.

Dalam proses penyusunan Skripsi ini, penulis telah berupaya dengan sungguh-sungguh untuk menghasilkan karya terbaik, baik dari segi isi maupun desain program. Namun, penulis menyadari keterbatasan kemampuan yang ada, sehingga masih banyak kekurangan. Oleh karena itu, penulis sangat terbuka terhadap saran dan kritik yang membangun demi kesempurnaan karya ini di masa mendatang.

Penulis juga ingin mengucapkan terima kasih yang sebesar-besarnya kepada pihak-pihak yang telah dengan tulus hati memberikan bantuan, dukungan, dan arahan dalam menyelesaikan Skripsi ini:

1. Kepada Tuhan Yang Maha Esa, yang selalu memberikan kekuatan, hikmat, dan cinta-Nya, dalam setiap langkah penulis.
2. Kepada keluarga tercinta yang selalu menjadi sumber semangat, doa, dan dukungan tanpa henti dalam setiap perjuangan.
3. Bapak Dr. H. Sardiyo, MM selaku Rektor Universitas Bina Insan Lubuklinggau.
4. Bapak Dr. Muhamad Akbar, S.T., M.IT selaku Wakil Rektor I Universitas Bina Insan Lubuklinggau
5. Bapak Mukhlis Nur Wahid, M.Pd Selaku Wakil Rektor II Universitas Bina Insan Lubuklinggau
6. Bapak Dr. Rudi Kurniawan, S.T., M. Kom Selaku Dekan Fakultas Ilmu Teknik Universitas Bina Insan Lubuklinggau

7. Bapak Budi Santoso, M.Kom selaku Ketua Prodi Informatika Universitas Bina Insan Lubuklinggau
8. Bapak Lukman Sunardi, M. Kom selaku Pembimbing I yang telah membimbing saya dalam penyusunan skripsi dan pengembangan program.
9. Ibu Nelly Khairani Daula, M.Kom selaku Pembimbing II yang telah banyak memberikan ilmu, waktu, dan bimbingan selama masa perkuliahan dan penulisan skripsi ini.
10. Media YouTube yang telah menjadi sarana bagi saya untuk melaksanakan penelitian selama waktu yang telah ditentukan.
11. Seluruh Staf Dosen dan Karyawan Universitas Bina Insan yang telah banyak memberikan ilmu pengetahuan dan bimbingan kepada penulis.
12. Kepada sahabat-sahabat terkasih yang selalu ada, menghibur, dan memberikan motivasi di setiap perjalanan ini.

Akhir kata, semoga penelitian ini dapat memberikan manfaat yang berarti bagi analisis elektabilitas Mochammad Ridwan Kamil pada Pilkada DKI 2024 melalui data media sosial YouTube. Penelitian ini diharapkan dapat menjadi referensi yang berguna bagi penelitian selanjutnya, serta memberikan wawasan tentang perbandingan algoritma *Naive Bayes* dan *Support Vector Machine (SVM)* dalam konteks analisis data. Semoga hasil dari penelitian ini tidak hanya mencerminkan usaha dan kerja keras, tetapi juga memberikan kontribusi positif dalam memahami dinamika elektabilitas di era digital ini.

Lubuklinggau, Januari 2025

Daniel Adless Erich



**Biodata**

Nama : Daniel Adless Erich  
Tempat / Tanggal Lahir : Medan / 27 Juni 2003  
Jenis Kelamin : Laki-Laki  
Agama : Kristen  
Alamat : Jl. Perumahan Griya Asri blok Q. 8, Rt.10, Kel.Muara  
Enim, Kec. Lubuklinggau Barat 1

**Pendidikan**

- SD : SD Negeri 28 Lubuklinggau
- SMP : SMP Negeri 5 Lubuklinggau
- SMA : SMA Negeri 8 Lubuklinggau

**DAFTAR ISI**  
**Protected by PDF Anti-Copy Free**

(Upgrade to Pro Version to Remove the Watermark)

HALAMAN JUDUL.....	i
HALAMAN PENGESAHAN.....	ii
HALAMAN PERSETUJUAN.....	iii
HALAMAN MOTTO.....	iv
HALAMAN PERNYATAAN.....	v
HALAMAN <i>ABSTRACT</i> .....	vi
HALAMAN ABSTRAK.....	vii
KATA PENGANTAR.....	viii
DAFTAR RIWAYAT HIDUP.....	x
DAFTAR ISI.....	xi
DAFTAR TABEL.....	xiii
DAFTAR GAMBAR.....	xiv
DAFTAR LAMPIRAN.....	xv
<b>BAB I PENDAHULUAN.....</b>	<b>1</b>
1.1 Latar Belakang Penelitian .....	1
1.2 Identifikasi Masalah.....	5
1.3 Rumusan Masalah .....	5
1.4 Batasan Masalah .....	6
1.5 Tujuan dan Manfaat Penelitian .....	7
1.5.1 Tujuan Penelitian .....	7
1.5.2 Manfaat Penelitian .....	7
1.6 Sistem Penulisan.....	8
<b>BAB II KAJIAN PUSTAKA .....</b>	<b>10</b>
2.1 Literatur.....	10
2.2 Penelitian Relevan .....	26
2.3 Kerangka Berfikir .....	28
<b>BAB III METODOLOGI PENELITIAN .....</b>	<b>32</b>
3.1 Analisa Sistem.....	32
3.1.1 Analisa Sistem yang Berjalan .....	32
3.1.2 Alternatif Pemecahan Masalah .....	32
3.1.3 Metode Analisa .....	33

3.2 Teknik Pemilihan Informan (Populasi, sampel dan sampling) .....	35
3.2.1 Teknik Pengumpulan data .....	35
3.2.2 Teknik analisa data .....	36
3.3 Tempat dan waktu penelitian .....	43
<b>BAB IV. HASIL PENELITIAN DAN PEMBAHASAN</b> .....	<b>45</b>
4.1 Gambaran Umum (Temuan penelitian) .....	45
4.2 Hasil .....	46
4.3 Pembahasan .....	63
<b>BAB V KESIMPULAN DAN SARAN</b> .....	<b>65</b>
5.1 Kesimpulan .....	65
5.2 Saran .....	66
<b>DAFTAR PUSTAKA</b> .....	<b>67</b>
<b>LAMPIRAN</b> .....	<b>73</b>

# Protected by PDF Anti-Copy Free

## DAFTAR TABEL

(Upgrade to Pro Version to Remove the Watermark)

<b>Tabel 2.1</b> Representasi <i>Confusion Matrix</i> .....	20
<b>Tabel 3.1</b> <i>Confusion Matrix</i> .....	36
<b>Tabel 3.2</b> Contoh Hasil Proses <i>Cleaning</i> .....	38
<b>Tabel 3.3</b> Contoh Hasil <i>Case Folding</i> .....	39
<b>Tabel 3.4</b> Contoh Hasil <i>Tokenize</i> .....	39
<b>Tabel 3.5</b> Contoh Hasil Normalisasi.....	40
<b>Tabel 3.6</b> Contoh Hasil <i>Stopword Removal</i> .....	41
<b>Tabel 3.7</b> Contoh Hasil <i>Stemming</i> .....	41
<b>Tabel 3.8</b> Pemberian <i>Labeling</i> pada Data.....	42
<b>Tabel 3.9</b> Jadwal Waktu Penelitian.....	43

# Protected by PDF Anti-Copy Free

## DAFTAR GAMBAR (Upgrade to Pro Version to Remove the Watermark)

<b>Gambar 1.1</b> Grafik jumlah artikel klasifikasi dan prediksi .....	3
<b>Gambar 2.1</b> SVM berusaha mencari <i>hyperplane</i> terbaik .....	19
<b>Gambar 2.2</b> Kerangka Berfikir .....	28
<b>Gambar 3.1</b> <i>Preprocessing</i> data .....	42
<b>Gambar 4.1</b> Hasil <i>Scraping</i> data .....	46
<b>Gambar 4.2</b> Hasil <i>cleaning</i> data .....	47
<b>Gambar 4.3</b> Hasil <i>Case folding</i> .....	47
<b>Gambar 4.4</b> Hasil <i>Tokenization</i> .....	48
<b>Gambar 4.5</b> Hasil Normalisasi .....	48
<b>Gambar 4.6</b> Hasil <i>Stopword removal</i> .....	49
<b>Gambar 4.7</b> Hasil <i>Stemming</i> .....	49
<b>Gambar 4.8</b> Hasil <i>Labelling</i> .....	50
<b>Gambar 4.9</b> <i>Wordcloud</i> sentiment negatif .....	51
<b>Gambar 4.10</b> <i>Wordcloud</i> sentiment positif .....	51
<b>Gambar 4.11</b> Visualisasi <i>barchart</i> .....	52
<b>Gambar 4.12</b> Visualisasi <i>pie chart</i> .....	52
<b>Gambar 4.13</b> Data training 90% dan testing 10% .....	54
<b>Gambar 4.14</b> Data training 80% dan testing 20% .....	54
<b>Gambar 4.15</b> Data training 70% dan testing 30% .....	55
<b>Gambar 4.16</b> Data training 75% dan testing 25% .....	56
<b>Gambar 4.17</b> Data training 60% dan testing 40% .....	56
<b>Gambar 4.18</b> Hasil klasifikasi <i>Naïve Bayes</i> dengan perbandingan 90:10 .....	57
<b>Gambar 4.19</b> Hasil klasifikasi <i>Naïve Bayes</i> dengan perbandingan 80:20 .....	58
<b>Gambar 4.20</b> Hasil klasifikasi <i>Naïve Bayes</i> dengan perbandingan 70:30 .....	58
<b>Gambar 4.21</b> Hasil klasifikasi <i>Naïve Bayes</i> dengan perbandingan 75:25 .....	59
<b>Gambar 4.22</b> Hasil klasifikasi <i>Naïve Bayes</i> dengan perbandingan 60:40 .....	59
<b>Gambar 4.23</b> Hasil klasifikasi metode <i>SVM</i> untuk perbandingan 90:10 .....	60
<b>Gambar 4.24</b> Hasil klasifikasi metode <i>SVM</i> untuk perbandingan 80:20 .....	61
<b>Gambar 4.25</b> Hasil klasifikasi metode <i>SVM</i> untuk perbandingan 70:30 .....	61
<b>Gambar 4.26</b> Hasil klasifikasi metode <i>SVM</i> untuk perbandingan 75:25 .....	61
<b>Gambar 4.27</b> Hasil klasifikasi metode <i>SVM</i> untuk perbandingan 60:40 .....	62

# Protected by PDF Anti-Copy Free

DAFTAR LAMPIRAN  
(Upgrade to Pro Version to Remove the Watermark)

<b>Lampiran 1.</b> Acc Judul Proposal Skripsi .....	74
<b>Lampiran 2.</b> Bimbingan Proposal P1 .....	75
<b>Lampiran 3.</b> Bimbingan Proposal P2 .....	76
<b>Lampiran 4.</b> Channel YouTube Deddy .....	78
<b>Lampiran 5.</b> Channel YouTube Najwa Shihab .....	79
<b>Lampiran 6.</b> Perbaikan Seminar Proposal Skripsi.....	80
<b>Lampiran 7.</b> Bimbingan Skripsi P1 .....	81
<b>Lampiran 8.</b> Bimbingan Skripsi P2 .....	82
<b>Lampiran 9.</b> Perbaikan Skripsi .....	84
<b>Lampiran 10.</b> Jurnal .....	84
<b>Lampiran 11.</b> Bebas Plagiasi .....	91

# Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

## BAB I



## DAFTAR ISI

### 1.1 Latar Belakang Penelitian

Tahun 2024 menjadi tahun politik yang signifikan bagi masyarakat Indonesia, di mana berbagai pemilihan umum, termasuk pemilihan kepala daerah (Pilkada), akan dilangsungkan secara serentak. Pemilihan ini tidak hanya memberikan kesempatan kepada masyarakat untuk menentukan pemimpin yang layak, tetapi juga menjadi panggung bagi para kandidat untuk menunjukkan visi dan misi mereka. Dalam konteks ini, masyarakat semakin kritis dalam menentukan pilihan, terutama dengan memanfaatkan berbagai sumber digital untuk menggali informasi lebih mendalam mengenai karakteristik dan rekam jejak para kandidat.

Elektabilitas menjadi indikator penting dalam mengukur tingkat dukungan masyarakat terhadap seorang kandidat, serta peluangnya untuk terpilih berdasarkan popularitas dan penerimaan masyarakat luas. Dalam dunia politik, elektabilitas juga merupakan perbincangan hangat yang menjadi fokus sentral masyarakat secara insidental, memberikan pengaruh kepada individu atau pihak tertentu. Dengan kemudahan akses terhadap informasi terkini dari berbagai sumber digital, masyarakat semakin sering memberikan tanggapan yang diartikan sebagai feedback untuk tokoh-tokoh atau pihak tertentu[1]. Khususnya dalam konteks pemilihan gubernur di DKI Jakarta, elektabilitas memiliki peran yang sangat strategis. Jakarta, sebagai pusat politik dan ekonomi Indonesia, menjadi wilayah yang selalu menarik perhatian publik saat pemilihan kepala daerah berlangsung. Figur yang terpilih sebagai gubernur sering kali dilihat sebagai tokoh yang berpotensi menjadi pemimpin di tingkat nasional. Oleh karena itu, pengukuran elektabilitas kandidat menjadi sangat krusial dalam menentukan strategi kampanye politik yang tepat.

Di era digital saat ini, media sosial telah menjadi alat utama yang digunakan para kandidat politik untuk menyebarkan pesan-pesan

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

kampanye mereka. Kemajuan teknologi dalam kehidupan modern menuntut masyarakat selalu bergerak cepat dan instan dalam memenuhi segala aspek kehidupan sehari-hari. Era revolusi industri 4.0 yang memanfaatkan teknologi internet telah berdampak pada transformasi digitalisasi di hampir segala aspek kehidupan, termasuk dalam penyampaian pesan dari individu ke individu maupun dari individu ke khalayak luas. Muncul berbagai macam platform media sosial yang diciptakan untuk mewadahi kebutuhan tersebut, salah satunya adalah YouTube, yang berfungsi sebagai media berbagi video untuk penyampaian informasi maupun hiburan[2]. Berdasarkan laporan dari *We Are Social dan Hootsuite* pada bulan Januari 2021, aplikasi media sosial yang populer digunakan di Indonesia adalah YouTube yaitu 93,8 persen dari total keseluruhan pengguna internet. Kemudian untuk Instagram yaitu 86,6 persen, Facebook yaitu 85,5 persen dan TikTok 38,7 persen[3]. Sebagai salah satu platform terbesar, YouTube memberikan ruang bagi para kandidat untuk berkomunikasi secara visual dengan audiens yang luas. Dengan lebih dari 2,49 miliar pengguna aktif bulanan per Maret 2024 dan 47% dari pengguna internet di seluruh dunia mengakses YouTube setiap bulan, platform ini dipilih karena merupakan saluran komunikasi yang paling berpengaruh dalam era digital saat ini. Komentar atau tanggapan dari pengguna YouTube dapat memberikan gambaran nyata mengenai persepsi masyarakat terhadap seorang kandidat.

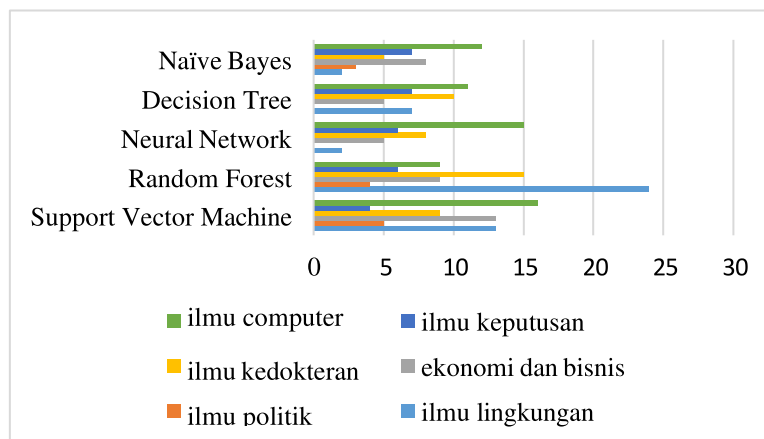
Dalam konteks ini, menilai elektabilitas Mochammad Ridwan Kamil melalui data dari komentar di YouTube menjadi sangat relevan. Sebagai pemimpin inovatif yang sebelumnya menjabat sebagai Wali Kota Bandung dan Gubernur Jawa Barat, ia kini muncul sebagai calon Gubernur dalam Pilkada DKI Jakarta 2024. Masyarakat Jakarta, yang dikenal dengan keberagaman dan dinamika sosialnya, semakin aktif dalam menyuarakan pendapat mereka melalui platform ini. Oleh karena itu, penting untuk memahami bagaimana mereka menanggapi

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

pencalonannya, agar strategi kampanye yang lebih tepat sasaran dapat dirancang.

Keberadaan YouTube sebagai salah satu situs web terpopuler di dunia menunjukkan bahwa platform ini tidak hanya berfungsi sebagai media hiburan, tetapi juga sebagai arena diskusi politik yang vital. Analisis terhadap komentar yang diberikan di YouTube akan membantu mengidentifikasi sentimen positif, negatif, dan netral yang mungkin mempengaruhi elektabilitas Ridwan Kamil. Dengan demikian, hasil analisis ini dapat memberikan wawasan yang berharga. Pada Gambar 1 bisa dilihat bahwa metode *Naïve Bayes*, *Decision Tree*, *Neural Network*, dan *Support Vector Machine* lebih sering dimanfaatkan pada bidang ilmu komputer dengan *Naïve Bayes* 12 artikel, *Decision Tree* 11 artikel, *Neural Network* 15 artikel, *Support Vector Machine* 16 artikel, sedangkan metode *Random Forest* pada lima tahun terakhir sering dimanfaatkan pada bidang ilmu lingkungan dengan jumlah 24 artikel.



**Gambar 1.1 Grafik jumlah artikel klasifikasi dan prediksi**

Sumber:[4]

Setelah dilakukan studi *literature* diketahui *Naïve Bayes* memiliki kelebihan yaitu: bisa digunakan dalam berbagai jenis data (kuantitatif, kualitatif) dan Data training yang perlukan hanya sedikit, dan bisa dipakai untuk mengklasifikasi jenis data biner maupun multiclass sementara *Support Vector Machine*, memiliki kelebihan tetap bisa mendapatkan

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

hasil yang maksimal meskipun hanya memiliki sedikit data training, karna untuk melatih metode ini hanya memerlukan data yang relative sedikit.

Akan tetapi, memproses jumlah komentar yang besar di YouTube, melakukan analisis secara manual akan memakan banyak waktu dan tenaga. Di sinilah teknologi machine learning, seperti algoritma *Naive Bayes* dan *Support Vector Machine (SVM)*, berperan. Algoritma ini memungkinkan klasifikasi sentimen masyarakat, baik positif maupun negatif, dilakukan secara otomatis dan efisien. *Naive Bayes* bekerja berdasarkan probabilitas dan terkenal dengan kesederhanaannya dalam menganalisis teks, sementara *Support Vector Machine (SVM)* mencari garis pemisah terbaik (hyperplane) yang memisahkan data sentimen menjadi kategori yang berbeda, seperti positif, negatif dan netral. Dengan kedua algoritma ini, analisis terhadap ribuan komentar di YouTube bisa diselesaikan secara lebih efektif. Proses ini dikenal sebagai *Sentiment Analysis* atau *Opinion Mining*, yang merupakan metode untuk memahami, mengekstrak, dan mengolah data tekstual secara otomatis guna mendapatkan informasi sentimen yang terkandung dalam opini. Analisis sentimen dilakukan untuk melihat pendapat atau kecenderungan opini terhadap sebuah masalah atau objek, apakah pandangan yang diberikan lebih cenderung positif atau negatif [5].

Penelitian ini tidak hanya penting untuk mengetahui elektabilitas Ridwan Kamil, tetapi juga memberikan wawasan mendalam mengenai pendekatan yang paling efektif dalam menganalisis sentimen publik di media sosial. Oleh karena itu, penelitian ini bertujuan untuk menemukan algoritma mana yang paling akurat dan efisien dalam menganalisis data sosial yang besar dan kompleks. Menjelang Pilkada DKI Jakarta 2024, urgensi penelitian ini semakin tinggi karena sentimen masyarakat di media sosial dapat mempengaruhi langkah-langkah kampanye yang akan diambil oleh para kandidat. Ridwan Kamil, sebagai salah satu kandidat potensial, dapat memanfaatkan hasil penelitian ini untuk memahami opini publik yang berkembang mengenai dirinya, sehingga ia bisa merancang

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

strategi kampanye yang lebih terarah dan efektif. Selain berkontribusi pada literatur akademik, penelitian ini diharapkan memberikan pemahaman mendalam tentang analisis sentimen tidak hanya bermanfaat untuk mengevaluasi pendapat publik, tetapi juga untuk merancang pesan-pesan kampanye yang sesuai dengan harapan masyarakat. Oleh karena itu, penulis tertarik untuk melakukan penelitian dengan judul "**Komparasi Algoritma *Naive Bayes* dan *Support Vector Machine (SVM)* dalam Menganalisis Elektabilitas Mochammad Ridwan Kamil pada Pilkada DKI 2024 melalui Data Media Sosial YouTube.**"

### 1.2 Identifikasi Masalah

Berdasarkan latar belakang yang telah dijelaskan, maka dapat diidentifikasi masalah yang ada sebagai berikut :

- a. Belum ada pendekatan yang efektif untuk menilai elektabilitas Mochammad Ridwan Kamil melalui analisis data komentar publik di YouTube secara komprehensif.
- b. Belum diketahui algoritma machine learning yang paling akurat dan efisien, antara *Naive Bayes* dan *Support Vector Machine (SVM)*, untuk mengolah data sentimen komentar publik di YouTube.

### 1.3 Rumusan Masalah

Rumusan masalah dalam penelitian ini adalah bagaimana efektivitas algoritma *Naive Bayes* dan *Support Vector Machine (SVM)* dalam menganalisis sentimen publik terhadap elektabilitas Mochammad Ridwan Kamil pada Pilkada DKI 2024. Penelitian ini bertujuan untuk membandingkan kedua algoritma tersebut dan menentukan mana yang lebih tepat untuk digunakan dalam konteks analisis sentimen di media sosial.

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

### 1.4 Batasan Masalah

Mengingat cakupan penelitian yang luas dan adanya keterbatasan waktu, penulis merasa perlu menetapkan batasan dalam penelitian ini agar fokus dan tujuan penelitian dapat tercapai dengan lebih efektif. Dengan demikian, batasan masalah yang ditetapkan dalam penelitian ini adalah sebagai berikut:

1. Penelitian ini hanya akan menggunakan data berupa komentar dan ulasan tentang Mochammad Ridwan Kamil yang diambil dari platform media sosial YouTube. Data yang dikumpulkan mencakup 10.843 entri yang berasal dari periode tahun 2024 pada bulan agustus hingga september, dengan pengumpulan dilakukan menggunakan bahasa pemrograman *Python* di *Jupyter Notebook*.
2. Fokus pengumpulan data akan terletak pada isi komentar dari pengguna YouTube mengenai Mochammad Ridwan Kamil, tanpa menganalisis elemen lain seperti video atau channel yang terkait.
3. Metode analisis yang diterapkan dalam penelitian ini adalah algoritma *Naïve Bayes* dan *Support Vector Machine (SVM)*. Proses pengolahan data dan penerapan algoritma dilakukan menggunakan bahasa pemrograman *Python* melalui platform *Jupyter Notebook*, dan tidak akan mencakup analisis menggunakan perangkat lunak lain.
4. Lingkup penelitian ini dibatasi pada analisis data dari media sosial YouTube untuk mengukur elektabilitas Mochammad Ridwan Kamil dalam pemilihan kepala daerah (Pilkada) DKI Jakarta 2024. Penelitian tidak akan mencakup faktor eksternal lainnya, seperti survei atau data dari media sosial lainnya, sehingga hasil yang diperoleh dapat lebih terfokus dan relevan.

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

5. Penelitian ini akan berfokus pada analisis sentimen positif dan negatif, yang dilakukan dengan komentar-komentar tersebut, serta tidak akan meneliti variasi faktor-faktor lain yang mungkin memengaruhi analisis elektabilitas.

### 1.5 Tujuan dan Manfaat Penelitian

#### 1.5.1 Tujuan Penelitian

##### a. Tujuan Umum

Penelitian ini bertujuan untuk menganalisis dan membandingkan efektivitas algoritma *Naive Bayes* dan *Support Vector Machine (SVM)* dalam menilai opini publik mengenai elektabilitas Mochammad Ridwan Kamil pada Pilkada DKI 2024, menggunakan data yang diperoleh dari media sosial YouTube.

##### b. Tujuan Khusus

Tujuan khusus penelitian ini adalah untuk memenuhi syarat penyusunan skripsi S-1 Program Studi Informatika, Fakultas Ilmu Teknik di Universitas Bina Insan Lubuklinggau, serta memberikan kontribusi dalam pengembangan ilmu pengetahuan dan teknologi, sekaligus meningkatkan pemahaman dan keterampilan mahasiswa dalam menerapkan teori-teori yang telah dipelajari.

#### 1.5.2 Manfaat Penelitian

##### a. Manfaat bagi Ilmu Pengetahuan

Penelitian ini diharapkan dapat memberikan kontribusi yang signifikan terhadap pengembangan ilmu pengetahuan, khususnya dalam bidang analisis sentimen. Dengan menerapkan Algoritma *Naive Bayes* dan *Support Vector Machine (SVM)*, penelitian ini akan memperkaya pemahaman mengenai cara menganalisis sentimen publik terhadap elektabilitas Mochammad Ridwan Kamil berdasarkan komentar di platform media sosial YouTube. Hasil dari penelitian ini juga dapat dijadikan referensi bagi peneliti lain yang ingin mendalami topik serupa.

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

### b. Manfaat bagi Masyarakat Umum

Dari sudut pandang masyarakat, penelitian ini akan menyediakan informasi yang berharga mengenai opini publik terhadap Mochammad Ridwan Kamil di Pilkada 2024. Dengan menganalisis komentar dan ulasan di YouTube, masyarakat dapat lebih memahami bagaimana publik menilai calon tersebut, yang pada gilirannya dapat memengaruhi pandangan mereka dalam menentukan pilihan di pemilihan kepala daerah.

### c. Manfaat bagi Peneliti

Bagi peneliti itu sendiri, penelitian ini berfungsi sebagai wadah untuk menerapkan dan mengintegrasikan ilmu yang telah diperoleh selama pendidikan. Proses penulisan skripsi ini memberikan kesempatan untuk memperdalam wawasan peneliti mengenai metodologi penelitian, analisis data, serta teknik penulisan akademis yang baik.

## 1.6 Sistem Penulisan

Penulisan skripsi ini, yang merupakan laporan hasil penelitian, dirancang terdiri dari lima bab. Setiap bab memiliki isi yang terstruktur sebagai berikut:

### **BAB I : PENDAHULUAN**

Bab ini berisi latar belakang penelitian yang menjelaskan alasan dilakukannya penelitian, identifikasi masalah untuk memahami persoalan utama, rumusan masalah sebagai bentuk pertanyaan penelitian, batasan masalah untuk memperjelas ruang lingkup, tujuan dan manfaat penelitian sebagai hasil yang ingin dicapai, serta sistematika penulisan sebagai panduan pembaca mengenai alur isi skripsi.

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

### **BAB II : KAJIAN PUSTAKA**

Bab ini membahas teori-teori yang relevan sebagai dasar penelitian, hasil penelitian terdahulu yang relevan (jika ada), dan kerangka berpikir sebagai panduan logis dalam menjawab permasalahan yang diteliti.

### **BAB III : METODOLOGI PENELITIAN**

Bab ini menjelaskan secara rinci metode yang digunakan dalam penelitian, teknik atau metode pengumpulan data, metode analisis data, lokasi dan waktu penelitian, alat dan bahan yang digunakan, serta metode pengujian dan pengolahan data (populasi).

### **BAB IV : HASIL DAN PEMBAHASAN**

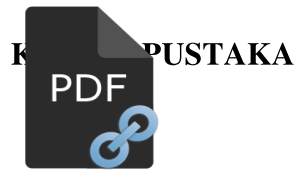
Bab ini menyajikan gambaran umum dari hasil penelitian yang diperoleh, analisis hasil penelitian yang dilakukan, pembahasan terkait interpretasi hasil, serta pengujian.

### **BAB V : KESIMPULAN DAN SARAN**

Bab ini merangkum kesimpulan dari seluruh penelitian berdasarkan hasil yang telah diperoleh, serta memberikan saran atau masukan yang diharapkan berguna untuk pengembangan penelitian di masa mendatang.

### **DAFTAR PUSTAKA**

### **LAMPIRAN**



## 2.1 Literatur

### 2.1.1 Text Mining

Text Mining atau Data mining adalah serangkaian proses untuk menggali nilai tambah berupa informasi yang selama ini tidak diketahui secara manual dari suatu basis data dengan melakukan penggalian pola-pola dari data dengan tujuan untuk memanipulasi data menjadi informasi yang lebih berharga yang diperoleh dengan cara mengekstraksi dan mengenali pola yang penting atau menarik dari data yang terdapat dalam basis data. Seperti halnya data mining, text mining adalah proses penemuan akan informasi atau trend baru yang sebelumnya tidak terungkap dengan memproses dan menganalisis data dalam jumlah besar. Text mining merupakan salah satu cabang ilmu data mining yang menganalisis data berupa dokumen teks. Menurut Han, Kamber, dan Pei (dalam Prilianti dan Wijaya), text mining adalah salah satu langkah otomatis oleh komputer untuk menggali informasi yang berkualitas dari suatu rangkaian teks yang terangkum dalam sebuah dokumen[6].

Pada intinya, text mining terdiri dari beberapa tahapan, dimulai dengan *preprocessing* atau pra-pemrosesan teks. *Preprocessing* menjadi tahap awal dalam klasifikasi teks untuk mempersiapkan data teks sebelum digunakan pada proses lainnya. Pada tahap ini, data teks akan diubah menjadi bentuk yang lebih baik sehingga menghasilkan informasi teks dengan kualitas yang baik dan siap digunakan pada proses selanjutnya[7]. Tahapan ini mencakup pemecahan teks menjadi bagian-bagian kecil yang disebut *tokenization*, mengubah kata menjadi bentuk dasarnya melalui proses *stemming*, serta menghilangkan kata-kata yang tidak penting atau sering muncul, yang dikenal sebagai *stopword removal*. Setelah tahap pra-

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

pemrosesan selesai, metode analisis lanjutan seperti klasifikasi, *clustering*, dan analisis asosiasi dapat diterapkan untuk mengolah teks tersebut. Klasifikasi digunakan untuk mengelompokkan teks ke dalam kategori yang sudah ditentukan sebelumnya, sedangkan *clustering* lebih berfokus pada identifikasi pola atau kelompok baru dalam data yang belum diberi label.

Dari berbagai definisi yang telah dijelaskan, dapat disimpulkan bahwa text mining merupakan proses analisis terhadap teks yang dilakukan secara otomatis oleh komputer untuk mengekstrak informasi berkualitas dari kumpulan data besar. Proses ini mencakup serangkaian tahapan, termasuk *tokenization*, *stemming*, dan *stopword removal*, yang pada akhirnya mempersiapkan data untuk analisis lebih lanjut.

### a. Analisis



Analisis adalah sikap atau perhatian terhadap sesuatu benda, fakta, dan fenomena, sehingga mampu menguraikan menjadi bagian-bagian serta mengenal kaitan bagian tersebut dalam keseluruhan[8]. Dalam konteks penelitian, analisis bertujuan untuk mengidentifikasi pola, hubungan, dan informasi yang relevan dari data yang tersedia. Metode analisis dapat bervariasi tergantung pada jenis data yang digunakan, seperti data kualitatif atau kuantitatif. Proses ini sering melibatkan langkah-langkah seperti pengumpulan data, pengolahan data, dan penggunaan teknik statistik untuk menghasilkan kesimpulan yang valid. Dengan menganalisis data, peneliti dapat memberikan wawasan yang mendalam, mendukung pengambilan keputusan, serta merumuskan rekomendasi yang berdasarkan bukti.

Dalam penelitian, analisis memiliki peran sentral untuk memperoleh informasi bermakna dari data yang tersedia. Tahapan-tahapan dalam proses analisis bertujuan untuk menguraikan data secara sistematis, sehingga peneliti dapat memahami pola-pola tersembunyi

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

dan menemukan keterkaitan yang signifikan antara variabel yang ada.

Dengan demikian,  ini berfungsi untuk mendapatkan sebuah informasi dari suatu  data yang ada[9].

### b. Analisis Sentimen

Analisis sentimen adalah proses untuk mengidentifikasi dan mengekstraksi opini, emosi, atau sikap yang tersirat atau tersurat dalam teks dan diklasifikasikan menjadi opini positif, negatif atau netral[10]. Dalam penelitian ini, analisis sentimen dapat digunakan untuk mengevaluasi bagaimana pandangan masyarakat terhadap calon pemimpin seperti Mochammad Ridwan Kamil, berdasarkan komentar dan interaksi di platform media sosial seperti YouTube. Dengan menggunakan teknik analisis sentimen, peneliti dapat mengkategorikan opini menjadi positif, negatif, atau netral, memberikan gambaran yang jelas mengenai persepsi publik terhadap topik yang diteliti. Proses analisis sentimen biasanya melibatkan beberapa tahap, dimulai dengan pra-pemrosesan teks untuk menyiapkan data agar lebih mudah dianalisis. Setelah itu, algoritma tertentu, seperti *Naive Bayes* dan *Support Vector Machine (SVM)*, dapat digunakan untuk mengklasifikasikan sentimen. Selain itu, visualisasi data juga sering digunakan untuk memperjelas hasil analisis dan membantu dalam pengambilan keputusan.

Analisis sentimen banyak digunakan di berbagai bidang seperti *e-commerce*, kesehatan, hiburan dan politik. Dalam bidang pemasaran analisis sentimen dapat digunakan untuk membedakan sikap pelanggan, pengguna atau pengikut terhadap suatu merek, topik ataupun produk dengan bantuan ulasan dari mereka[11]. Dalam konteks politik, analisis sentimen memberikan wawasan tentang penerimaan publik terhadap calon pemimpin. Seiring dengan perkembangan teknologi, analisis sentimen semakin penting karena mampu menyediakan informasi

## Protected by PDF Anti-Copy Free

[\(Upgrade to Pro Version to Remove the Watermark\)](#)

berguna dari data besar dan beragam, membantu peneliti dan pengambil keputusan untuk membuat keputusan yang lebih baik berdasarkan pemahaman yang lebih dalam tentang opini masyarakat.



### c. YouTube

YouTube adalah platform berbagi video terbesar di dunia yang didirikan pada tahun 2005 oleh tiga mantan karyawan PayPal, yaitu Chad Hurley, Steve Chen, dan Jawed Karim. Saat ini, YouTube dimiliki oleh Google dan menjadi salah satu situs web paling populer dengan jutaan konten video yang mencakup beragam topik, mulai dari hiburan, pendidikan, tutorial, berita, hingga vlog pribadi. Pengguna dapat mengunggah, menonton, membagikan, serta memberikan respons terhadap video melalui fitur komentar, likes, dan subscribe. Salah satu layanan penyiaran populer berbasis audio-visual ini memiliki pengguna yang mencapai 95% dari total pengguna internet, dan konten yang tersedia di YouTube mencakup 80 bahasa yang berbeda[12]. Menjadikannya media yang sangat efektif untuk berbagi informasi dan hiburan. Platform ini juga sering dimanfaatkan oleh berbagai kalangan, termasuk tokoh publik, politisi, serta perusahaan untuk memperluas jangkauan mereka.

Media sosial, seperti YouTube, memiliki banyak pengguna pada era saat ini. Berdasarkan laporan data dari Hootsuite (We Are Social) pada 2020, YouTube merupakan platform media sosial yang paling banyak dipakai oleh masyarakat Indonesia, yakni 88% dari jumlah populasi. Pengguna media sosial terbiasa untuk mengutarakan opininya melalui komentar pada video YouTube yang menarik perhatiannya[13]. Data yang dihasilkan dari platform ini sering digunakan dalam analisis sentimen, yang memberikan gambaran mengenai reaksi pengguna terhadap isu-isu tertentu. Interaksi langsung antara pembuat konten dan audiens menjadikan YouTube sebagai sumber data yang kaya untuk

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

penelitian di berbagai bidang, termasuk analisis politik, pemasaran digital, dan perilaku

### d. *Preprocessing*



*Preprocessing* merupakan langkah krusial dalam analisis data yang bertujuan untuk menyiapkan data sebelum melakukan analisis lebih lanjut. Tahapan ini merupakan proses ekstraksi teks, dari data tidak terstruktur diubah menjadi data yang terstruktur agar bisa diolah lebih lanjut untuk proses klasifikasi. Tujuannya adalah agar dokumen teks yang akan dianalisis tidak mengandung banyak kata yang tidak diinginkan. Tahapan yang biasa dilakukan adalah penghapusan karakter yang tidak perlu (*cleansing*), mengubah kata yang tidak formal menjadi formal (*formalization*), dan mengambil kata-kata yang dirasa penting (*filtering*)[14]. Secara umum, terdapat enam tahapan dalam preprocessing yang akan dilaksanakan yaitu:

#### 1) *Data Cleaning*

Proses identifikasi, koreksi, dan penghapusan ketidaksesuaian atau anomali dalam dataset untuk meningkatkan kualitas data, memastikan keakuratan analisis, dan memungkinkan visualisasi data yang informatif[15]. Pembersihan data memastikan bahwa hanya informasi yang bermanfaat yang tersisa, sehingga memudahkan analisis selanjutnya.

#### 2) *Case Folding*

*Transform case* atau *case folding* merupakan proses pemerataan huruf dari huruf kapital menjadi huruf kecil atau sebaliknya[16].

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

### 3) *Tokenization*

Tokenization adalah proses yang dilakukan untuk memecah konten tekstual menjadi kata-kata, istilah, simbol, atau beberapa elemen bermakna lainnya. Hasil pemecahan ini kemudian disebut dengan token[17].



### 4) *Normalisasi*

Normalisasi mengubah semua bentuk kata yang diberikan menjadi satu bentuk yang baku[18]. Dengan melakukan normalisasi, kata-kata yang memiliki makna sama dapat diperlakukan secara seragam, sehingga mempermudah analisis.

### 5) *Stopword Removal*

Teknik menghapus kata-kata yang umum digunakan dan kata yang tidak memiliki arti khusus seperti kata ganti, preposisi, dan konjungsi[19]. Dengan demikian, fokus dapat ditempatkan pada kata-kata yang lebih penting dan relevan dalam konteks analisis.

### 6) *Stemming*

Bertujuan untuk menghilangkan imbuhan awalan, akhiran dan sisipan untuk kembali dalam bentuk kata dasar sesuai KBBI[20]. Hal ini membantu mengurangi variasi dalam data, memungkinkan analisis yang lebih akurat dengan fokus pada makna inti dari kata-kata yang digunakan.

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

### e. Klasifikasi

Klasifikasi adalah proses yang dilakukan sebagai teknik untuk membentuk model dari contoh data pelatihan. Klasifikasi akan menganalisis data dan membentuk model dengan menggambarkan kelas data[21]. Dalam pembelajaran mesin, algoritma klasifikasi memprediksi kategori atau label dari data baru dengan menggunakan model yang telah dilatih. Prosesnya melibatkan penggunaan data historis yang sudah terklasifikasi untuk melatih model, sehingga model mampu mengenali pola dan karakteristik pada data baru. Contoh penerapannya adalah dalam analisis sentimen, di mana komentar atau ulasan dapat dikategorikan sebagai positif, negatif, atau netral berdasarkan pola teks.

### f. Ulasan

Ulasan adalah tanggapan atau pendapat yang diberikan oleh pengguna atau audiens terhadap produk, layanan, atau konten yang mereka konsumsi. Jutaan individu aktif memanfaatkan platform media sosial youtube untuk berbagi pandangan, opini, dan pengalaman mereka terkait produk, layanan, dan merek tertentu. Ulasan ini tidak hanya menjadi sumber informasi berharga secara kuantitatif dan kualitatif bagi bisnis, tetapi juga memperlihatkan dampak signifikan media sosial pada dinamika interaksi antara pelanggan dan Perusahaan[22].

Selain sebagai umpan balik, komentar dan ulasan juga berperan penting dalam membentuk opini publik. Opini publik ini bisa memengaruhi persepsi dan preferensi masyarakat terhadap calon pemimpin, seperti dalam pemilihan kepala daerah, termasuk Pilkada DKI 2024. Komentar positif atau negatif mengenai seorang kandidat, seperti Mochammad Ridwan Kamil, dapat membentuk citra publiknya dan, pada akhirnya, memengaruhi elektabilitasnya.

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)  
g. *Naive Bayes*

*Naive Bayes* adalah algoritma yang dapat mengklasifikasikan variabel menggunakan metode probabilistik dan statistik[23]. *Naive Bayes* adalah metode yang tidak memiliki aturan dan menggunakan cabang matematika yang disebut teori probabilitas untuk mendapatkan peluang setinggi mungkin dengan melihat frekuensi atau jumlah kemunculan setiap klasifikasi dalam data pelatihan[24]. Algoritma ini sering digunakan dalam berbagai aplikasi, seperti klasifikasi teks, analisis sentimen, dan pengenalan pola, karena kesederhanaannya dan kemampuannya dalam menangani dataset besar dengan cepat. Berikut contoh dari penerapan Algoritma *Naive Bayes* :

1. Analisis Sentimen Aplikasi Brimo Pada Ulasan Pengguna Di Google Play Menggunakan Algoritma *Naive Bayes* (Studi kasus Program Studi Teknik Informatika).
2. Analisis Sentimen Terhadap *Cryptocurrency* Berbasis *Python* TextBlob Menggunakan Algoritma *Naive Bayes* (Studi kasus Program Studi Magister Sistem Informasi, STMIK LIKMI Bandung).
3. Analisis Data Minat Calon Mahasiswa Universitas Pamulang dengan Menggunakan Algoritma *Naive Bayes Classifier* (Studi kasus Program Studi Sistem Informasi Universitas Pamulang) .

Adapun rumus umum dari Algoritma *Naive Bayes* sebagai berikut:

$$P(H|X) = \frac{P(H) \prod P(X_i|H)}{P(X)} \quad (1)$$

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

Keterangan:

$X$	=	D	gan kelas yang belum diketahui
$H$	=	H	PDF ata X
$P(H X)$	=	P	s hipotesis
$P(H)$	=		Probabilitas hipotesis H (prior probability)
$P(H H)$	=		Probabilitas X/H
$P(X)$	=		Probabilitas dari X

Adapun Aturan Bayes adalah:

Jika  $P(h_1|x) < P(h_2|x)$  maka  $x$  diklasifikasikan sebagai  $h_2$ . Pernyataan  $P(h_1|x)$  mengindikasikan probabilitas hipotesis  $h_1$  berdasarkan kondisi  $x$  terjadi, begitu juga dengan  $h_2$ . Sehingga dapat klasifikasi dari  $x$  sesuai dengan probabilitas terbesar di antara probabilitas  $x$  terhadap semua kelas.

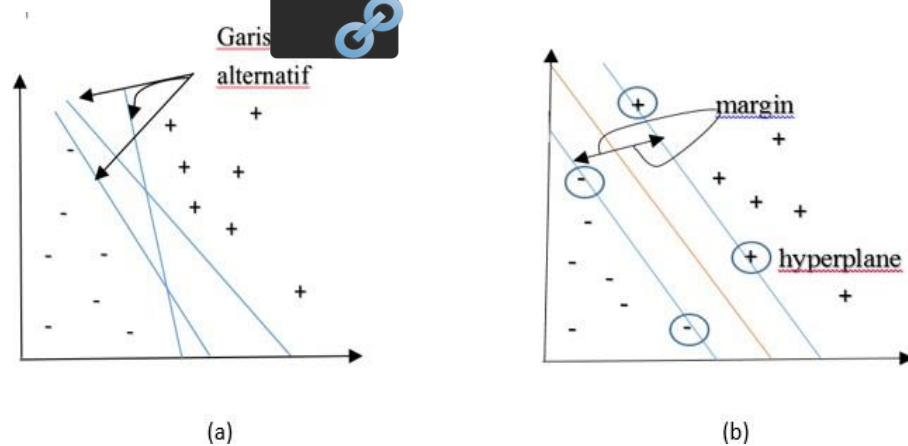
### h. Support Vector Machine

*Support Vector Machine (SVM)* merupakan suatu metode klasifikasi yang dapat mengidentifikasi titik baru dengan akurasi matematis yang tinggi. Jika suatu dataset dapat dilakukan pemisahan secara linier, maka bisa menggunakan pendekatan Hard margin sehingga jarak antara keduanya sebesar mungkin, dengan demikian dapat diidentifikasi kelas mana yang dimiliki setiap titik dataset. Tetapi sebagian besar waktu kumpulan data tidak dapat dipisahkan secara linier sehingga dapat dilakukan dua cara yaitu menggunakan pendekatan Linear menggunakan Soft Margin, atau dengan menggunakan Non-Linear Kernel (yang harus memenuhi Mercer Condition) atau lebih tepatnya pemetaan data pada dimensi di atasnya. SVM merupakan metode machine learning yang dapat digunakan untuk klasifikasi baik data linear maupun non-linear. SVM masuk ke dalam kelas supervised learning. Ide dasar SVM adalah memaksimalkan batas *hyperplane*[25]. *Model Support Vector Machine (SVM)* membangun

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

*hyperplane* (bidang pemisah) yang memiliki margin maksimum antara dua kelas, sehingga memaksimalkan pemisahan antara kelas positif dan kelas negatif.



**Gambar 2.1 SVM berusaha mencari *hyperplane* terbaik**

Sumber:[27]

Gambar 1 (a) menunjukkan *hyperplane* pemisah terbaik antara kedua kelas dapat ditemukan dengan mengukur margin *hyperplane* tersebut dan mencari titik maksimalnya. Margin adalah jarak antara *hyperplane* tersebut dengan pola terdekat dari masing-masing kelas. Pola yang paling dekat ini disebut sebagai *support vector*. Garis solid pada Gambar 1 (b) menunjukkan *hyperplane* yang terbaik, yaitu yang terletak tepat pada tengah-tengah kedua kelas, sedangkan titik merah dan kuning yang berada dalam lingkaran hitam adalah *support vector*.

### i. *Confusion Matrix*

*Confusion matrix* adalah alat evaluasi yang digunakan untuk mengukur kinerja model klasifikasi dalam machine learning. Matriks ini berbentuk tabel yang memperlihatkan jumlah prediksi benar dan salah dari suatu model klasifikasi, dengan membandingkan hasil prediksi model terhadap nilai aktual yang sebenarnya. *Confusion matrix* menyajikan informasi tentang jumlah prediksi yang benar dan yang salah yang dibuat oleh model pada suatu dataset[28]. Pada dasarnya,

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

*confusion matrix* terdiri dari empat elemen: *True Positive (TP)*, *True Negative (TN)*, *False Positive (FP)*, dan *False Negative (FN)*. *True Positive (TP)* menunjukkan jumlah kasus di mana model memprediksi kelas positif dengan benar. Sementara *True Negative (TN)* adalah jumlah kasus di mana model memprediksi kelas negatif dengan benar. Sebaliknya, *False Positive (FP)* adalah jumlah kesalahan ketika model memprediksi positif, padahal seharusnya negatif, dan *False Negative (FN)* adalah ketika model memprediksi negatif padahal seharusnya positif.

Berdasarkan data yang diberikan oleh *confusion matrix*, kita dapat menghitung berbagai metrik evaluasi seperti akurasi, *presisi*, *recall*, dan *F1-score*. Evaluasi yang dilakukan dalam *confusion matrix* akan menghasilkan nilai akurasi, *presisi*, dan *recall*. Akurasi dalam klasifikasi adalah persentase dalam ketepatan pada record data yang telah diklasifikasi dengan benar setelah dilakukan pengujian pada hasil klasifikasi[29]. Di samping itu, *presisi* menghitung proporsi prediksi positif yang benar, sedangkan *recall* menunjukkan seberapa baik model dalam menemukan semua kasus positif. *F1-score* menggabungkan *presisi* dan *recall* menjadi satu metrik yang seimbang. *Confusion matrix* memberikan gambaran visual tentang kekuatan dan kelemahan model, memungkinkan kita untuk memahami pola kesalahan dan meningkatkan performa model melalui tuning parameter atau pemilihan fitur yang lebih tepat. Beberapa istilah yang terdapat dalam *confusion matrix* seperti yang direpresentasikan pada Tabel 1 antara lain:

**Tabel 2.1** Representasi *Confusion Matrix*

<i>Class/Actual</i>	<i>Positive</i>	<i>Negative</i>
<i>Positive</i>	<i>True Positive (TP)</i>	<i>False Negative (FN)</i>
<i>Negative</i>	<i>False Positive (FP)</i>	<i>True Negative (TN)</i>

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

Keterangan:

- a) TP (True Positive) adalah jumlah prediksi benar pada kelas sentimen positif
- b) TN (True Negative) adalah jumlah prediksi benar pada kelas sentimen negatif
- c) FP (False Positive) adalah jumlah prediksi salah pada kelas sentimen positif
- d) FN (False Negative) adalah jumlah prediksi salah pada kelas sentimen negatif

Akurasi, *precision*, dan *recall* adalah metrik evaluasi yang dihitung menggunakan informasi dalam *confusion matrix*, dengan rumus sebagai berikut:

### 1) Akurasi

Akurasi mengukur sejauh mana hasil prediksi yang dihasilkan oleh suatu model sesuai dengan nilai sebenarnya. Ini merupakan persentase dari jumlah keseluruhan prediksi yang benar dibandingkan dengan total data yang diujikan. Dalam konteks analisis sentimen atau klasifikasi lainnya, akurasi menunjukkan seberapa baik model dalam memprediksi kelas dengan benar, baik itu kelas positif maupun negatif. Berikut contoh Penerapan proses perhitungan akurasi:

1. Pada jurnal "Analisis Sentimen Aplikasi TikTok Menggunakan Algoritma *Naïve Bayes* dan *Support Vector Machine*", diperoleh hasil akurasi *Naïve Bayes* sebesar 79% dan akurasi SVM 84%[30].

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

$$\text{Accuracy Naive Bayes} = \frac{15 + 283}{15 + 1 + 78 + 283} \times 100\% = 0.79$$

$$\text{Accuracy Svm} = \frac{43 + 274}{38 + 22 + 274} \times 100\% = 0.84$$

Rumus akurasi

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (2)$$

### 2) Precision

*Precision* adalah metrik yang menggambarkan proporsi prediksi yang benar untuk suatu kelas tertentu dari keseluruhan prediksi yang diklasifikasikan ke dalam kelas tersebut. Dengan kata lain, *precision* menunjukkan seberapa banyak prediksi positif yang benar-benar relevan dibandingkan dengan total prediksi positif yang dihasilkan oleh model. Berikut contoh penerapan proses perhitungan *Precision*:

1. Pada jurnal "Analisis Sentimen Aplikasi TikTok Menggunakan Algoritma *Naive Bayes* dan *Support Vector Machine*".

$$\text{Precision Naive Bayes} = \frac{15}{15 + 1} = 0.93$$

$$\text{Precision Svm} = \frac{43}{43 + 22} = 0.66$$

Rumus *Precision*:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (3)$$

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

### 3) Recall

*Recall* yang dikenal sebagai sensitivitas atau true positive rate, adalah metrik yang digunakan untuk mengukur seberapa baik model dalam mengidentifikasi semua contoh positif yang sebenarnya. Dengan kata lain, *recall* menunjukkan proporsi dari total kasus positif yang berhasil diidentifikasi oleh model dari keseluruhan kasus positif yang ada. Berikut contoh penerapan proses perhitungan *Recall*:

1. Pada jurnal "Analisis Sentimen Aplikasi TikTok Menggunakan Algoritma *Naïve Bayes* dan *Support Vector Machine*".

$$\text{Recall Naive Bayes} = \frac{15}{15 + 78} = 0.16$$

$$\text{Recall Svm} = \frac{43}{43 + 38} = 0.53$$

Rumus *Recall*:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (4)$$

### 4) F- Measure (F1 Score) FI

F1- skor adalah ukuran yang paling banyak digunakan dalam evaluasi berbasis batas, memperkenalkan kerangka presisi-*recall*[31]. F1 Score merupakan rata-rata harmonis dari *precision* dan *recall*, dua metrik penting dalam evaluasi model. *Precision* menunjukkan proporsi prediksi positif yang benar, sedangkan *recall* mengukur sejauh mana model dapat mendeteksi semua kasus positif dalam data. Dengan menghitung rata-rata harmonis, F1 Score memberikan nilai seimbang, terutama saat *precision* dan *recall* memiliki perbedaan signifikan.

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

Nilai F1 Score berkisar antara 0 hingga 1, di mana nilai 1 menunjukkan performa sempurna dari model dalam memprediksi kelas positif dengan presisi tinggi, sementara nilai 0 menandakan bahwa model gagal mendeteksi atau memprediksi kelas positif dengan benar. Metrik ini sangat berguna untuk dataset yang tidak seimbang karena memastikan bahwa baik precision maupun recall dipertimbangkan secara proporsional dalam evaluasi, memberikan gambaran performa model yang lebih akurat dan dapat dipercaya.

### j. Python

*Python* adalah *scripting language* yang berorientasi objek. Bahasa pemrograman ini dapat digunakan untuk pengembangan perangkat lunak dan bisa dijalankan melalui berbagai sistem operasi. Saat ini, *Python* juga merupakan bahasa yang populer bagi bidang data science dan analisis[32]. Dikenal karena sintaksisnya yang sederhana dan mudah dipahami, *Python* menjadi pilihan ideal bagi pemula untuk mempelajari dasar-dasar pemrograman. Selain itu, Python mendukung paradigma pemrograman berorientasi objek dan fungsional, memberikan fleksibilitas kepada pengembang dalam memilih pendekatan yang sesuai untuk kebutuhan proyek mereka.

*Python* menjadi pilihan utama di kalangan data scientist berkat ekosistem pustakanya yang luas. Berikut adalah beberapa tools dan library Python yang umum digunakan dalam analisis data:

1. *Pandas* : digunakan untuk manipulasi dan analisis data dalam format tabel (DataFrame).
2. *NumPy* : digunakan untuk komputasi numerik dan operasi array multidimensi.
3. *Matplotlib* : digunakan untuk membuat berbagai jenis grafik dan visualisasi data.

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

4. *Scikit-learn* : digunakan untuk penerapan algoritma machine learning seperti klasifikasi dan regresi.
5. *Scrapers API* : digunakan untuk mengambil data dari YouTube, seperti video dan komentar.
6. *Seaborn* : digunakan untuk visualisasi data statistik yang lebih estetik dan kompleks.

*Python* juga memiliki kemampuan integrasi yang sangat baik dengan berbagai platform dan teknologi, memudahkan pengembangan aplikasi yang kompleks. *Python* dapat dengan mudah dihubungkan dengan database, API, dan alat lainnya untuk menciptakan solusi yang terintegrasi. Dengan fleksibilitas dan dukungan komunitas yang luas, *Python* tetap menjadi bahasa pemrograman yang relevan dan populer di dunia teknologi saat ini.

Keunggulan lain dari *Python* adalah kemampuannya untuk berjalan di berbagai platform sistem operasi, seperti Windows, macOS, Linux, dan bahkan platform mobile seperti Android dan iOS melalui framework tertentu. Aplikasi yang dikembangkan menggunakan Python dapat dengan mudah dipindahkan dan dijalankan pada platform yang berbeda tanpa banyak modifikasi. *Python* juga mendukung berbagai perangkat keras, termasuk server, komputer pribadi, dan perangkat tertanam, menjadikannya bahasa pemrograman lintas platform yang fleksibel.

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

### 2.2 Penelitian Relevan

Penelitian yang relevan merupakan kajian yang memiliki keterkaitan atau kemiripan dengan penelitian yang sedang dilakukan, baik dari segi topik, metodologi, maupun pendekatan yang digunakan. Dalam konteks skripsi ini, penelitian yang relevan dapat mencakup studi-studi sebelumnya yang juga menggunakan algoritma *machine learning*, seperti *Naive Bayes* dan *Support Vector Machine (SVM)*, untuk analisis sentimen politik atau pengukuran elektabilitas calon kepala daerah. Oleh karena itu, penulis melakukan kajian terhadap beberapa hasil penelitian yang diterbitkan dalam jurnal-jurnal yang tersedia di internet, antara lain sebagai berikut:

Peneliti pertama yaitu Iin dkk, mengenai Penerapan Natural Language Processing (NLP) dalam Analisis Sentimen CAWAPRES 2024 menggunakan algoritma *Naive Bayes*. Penelitian tersebut menunjukkan bahwa metode *Naive Bayes* mencapai akurasi sebesar 77,37%. Nilai *recall* dari metode tersebut adalah 73,98%, sementara nilai presisinya mencapai 100%[16].

Peneliti kedua yaitu Chely Aulia Misrun dkk, mengenai Analisis Sentimen Komentar YouTube Terhadap Anies Baswedan Sebagai Bakal Calon Presiden 2024 Menggunakan Metode *Naive Bayes Classifier*. Dalam penelitian ini. Dari total 1.009 komentar yang dianalisis, terdapat 610 komentar positif dan 399 komentar negatif. Hasil akurasi yang diperoleh dari model *Naive Bayes Classifier* ini adalah 78%, menggunakan perbandingan 10% data uji dan 90% data latihan[20].

Peneliti ketiga yaitu Alberi Meidharma Fadli Hulu dan Kemas Muslim Lhaksana, mengenai Analisis Sentimen Politik pada Twitter Menggunakan Metode *Support Vector Machine* (Studi Kasus: Pilpres 2019). Hasil dari penelitian menunjukkan bahwa akurasi tertinggi yang dicapai oleh model adalah 62,88%, dengan pendekatan gabungan Unigram, Bigram, dan Trigram[5].

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

Peneliti keempat yaitu Arfina Handayani dan Ilka Zufria, mengenai Analisis Sentimen Terhadap Bakal Capres RI 2024 di Twitter Menggunakan Algoritma SVM. Studi ini memanfaatkan media sosial Twitter untuk menganalisis sentimen masyarakat terhadap tiga bakal calon presiden: Anies Baswedan, Ganjar Pranowo, dan Prabowo Subianto. Metode yang digunakan adalah *Support Vector Machine (SVM)* untuk mengklasifikasikan sentimen tweet, dan penelitian ini menghasilkan akurasi sekitar 78,3%[33].

Peneliti kelima yaitu Dafa Rosi Berliana dan Budi Santoso mengenai Elektabilitas Ridwan Kamil dan Anies Baswedan dalam Simulasi Pilpres 2024 di Twitter. Penelitian ini menggunakan Sentiment Analysis dan Social Media Network Analysis untuk menganalisis tweet yang mengandung hashtag #RidwanKamil dan #AniesBaswedan. Berdasarkan data yang diambil dari 10.000 tweet untuk Anies dan 9.815 tweet untuk Ridwan, hasilnya menunjukkan bahwa Anies memperoleh 29% sentimen negatif, 65% sentimen positif, dan 6% sentimen netral. Sementara itu, Ridwan Kamil menerima 8% sentimen negatif, 89% sentimen positif, dan 3% sentimen netral[34].

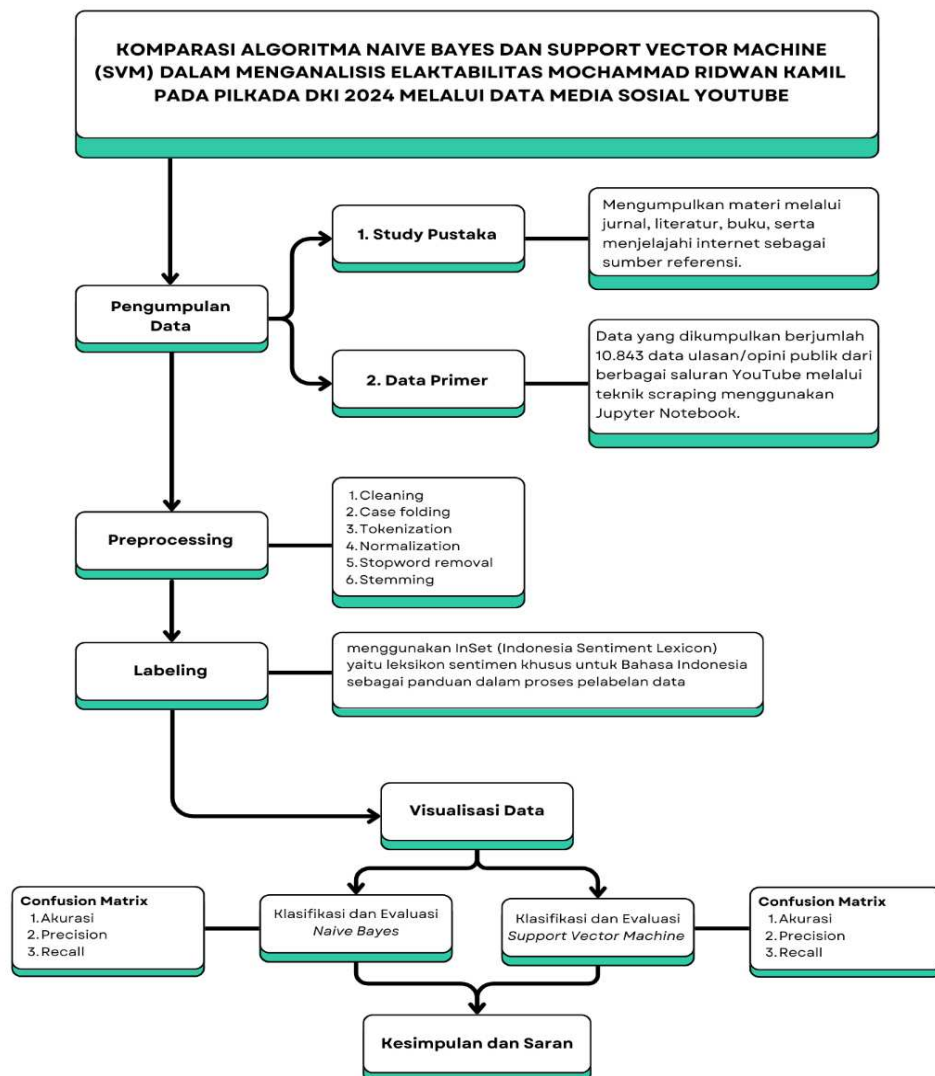
Peneliti keenam Dede Sandi dkk, mengenai Analisis Sentimen Publik Terhadap Elektabilitas Ganjar Pranowo di Tahun Politik 2024 di Twitter dengan Algoritma KNN dan *Naïve Bayes*. Studi ini menganalisis sentimen masyarakat terhadap Ganjar Pranowo, calon presiden untuk pemilihan 2024, menggunakan data dari Twitter. Metode yang digunakan adalah K-Nearest Neighbor (KNN) dan *Naïve Bayes* untuk klasifikasi sentimen dari data yang diambil. Penelitian ini menunjukkan bahwa model KNN mencapai akurasi tertinggi sebesar 99%, sedangkan *Naïve Bayes* mendapatkan akurasi sebesar 96%. Hasil analisis menunjukkan bahwa 96,6% dari sentimen yang diekstrak adalah positif, sementara 3,4% bersifat negatif, yang mencerminkan dominasi pandangan positif masyarakat terhadap Ganjar Pranowo menjelang pemilihan umum[1].

# Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

## 2.3 Kerangka Berfikir

Kerangka berfikir adalah struktur yang membantu peneliti dalam merancang dan melaksanakan suatu penelitian. Dalam konteks penelitian, kerangka berfikir berfungsi sebagai panduan yang menghubungkan antara teori, konsep, dan variabel yang akan diteliti. Dengan mendefinisikan hubungan antar elemen ini, peneliti dapat lebih mudah memahami tujuan, fokus, serta arah penelitian yang akan dilakukan. Kerangka berfikir penelitian dapat dilihat pada Gambar 2.2.



Gambar 2.2 Kerangka Berfikir

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

Kerangka berpikir penelitian ini menjelaskan langkah-langkah yang ditempuh dalam menganalisis elektabilitas Mochammad Ridwan Kamil pada Pilkada DKI 2020 menggunakan algoritma *Naive Bayes* dan *Support Vector Machine (SVM)* dengan data dari komentar YouTube. Berikut penjelasan kerangka berpikir tersebut:

### 1. Pengumpulan Data dibagi menjadi 2 bagian yaitu:


*Studi Pustaka Tahap* ini melibatkan pengumpulan informasi melalui jurnal, literatur, buku, serta sumber online sebagai referensi dan landasan teori.

*Data Primer* Data primer berjumlah 10.843 komentar atau opini publik yang diambil dari berbagai saluran YouTube. Pengumpulan data dilakukan dengan metode *scraping* menggunakan *Jupyter Notebook*.

### 2. *Preprocessing* menjadi tahap awal dalam klasifikasi teks untuk mempersiapkan data teks sebelum digunakan pada proses lainnya. Pada tahap ini akan mengubah data teks menjadi bentuk yang lebih baik sehingga menghasilkan informasi teks dengan kualitas yang baik dan siap digunakan pada proses selanjutnya[5]. Data yang dikumpulkan dari media sosial sering kali mengandung informasi tidak relevan, seperti simbol, emotikon, atau huruf kapital yang tidak seragam. Melalui *preprocessing*, data ini difilter dan dinormalisasi agar menjadi konsisten, termasuk dengan mengubah semua huruf menjadi kecil, menghapus kata-kata umum yang tidak memiliki makna penting (*stopword*), dan mengubah kata ke bentuk dasar (*stemming*). Proses ini memastikan data yang dianalisis bersih dan siap diproses secara optimal.

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

3. *Labeling* pada intinya adalah sebuah proses melabeli atau menandai sebuah informasi yang relevan pada sebuah dataset untuk kemudian dicari mengenai makna yang terkandung didalamnya[28].  ini dibuat untuk mengklasifikasikan sentimen yang terkandung dalam komentar atau opini yang ada. Komentar yang mendukung atau memuji topik akan diberi label "positif," sedangkan komentar yang mengkritik atau menunjukkan ketidakpuasan akan diberi label "negatif." Dengan adanya *labeling* ini, proses klasifikasi data dapat dilakukan lebih mudah dan akurat.
4. Visualisasi dilakukan dengan tujuan untuk membantu semua orang dalam memahami informasi dan menginterpretasi arti dari informasi tersebut. Selain itu, visualisasi juga bertujuan untuk menganalisis informasi mengenai hubungan antar data secara grafis[29]. Melalui grafik atau diagram, pola-pola dan distribusi data dapat terlihat dengan jelas, memudahkan peneliti dalam mengenali tren atau kecenderungan tertentu. Misalnya, visualisasi dapat menunjukkan perbandingan jumlah sentimen positif dan negatif atau kata-kata yang paling sering muncul dalam komentar. Dengan cara ini, hasil analisis menjadi lebih informatif dan mudah ditafsirkan.
5. Klasifikasi dan Evaluasi Algoritma Tahap ini melibatkan penerapan algoritma tertentu, seperti *Naive Bayes* dan *Support Vector Machine (SVM)*, untuk mengklasifikasikan data yang telah diberi label. Evaluasi dapat dilakukan dengan cara mengamati dan menganalisa hasil dari algoritma yang digunakan untuk memastikan bahwa hasil pengujian itu benar atau tidak sesuai dengan pembahasan. Sehingga dapat dipilih model yang paling efektif dan akurat dalam mengklasifikasikan sentimen publik.

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

6. Setelah tahap klasifikasi dan evaluasi algoritma maka tahap akhir ialah Kesimpulan dan Saran. Pada bagian akhir penelitian, peneliti merangkum temuan utama dari hasil analisis dan menyajikan kesimpulan yang menggambarkan kondisi umum elektabilitas tokoh atau isu yang diteliti. Kesimpulan ini juga disertai dengan rekomendasi atau saran yang relevan berdasarkan hasil yang diperoleh, untuk mendukung pengambilan keputusan atau penelitian lebih lanjut. Dengan demikian, penelitian ini tidak hanya memberikan hasil analisis tetapi juga kontribusi yang bermanfaat untuk studi di masa mendatang.



### **3.1 Analisa Sistem**

#### **3.1.1 Analisa sistem yang berjalan**

Penelitian ini menganalisis ulasan dari dua saluran YouTube, yaitu Najwa Shihab dan Deddy Corbuzier Close the Door. Data yang digunakan berupa komentar dan ulasan pengguna YouTube terkait Pilgub DKI, khususnya mengenai Mohammad Ridwan Kamil, yang dikumpulkan selama periode Agustus hingga September 2024. Total data yang berhasil dihimpun mencapai 10.843 komentar. Data primer dalam penelitian ini diperoleh melalui proses *scraping* dari media sosial YouTube, sedangkan data sekunder berupa literatur dan dokumen terkait dengan pemilihan model untuk klasifikasi data. Penelitian ini menggunakan pendekatan kuantitatif untuk menganalisis data yang telah dikumpulkan. Data yang diperoleh kemudian diproses menggunakan *Jupyter Notebook* dengan bahasa pemrograman *Python*, menerapkan algoritma *Naive Bayes* dan *Support Vector Machine (SVM)* untuk klasifikasi. Metode ini digunakan untuk memperoleh hasil analisis yang relevan dalam menyelesaikan permasalahan penelitian.

#### **3.1.2 Alternatif pemecahan masalah**

Untuk mengevaluasi kinerja sistem klasifikasi, penelitian ini menggunakan metode *Confusion Matrix*. *Confusion Matrix* digunakan untuk mengevaluasi sejauh mana sistem mampu mengenali dengan baik kelas-kelas yang berbeda dari data yang telah dikumpulkan. Sistem ini diuji menggunakan data dari YouTube yang diproses dengan bahasa pemrograman *Python*. Data yang telah diolah kemudian dianalisis untuk mengidentifikasi opini masyarakat tentang Mohammad Ridwan Kamil di media sosial YouTube.

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

### 3.1.3 Metode Analisa

Dalam penelitian ini penulis menggunakan metode analisis yang mencakup analisis sentimen dan klasifikasi dengan memanfaatkan algoritma *Naive Bayes* dan *Support Vector Machine*. Metode tersebut bertujuan untuk mengelompokkan ulasan menjadi dua kategori sentimen, yaitu positif, negatif dan netral. Untuk memperkuat analisis ini, penulis menggunakan InSet (Indonesia Sentiment Lexicon), yaitu leksikon sentimen khusus untuk Bahasa Indonesia, sebagai panduan dalam proses pelabelan data. Leksikon ini memuat daftar kata-kata dengan muatan sentimen positif dan negatif yang banyak digunakan dalam Bahasa Indonesia. Melalui pendekatan ini, pelabelan dilakukan dengan mencocokkan kata-kata yang muncul dalam ulasan pengguna pada saluran YouTube Najwa Shihab dan Deddy Corbuzier Close the Door dengan kata-kata yang ada di dalam InSet. Setelah pelabelan selesai, penulis akan melanjutkan dengan visualisasi data menggunakan *library Matplotlib*, di mana data akan disajikan dalam bentuk *barchart* dan *pie chart*.

Setelah proses pelabelan dan visualisasi data selesai, langkah selanjutnya adalah melakukan klasifikasi menggunakan algoritma *Naive Bayes* dan metode *Support Vector Machine (SVM)*. Algoritma *naive bayes* adalah teknik klasifikasi berdasarkan *Teorema Bayes* dengan asumsi independensi antar predictor[35]. Model *Naive Bayes* mudah dibuat dan sangat berguna untuk kumpulan data yang sangat besar. Bersamaan dengan kesederhanaan, *Naive Bayes* dikenal mengungguli bahkan metode klasifikasi yang sangat canggih. Sementara itu, *Support Vector Machine (SVM)* yakni suatu model pembelajaran yang diawasi, memakai algoritma terkait untuk menganalisis data dalam konteks analisis klasifikasi dan regresi. SVM melakukan pelatihan dengan menyediakan dataset dan menentukan dua kategori berbeda di dalamnya[36].

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

Proses pelatihan SVM melibatkan pembuatan garis pemisah, yang dikenal sebagai *hyperplane* serta pembentukan margin untuk setiap label dengan menentukan vektor yang membatasi margin untuk kedua label. Proses klasifikasi lainnya dilakukan dengan membagi dataset menjadi dua bagian, yaitu data pelatihan (*training*) dan data pengujian (*testing*). Pembagian ini bertujuan untuk memastikan bahwa model yang dibangun dapat dievaluasi dengan akurat menggunakan data yang belum pernah dilihat sebelumnya. Dalam penelitian ini, penulis menggunakan lima nilai perbandingan untuk pembagian data *training* dan *testing*, yaitu:

1. Data training 90% dan data testing 10%
2. Data training 80% dan data testing 20%
3. Data training 70% dan data testing 30%
4. Data training 75% dan data testing 25%
5. Data training 60% dan data testing 40%

Setelah pembagian data, langkah berikutnya adalah mengubah data tersebut menjadi format vektor. Proses konversi ini dilakukan menggunakan *library CountVectorizer* yang ada dalam Bahasa pemrograman *Python*. *CountVectorizer* yaitu metode ini menghitung frekuensi kemunculan setiap kata dalam dokumen, menghasilkan representasi vektor yang menunjukkan jumlah kemunculan setiap kata[37]. Setelah data diubah menjadi bentuk vektor, penulis melanjutkan ke tahap pelatihan dengan menggunakan *library Confusion Matrix* untuk menganalisis hasil klasifikasi dari kedua algoritma yang diteliti. Proses ini bertujuan untuk mendapatkan hasil klasifikasi dan evaluasi dari *confusion matrix*, yang akan memberikan informasi penting tentang performa kedua metode. Hasil akhir dari klasifikasi dan evaluasi ini akan memungkinkan penulis untuk menyimpulkan serta memberikan rekomendasi berdasarkan perbandingan hasil yang diperoleh dari kedua metode yang telah digunakan.

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

### 3.2. Teknik Pemilihan Informasi (Populasi, sampel dan sampling)

Teknik pemilihan informasi adalah metode yang digunakan untuk memilih bagian-bagian dari sejumlah besar data, sehingga kita dapat memperoleh gambaran yang akurat tentang keseluruhan data tersebut.

#### a. Populasi

Populasi adalah keseluruhan kelompok, objek, individu, atau data yang menjadi subjek dalam suatu penelitian atau studi. Dalam konteks penelitian, populasi mencakup semua elemen atau unit yang memiliki karakteristik atau atribut tertentu yang ingin diteliti oleh peneliti.

#### 3.2.1 Teknik pengumpulan data

Pengumpulan data dilakukan untuk memperoleh informasi yang diperlukan penulis dalam menyelesaikan permasalahan penelitian. Dalam penelitian ini, penulis menggunakan beberapa metode pengumpulan data yang dirancang untuk memastikan kelengkapan dan ketepatan data yang diperoleh. Berikut metode pengumpulan data dalam penelitian ini:

##### a. Studi Pustaka

Studi pustaka (*library research*) yaitu metode dengan pengumpulan data dengan cara memahami dan mempelajari teori-teori dari berbagai literature yang berhubungan dengan penelitian tersebut[38]. Langkah ini bertujuan untuk memahami konsep, teori, dan temuan yang dapat mendasari serta memperkaya penelitian yang sedang berlangsung. Melalui studi pustaka, penulis dapat meninjau penelitian terdahulu, mengidentifikasi kesenjangan dalam pengetahuan yang ada, dan membangun kerangka konseptual yang kuat. Hal ini membantu penulis merumuskan masalah penelitian dengan jelas dan memastikan penelitian sesuai dengan analisis sentiment menggunakan algoritma *Naive Bayes* dan *Support Vector Machine*.

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

### b. Data Primer

Data yang digunakan dalam penelitian ini adalah data primer yang diambil dari ulasan pengguna pada saluran YouTube Najwa Shihab dan Deddy Corbuzier *Close the Door*. Pengambilan data dilakukan dengan menggunakan *library Python* untuk melakukan *scraping*, yang dilakukan melalui *Jupyter Notebook*. Data yang dikumpulkan mencakup atribut seperti *publishedAt*, *authorDisplayName*, *text*, dan *likeCount* dari komentar pengguna di YouTube. Total data yang terkumpul sebanyak 10.843 ribu data ulasan pengguna terkait Pilgub DKI 2024 pada kedua saluran YouTube ini, yang mencakup periode dari bulan Agustus hingga September 2024.

### 3.2.2 Teknik analisa data

#### a. Metode Pengujian

Pada penelitian ini, metode pengujian dilakukan dengan menggunakan *library Confusion Matrix* pada algoritma *Naïve Bayes* dan *Support Vector Machine*. *Confusion matrix* merupakan tabel yang menampilkan hasil dari proses klasifikasi dengan mengidentifikasi prediksi yang benar dan salah dalam setiap kategori kelas sentimen. Tabel 2 memperlihatkan struktur *confusion matrix* yang digunakan dalam penelitian ini:

**Tabel 3.1** *Confusion Matrix*

<i>Class/Aktual</i>	<i>Positif</i>	<i>Negatif</i>
<i>Positif</i>	<i>Positive (TP)</i>	<i>False Negative (FN)</i>
<i>Negatif</i>	<i>False Positive (FP)</i>	<i>True Negative (TN)</i>

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

Keterangan:

TP (*True Positive*) : Jumlah prediksi benar pada kelas sentimen positif.

TN (*True Negative*) : Jumlah prediksi benar pada kelas sentimen negatif.

FP (*False Positive*) : Jumlah prediksi salah pada kelas sentimen positif.

FN (*False Negative*) : Jumlah prediksi salah pada kelas sentimen negatif.

Berdasarkan hasil dari klasifikasi, diperoleh nilai *accuracy*, *precision*, dan *recall* sebagai berikut:

- 1) *Accuracy* mengukur sejauh mana model dapat mengklasifikasikan data dengan benar secara keseluruhan. Dengan kata lain, akurasi adalah nilai yang menunjukkan tingkat kedekatan antara nilai prediksi sistem dengan nilai prediksi manusia. Nilai akurasi dapat dicari dengan persamaan[35].

Berikut rumus *accuracy*:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

- 2) *Precision* adalah nilai sensitifitas atau nilai ketepatan sistem antara informasi yang diberikan oleh sistem untuk menunjukkan secara benar data kelas negatif atau kelas positif[35].

Berikut rumus *precision*:

$$precision = \frac{TP}{TP + FP} \quad (3)$$

- 3) *Recall* adalah nilai yang menunjukkan tingkat keberhasilan atau spesifisitas untuk mengetahui kembali sebuah informasi secara benar tentang data yang kelas negatif atau pun konten teks positif[35]. Berikut rumus *recall*:

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

$$\text{Recall} = \frac{TP}{TP + FN} \quad (4)$$



### b. Pengolahan Data

Penelitian ini menggunakan data yang diambil dari kanal YouTube Mata Najwa dan Deddy Corbuzier Close the Door, yang membahas Ridwan Kamil dalam konteks Pilgub DKI 2024. Data ini akan diproses melalui tahapan *preprocessing* dan *labeling*. Adapun tahap-tahap dalam proses ini adalah sebagai berikut:

#### 1) *Cleaning*

Proses identifikasi, koreksi, dan penghapusan ketidaksesuaian atau anomali dalam dataset untuk meningkatkan kualitas data, memastikan keakuratan analisis, dan memungkinkan visualisasi data yang informative[15]. Dalam konteks ini, Pembersihan kalimat dilakukan untuk menghilangkan informasi yang tidak relevan atau ambigu yang dapat mempengaruhi analisis sentiment[39]. Tabel 3 adalah contoh hasil dari proses *cleaning* data:

**Tabel 3.2** Contoh Hasil Proses *Cleaning* Data

Text	Hasil <i>Cleaning</i>
Saya warga jakafta tdk akan pilih RK	saya warga jakafta tidak akan pilih rk
sangat kurang yakin kalo dia mimpin jakarta..	warga jakarta anti orang jabarjakarta cocok di pimpin org jawa
Saya Jawa Barat aja... Geleng geleng kepala cara orang ini ngomong. Muak aku mendengarnya. Jabar aja amburadul dipimpin dari dia	saya jawa barat aja geleng geleng kepala cara orang ini ngomong muak aku mendengarnya jabar aja amburadul dipimpin dari dia

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

### 2) Case Folding

Case Folding, dilakukan untuk mengkonverensi teks dalam dokumen data menjadi lower case[40]. Proses ini bertujuan untuk menyamakan format penulisan, sehingga analisis data menjadi lebih konsisten dan efisien. Tabel 4 adalah contoh tabel hasil Case Folding:

**Tabel 3.3** Contoh Hasil Case Folding

Text	Hasil Case Folding
Kapan paslon yg independent diundang	kapan paslon yg independent diundang
Tetap Abah	tetap abah
Jak mania ketakutan gbk dn jis d bikin cabang gedung sate	jak mania ketakutan gbk dn jis d bikin cabang gedung sate

### 3) Tokenize

Tokenize, dilakukan untuk memecah teks menjadi per kata[40]. Langkah ini juga dapat digunakan untuk menghapus tanda baca, simbol, atau karakter non-alfabet yang tidak diperlukan dalam analisis. Tabel adalah contoh tabel hasil tokenize:

**Tabel 3.4** Contoh Hasil Tokenize

Text	Hasil Tokenize
sangat kurang yakin kalo dia mimpin jakarta	['sangat', 'kurang', 'yakin', 'kalo', 'dia', 'mimpin', 'jakarta']
Kapan paslon yg independent diundang	['kapan', 'paslon', 'yg', 'independent', 'diundang']
Kami sudah tinggal Jawabarat selama tahun masih belum mendapatkan air bersih seperti air PDAM atau air ledeng Jadi	['kami', 'sudah', 'tinggal', 'jawabarat', 'selama', '18', 'mendapatkan', 'air', 'bersih', 'air', 'pdam', 'atau', 'air', 'ledeng', 'jadi']

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

bak kamar mandi cepat kotor 'seperti', 'air', 'pdam', 'atau', 'air',  
 kuning baju ce... am dekil 'ledeng', 'jadi', 'bak', 'kamar',  
 air minum be... hari 'mandi', 'cepat', 'kotor', 'kuning',  
 sekali 'baju', 'cepat', 'kusam', 'dekil',  
 'air', 'minum', 'beli', 'setiap', '2',  
 'hari', 'sekali']

#### 4) Normalisasi

Pada tahap Normalisasi adalah proses mengubah dan mengoreksi singkatan menjadi kata yang lengkap dan memiliki arti yang sama menurut Kamus Besar Bahasa Indonesia (KBBI)[41]. Hal ini bertujuan agar informasi dapat diproses dengan lebih mudah. Tabel 6 adalah contoh tabel hasil normalisasi.

**Tabel 3.5** Contoh Hasil Normalisasi

Text	Normalisasi
Saya warga jakafta tdk akan pilih RK	saya warga jakarta tidak akan pilih rk
Mf kanng emil sy dan keluarga termasuk tetangga sy satu RT pilh pramono rano	maaf kang emil saya dan keluarga termasuk tetangga saya satu rt pilh pramono rano
Setelah nonton bang rano dan pasangannya kang emil dan pasangannya dan pak anis mana yg lebih seru	setelah menonton abang rano dan pasangannya kang emil dan pasangannya dan pak anis mana yang lebih seru

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

### 5) Stopword Removal

*Stopword removal* dilakukan penghilangan kata tidak begitu penting[40]. Proses ini bertujuan untuk menyederhanakan teks dan meningkatkan analisis, sehingga fokus dapat lebih diarahkan pada kata-kata yang memiliki makna signifikan dalam konteks analisis yang dilakukan. Tabel 7 adalah contoh tabel hasil *stopword removal*

**Tabel 3.6** Contoh Hasil *Stopword Removal*

Text	Hasil <i>Stopword Removal</i>
Kapan paslon yg independent diundang?	[paslon, independent, diundang]
Tetap Abah	[abah]
sangat kurang yakin kalo dia mimpin jakarta	[mimpin, jakarta]

### 6) Stemming

*Stemming* merupakan proses mengubah kata berimbuhan menjadi kata dasar[42]. Serta menghapus imbuhan pada setiap kata-kata yang kemudian menjadi kata inti, dan dalam proses ini pun bermaksud untuk menghilangkan suatu kalimat dari penyebutan kata yang kurang baik. Tabel 8 adalah contoh tabel hasil *stemming*.

**Tabel 3.7** Contoh Hasil *Stemming*

Text	Hasil <i>Stemming</i>
sangat kurang yakin kalo dia mimpin jakarta	mimpin jakarta
Saya warga jakarta tdk akan pilih RK	warga jakarta pilih rk
Kapan paslon yg independent diundang	paslon independent undang

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

### b. Labeling Data

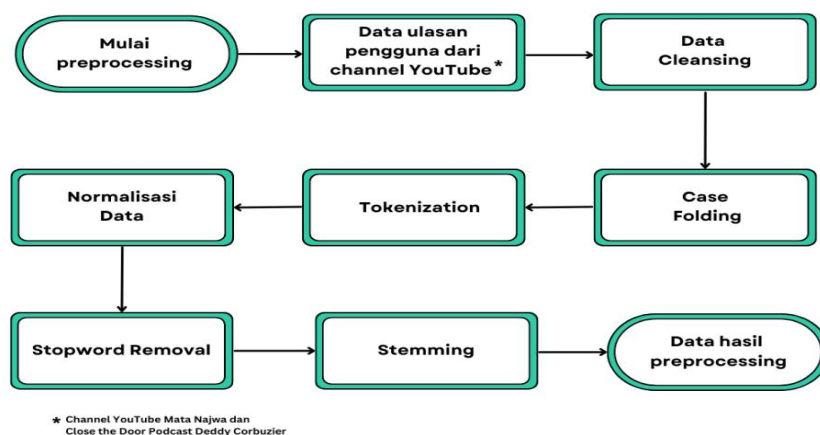
*Labeling* pada intinya adalah sebuah proses melabeli atau menandai sebuah informasi yang ada pada sebuah dataset untuk kemudian dipelajari mengenai informasi yang terkandung didalamnya[43]. Tahap pemberian label pada data dilakukan dengan menggunakan InSet (Indonesia Sentiment Lexicon), yaitu leksikon sentimen khusus untuk Bahasa Indonesia, sebagai panduan dalam proses pelabelan. Leksikon ini memuat daftar kata-kata dengan muatan sentimen positif dan negatif yang umum digunakan dalam Bahasa Indonesia.

Label yang dihasilkan digunakan untuk mengidentifikasi apakah setiap ulasan termasuk dalam kategori sentimen positif, negatif, atau netral. Seperti yang terlihat pada tabel 9 berikut:

**Tabel 3.8** Pemberian *Labeling* pada Data

Score	Hasil <i>Labeling</i>
(-1) – (-5)	Negatif
1 - 5	Positif
0	Netral

Langkah-langkah *preprocessing* yang telah dijelaskan dapat dilihat pada Gambar 3.1:



**Gambar 3.1** *Preprocessing* Data





**HASIL PENELITIAN DAN PEMBAHASAN**



**4.1 Gambaran Umum**

YouTube merupakan platform berbagi video yang sangat digemari di seluruh dunia, memungkinkan penggunanya untuk menonton, mengunggah, serta berbagi video secara gratis. Sejak diluncurkan pada tahun 2005, YouTube telah berkembang menjadi salah satu situs yang paling banyak digunakan, dengan jutaan video yang tersedia dalam berbagai kategori, seperti hiburan, edukasi, dan berita. Keberagaman jenis konten di YouTube membuatnya menjadi pusat informasi dan hiburan yang sangat dicari oleh pengguna di berbagai negara, mempermudah orang untuk mencari informasi tentang berbagai topik, termasuk peristiwa-peristiwa politik dan sosial terkini. Sekarang ini, terdapat banyak sekali media sosial yang populer di internet, seperti Facebook, Twitter, Instagram, dan tentu saja YouTube. Menurut laporan HootSuite Indonesia tahun 2021, YouTube menempati posisi pertama sebagai media sosial dengan jumlah pengguna terbanyak, mencapai 107 juta pengguna[44].

Media platform seperti YouTube menyediakan fitur komentar yang memungkinkan pengguna untuk memberikan pendapat mereka tentang video maupun siaran langsung yang diunggah. Melalui fitur ini, pengguna dapat berkomentar secara terbuka, seperti memberikan masukan, pujian, serta kritik terhadap video dan live streaming yang mereka tonton[45]. YouTube tidak hanya menjadi platform untuk menonton video, tetapi juga memungkinkan interaksi melalui komentar, tombol suka, berbagi, dan berlangganan saluran. Fitur-fitur ini membantu pembuat konten membangun audiens dan meningkatkan interaksi. Dengan sistem rekomendasi canggih, YouTube efektif membentuk opini publik dan persepsi masyarakat, termasuk dalam perhelatan politik seperti Pilkada.


## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

### 4.2 Hasil

#### 4.2.1 Pengumpulan data

Pengumpulan data dilakukan melalui proses scraping menggunakan API YouTube untuk mengekstrak komentar pengguna. Proses scraping ini dilakukan dengan menggunakan *Jupyter Notebook*. Data dikumpulkan dalam rentang waktu Agustus hingga September 2024 dari kanal YouTube Najwa Shihab dan Deddy Corbuzier, yang membahas pencalonan Mohammad Ridwan Kamil sebagai Gubernur DKI Jakarta, dengan total sebanyak 10.843 entri. Gambar 4.1 menampilkan hasil dari proses *scraping* data.



```
# Ubah Hasil Crawling ke Dataframe
df = pd.DataFrame(comments, columns=['publishedAt', 'authorDisplayName', 'textDisplay', 'likeCount'])
df
```

	publishedAt	authorDisplayName	textDisplay	likeCount
0	2024-09-29T14:14:52Z	@fahriahmad3543	Pertama sy tau RK tuh pas dia ada di iklan rok...	0
1	2024-09-29T13:52:16Z	@tjenchandra1082	<a href="https://www.youtube.com/watch?v=SifWI...	1
2	2024-09-29T13:43:28Z	@ade_burhanudin	Mobil curhat di bogor??? Ngadi2 wae ??	0
3	2024-09-29T13:31:39Z	@EduPlant_23	Tim editor ini gimana ko ada bagian yang diula...	0
4	2024-09-29T13:31:18Z	@hatakiwarehouse578	Bapaak, suka bangettt, luas banget wawasannya...	1
...	...	...	...	...

Gambar 4.1 Hasil *Scraping* data

#### 4.2.2 Preprocessing

*Preprocessing* adalah tahap awal dalam analisis data yang bertujuan untuk mempersiapkan data mentah agar siap digunakan dalam pemodelan lebih lanjut. Adapun langkah-langkah *preprocessing* yang dilakukan dalam penelitian ini adalah sebagai berikut.

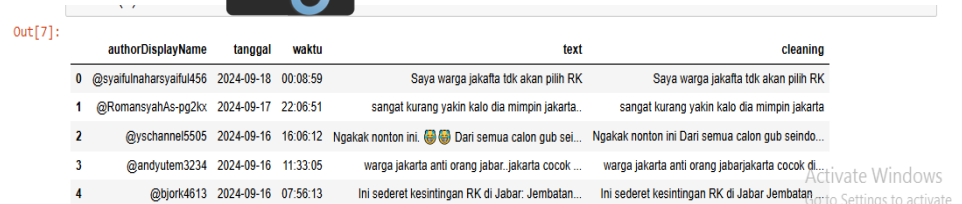
##### 1) *Cleaning*

*Cleaning* adalah tahap yang melibatkan penghilangan karakter non-alfabet untuk mengurangi gangguan dalam data. Karakter yang dihilangkan meliputi tanda baca seperti titik (.), koma

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

(.), tanda tanya (?), dan tanda seru (!), serta simbol-simbol seperti tanda '@' untuk nama pengguna, hashtag (#), dan emotikon. Gambar 4.2 menampilkan hasil proses *cleaning* data.



Out[7]:

	authorDisplayName	tanggal	waktu	text	cleaning
0	@syailunaharsyaiful456	2024-09-18	00:08:59	Saya warga jakarta tdk akan pilih RK	Saya warga jakarta tdk akan pilih RK
1	@RomansyahAs-pg2kx	2024-09-17	22:06:51	sangat kurang yakin kalo dia mimpin jakarta..	sangat kurang yakin kalo dia mimpin jakarta
2	@yschanel5505	2024-09-16	16:06:12	Ngakak nonton ini. 🤔 Dari semua calon gub sei...	Ngakak nonton ini Dari semua calon gub seindo...
3	@andyutem3234	2024-09-16	11:33:05	warga jakarta anti orang jabar. jakarta cocok ...	warga jakarta anti orang jabarjakarta cocok di...
4	@bjork4613	2024-09-16	07:56:13	Ini sederet kesingkatan RK di Jabar. Jembatan...	Ini sederet kesingkatan RK di Jabar. Jembatan...

Gambar 4.2 Hasil *cleaning* data

## 2) Casefolding

*Case folding* adalah proses mengubah teks menjadi huruf kecil (lowercase) dengan tujuan menghilangkan perbedaan yang disebabkan oleh variasi penulisan huruf besar dan kecil dalam dokumen teks. Gambar 4.3 menunjukkan hasil dari proses *case folding* tersebut.

Out[8]:

	authorDisplayName	tanggal	waktu	text	cleaning	case_folding
0	@syailunaharsyaiful456	2024-09-18	00:08:59	Saya warga jakarta tdk akan pilih RK	Saya warga jakarta tdk akan pilih RK	saya warga jakarta tdk akan pilih rk
1	@RomansyahAs-pg2kx	2024-09-17	22:06:51	sangat kurang yakin kalo dia mimpin jakarta..	sangat kurang yakin kalo dia mimpin jakarta	sangat kurang yakin kalo dia mimpin jakarta
2	@yschanel5505	2024-09-16	16:06:12	Ngakak nonton ini. 🤔 Dari semua calon gub sei...	Ngakak nonton ini Dari semua calon gub seindo...	ngakak nonton ini dari semua calon gub seindo...
3	@andyutem3234	2024-09-16	11:33:05	warga jakarta anti orang jabar. jakarta cocok ...	warga jakarta anti orang jabarjakarta cocok di...	warga jakarta anti orang jabarjakarta cocok di...
4	@bjork4613	2024-09-16	07:56:13	Ini sederet kesingkatan RK di Jabar. Jembatan...	Ini sederet kesingkatan RK di Jabar. Jembatan...	ini sederet kesingkatan rk di jabar jembatan ...

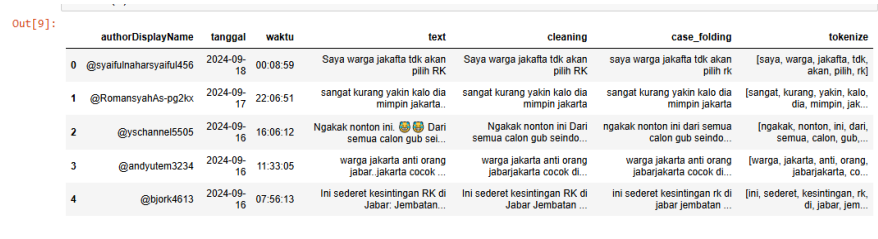
Gambar 4.3 Hasil *Case folding*

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

### 3) Tokenization

Tokenization adalah proses pemisahan kalimat menjadi setiap kata yang membentuknya, yang disebut term atau token. Tokenization dilakukan dengan memisahkan kata-kata berdasarkan spasi. Gambar 4.4 menunjukkan hasil dari proses tokenisasi.

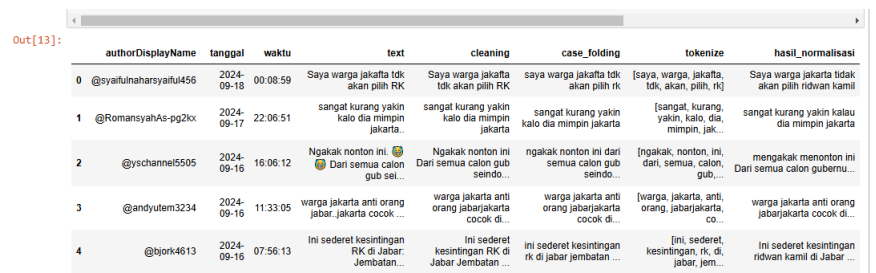


authorDisplayName	tanggal	waktu	text	cleaning	case_folding	tokenize
0 @syafuInaharsyafu456	2024-09-18	00:08:59	Saya warga jakarta tdk akan pilih RK	Saya warga jakarta tdk akan pilih RK	saya warga jakarta tdk akan pilih rk	[saya, warga, jakarta, tdk, akan, pilih, rk]
1 @RomansyahAs-pg2kx	2024-09-17	22:06:51	sangat kurang yakin kalo dia mimpin jakarta.	sangat kurang yakin kalo dia mimpin jakarta	sangat kurang yakin kalo dia mimpin jakarta	[sangat, kurang, yakin, kalo, dia, mimpin, jak...
2 @yschannel5505	2024-09-16	16:06:12	Ngakak nonton ini. 🤔 Dari semua calon gub sel...	Ngakak nonton ini Dari semua calon gub seindo...	ngakak nonton ini dari semua calon gub seindo...	[ngakak, nonton, ini, dari, semua, calon, gub, ...
3 @andyulem3234	2024-09-16	11:33:05	warga jakarta anti orang jabar, jakarta cocok ...	warga jakarta anti orang jabarjakarta cocok di...	warga jakarta anti orang jabarjakarta cocok di...	[warga, jakarta, anti, orang, jabarjakarta, co...
4 @bjork4613	2024-09-16	07:56:13	Ini sederet kesintingan RK di Jabar. Jembatan...	Ini sederet kesintingan RK di Jabar. Jembatan ...	ini sederet kesintingan rk di jabar jembatan ...	[ini, sederet, kesintingan, rk, di, jabar, jem...

Gambar 4.4 Hasil Tokenization

### 4) Normalisasi

Normalisasi adalah proses untuk membuat teks lebih seragam dan konsisten. Tujuannya adalah untuk mengurangi variasi kata yang tidak diperlukan, seperti mengubah bentuk kata yang berbeda tetapi memiliki arti yang sama. Normalisasi juga mencakup penghapusan karakter yang tidak relevan atau tidak penting untuk analisis. Proses ini bertujuan untuk mempermudah analisis teks selanjutnya. Gambar 4.5 menunjukkan hasil dari proses normalisasi.



authorDisplayName	tanggal	waktu	text	cleaning	case_folding	tokenize	hasil_normalisasi
0 @syafuInaharsyafu456	2024-09-18	00:08:59	Saya warga jakarta tdk akan pilih RK	Saya warga jakarta tdk akan pilih RK	saya warga jakarta tdk akan pilih rk	[saya, warga, jakarta, tdk, akan, pilih, rk]	Saya warga jakarta tidak akan pilih rdwan kami
1 @RomansyahAs-pg2kx	2024-09-17	22:06:51	sangat kurang yakin kalo dia mimpin jakarta.	sangat kurang yakin kalo dia mimpin jakarta	sangat kurang yakin kalo dia mimpin jakarta	[sangat, kurang, yakin, kalo, dia, mimpin, jak...	sangat kurang yakin kalau dia mimpin jakarta
2 @yschannel5505	2024-09-16	16:06:12	Ngakak nonton ini. 🤔 Dari semua calon gub sel...	Ngakak nonton ini Dari semua calon gub seindo...	ngakak nonton ini dari semua calon gub seindo...	[ngakak, nonton, ini, dari, semua, calon, gub, ...	mengakak menonton ini Dari semua calon gubernu...
3 @andyulem3234	2024-09-16	11:33:05	warga jakarta anti orang jabar, jakarta cocok ...	warga jakarta anti orang jabarjakarta cocok di...	warga jakarta anti orang jabarjakarta cocok di...	[warga, jakarta, anti, orang, jabarjakarta, co...	warga jakarta anti orang jabarjakarta cocok di...
4 @bjork4613	2024-09-16	07:56:13	Ini sederet kesintingan RK di Jabar. Jembatan...	Ini sederet kesintingan RK di Jabar. Jembatan ...	ini sederet kesintingan rk di jabar jembatan ...	[ini, sederet, kesintingan, rk, di, jabar, jem...	Ini sederet kesintingan rdwan kami di Jabar ...

Gambar 4.5 Hasil Normalisasi

# Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

## 5) Stopword removal

Stopword removal adalah proses menghapus kata-kata umum yang tidak memberikan informasi signifikan dalam analisis teks, seperti 'dan', 'atau', 'yang', 'di', dan kata-kata lainnya. Meskipun sering muncul dalam teks, kata-kata ini tidak membantu dalam membedakan makna atau konteks. Gambar 4.6 menunjukkan hasil dari proses *Stopword removal*.

Out[15]:

	authorDisplayName	tanggal	waktu	text	cleaning	case_folding	tokenize	hasil_normalisasi	stopword removal
0	@syairfuhaharsyairfu456	2024-09-18	00:08:59	Saya warga jakarta tdk akan pilih RK	Saya warga jakarta tdk akan pilih RK	saya warga jakarta tdk akan pilih rk	[saya, warga, jakarta, tdk, akan, pilih, rk]	Saya warga jakarta tdk akan pilih rdwan kami	[Saya, warga, jakarta, pilih, rdwan, kami]
1	@RomansyahAs-pg2kx	2024-09-17	22:06:51	sangat kurang yakin kalo dia mimpin jakarta.	sangat kurang yakin kalo dia mimpin jakarta	sangat kurang yakin kalo dia mimpin jakarta	[sangat, kurang, yakin, kalo, dia, mimpin, jak...	sangat kurang yakin jabarjarkarta	[mimpin, jakarta]
2	@yschanne5505	2024-09-16	16:06:12	Ngakak nonton ini. @ Dari semua calon gub sei...	Ngakak nonton ini Dari semua calon gub seindo...	ngakak nonton ini dari semua calon gub seindo...	[ngakak, nonton, ini, dari, semua, calon, gub...]	mengakak menonton ini Dari semua calon gubernu...	[mengakak, menonton, Dari, calon, gubernu, se...]
3	@andytem3234	2024-09-16	11:33:05	warga jakarta anti orang jabar. jakarta cocok di...	warga jakarta anti orang jabarjarkarta cocok di...	warga jakarta anti orang jabarjarkarta cocok di...	[warga, jakarta, anti, orang, jabarjarkarta, co...]	warga jakarta anti orang jabarjarkarta cocok di...	[warga, jakarta, anti, orang, jabarjarkarta, co...]
4	@bjork4613	2024-09-16	07:56:13	Ini sederet kesintingan RK di Jabar. Jembatan...	Ini sederet kesintingan RK di Jabar. Jembatan ...	ini sederet kesintingan rk di jabar jembatan ...	[ini, sederet, kesintingan, rk, di, jabar, jem...]	Ini sederet kesintingan rdwan kami di Jabar ...	[ini, sederet, kesintingan, rdwan, kami, jab...]

Gambar 4.6 Hasil *Stopword removal*

## 6) Stemming

Stemming adalah proses mengubah kata-kata menjadi bentuk dasar dengan menghilangkan kata sambung, kata ganti, dan kata depan. Proses ini dilakukan dengan menghapus awalan atau akhiran dari kata tersebut. *Stemming* dilakukan menggunakan library Sastrawi dalam bahasa pemrograman *Python*. Gambar 4.7 menunjukkan hasil dari proses *Stemming*.

Out[18]:

	authorDisplayName	tanggal	waktu	text	cleaning	case_folding	tokenize	hasil_normalisasi	stopword removal	stemming_data
0	@syairfuhaharsyairfu456	2024-09-18	00:08:59	Saya warga jakarta tdk akan pilih RK	Saya warga jakarta tdk akan pilih RK	saya warga jakarta tdk akan pilih rk	[saya, warga, jakarta, tdk, akan, pilih, rk]	Saya warga jakarta tdk akan pilih rdwan kami	[Saya, warga, jakarta, pilih, rdwan, kami]	saya warga jakarta pilih rdwan kami
1	@RomansyahAs-pg2kx	2024-09-17	22:06:51	sangat kurang yakin kalo dia mimpin jakarta.	sangat kurang yakin kalo dia mimpin jakarta	sangat kurang yakin kalo dia mimpin jakarta	[sangat, kurang, yakin, kalo, dia, mimpin, jak...	sangat kurang yakin jabarjarkarta	[mimpin, jakarta]	mimpin jakarta
2	@yschanne5505	2024-09-16	16:06:12	Ngakak nonton ini. @ Dari semua calon gub sei...	Ngakak nonton ini Dari semua calon gub seindo...	ngakak nonton ini dari semua calon gub seindo...	[ngakak, nonton, ini, dari, semua, calon, gub...]	mengakak menonton ini Dari semua calon gubernu...	[mengakak, menonton, Dari, calon, gubernu, se...]	akak nonton dari calon gubernu seindo kykny f...
3	@andytem3234	2024-09-16	11:33:05	warga jakarta anti orang jabar. jakarta cocok di...	warga jakarta anti orang jabarjarkarta cocok di...	warga jakarta anti orang jabarjarkarta cocok di...	[warga, jakarta, anti, orang, jabarjarkarta, co...]	warga jakarta anti orang jabarjarkarta cocok di...	[warga, jakarta, anti, orang, jabarjarkarta, co...]	warga jakarta anti orang jabarjarkarta cocok pi...
4	@bjork4613	2024-09-16	07:56:13	Ini sederet kesintingan RK di Jabar. Jembatan...	Ini sederet kesintingan RK di Jabar. Jembatan ...	ini sederet kesintingan rk di jabar jembatan ...	[ini, sederet, kesintingan, rk, di, jabar, jem...]	Ini sederet kesintingan rdwan kami di Jabar ...	[ini, sederet, kesintingan, rdwan, kami, jab...]	ini deret sinting rdwan kami jabar jembatan ...

Gambar 4.7 Hasil *Stemming*

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

### 4.2.3 Labelling

*Labelling* merupakan proses pemberian label sentimen pada data teks berdasarkan analisis kata-kata yang ada. Dalam penelitian ini, *labelling* dilakukan menggunakan Inset (Indonesian Sentiment Lexicon), yang merupakan kamus sentimen bahasa Indonesia yang membantu dalam mengidentifikasi dan memberi label pada teks berdasarkan nilai sentimennya, apakah positif, negatif, atau netral. Gambar 4.8 menunjukkan hasil dari proses *Labelling*.



```
Out[26]:
```

	authorDisplayName	tanggal	waktu	hasil_normalisasi	Score	Sentiment
0	@syaifulnaharsyaiful456	2024-09-18	00:08:59	Saya warga jakarta tidak akan pilih ridwan kamil	-2	Negatif
1	@RomansyahAs-pg2lox	2024-09-17	22:06:51	sangat kurang yakin kalau dia mimpin jakarta	-2	Negatif
2	@yschannel5505	2024-09-16	16:06:12	mengakak menonton ini Dari semua calon gubernu...	-6	Negatif
3	@andyutem3234	2024-09-16	11:33:05	warga jakarta anti orang jabarjakarta cocok di...	-1	Negatif
4	@bjork4613	2024-09-16	07:56:13	Ini sederet kesintingan ridwan kamil di Jabar ...	-16	Negatif

**Gambar 4.8** Hasil *Labelling*

### 4.2.4 Visualisasi

Visualisasi adalah proses penyajian data dalam bentuk grafik atau diagram untuk memudahkan pemahaman dan analisis. Dalam analisis teks, visualisasi digunakan untuk menggambarkan hasil dari proses seperti distribusi sentimen, frekuensi kata, atau hubungan antar variabel. Teknik visualisasi yang umum digunakan termasuk wordcloud, bar chart, dan pie chart, yang dapat membantu menampilkan pola atau tren yang ada dalam data. Gambar 4.9, 4.10, 4.11, 4.12 dibawah ini menunjukkan hasil dari proses visualisasi.

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)



Gambar 4.9 Wordcloud sentiment negatif

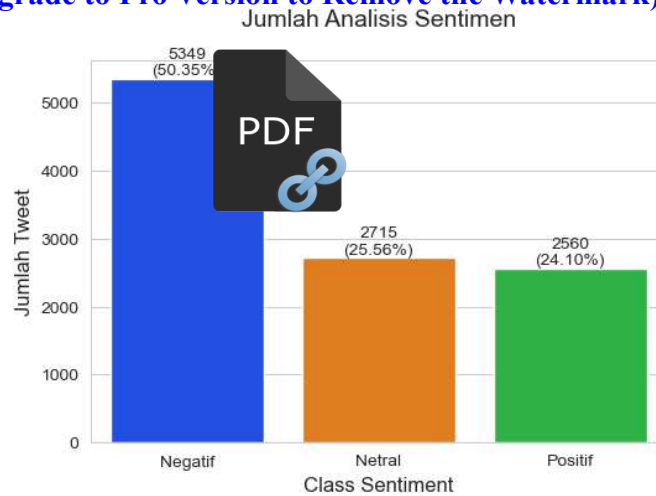


Gambar 4.10 Wordcloud sentiment positif

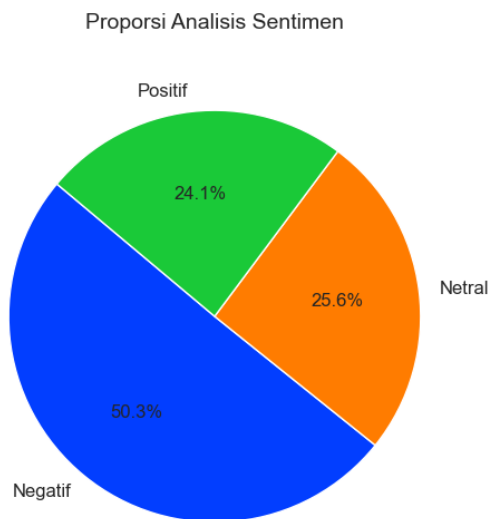
Berdasarkan hasil visualisasi WordCloud, sentimen negatif lebih fokus pada kekecewaan dan ketidakpuasan terhadap Ridwan Kamil, dengan kata-kata seperti "tidak", "buruk", "kalah", "masalah". Sebaliknya, sentimen positif menonjolkan dukungan dan apresiasi, dengan kata-kata seperti "ridwan kamil", "jakarta", "baik", "bagus", "mantap".

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)



**Gambar 4.11** Visualisasi *barchart*



**Gambar 4.12** Visualisasi *pie chart*


Berdasarkan hasil visualisasi Bar Chart dan Pie Chart, distribusi sentimen masyarakat terhadap Mochammad Ridwan Kamil menunjukkan bahwa sentimen negatif mendominasi dengan persentase 50,35%. Sementara itu, sentimen netral mencakup 25,56% dan sentimen positif hanya mencapai 24,10%. Ini menunjukkan bahwa mayoritas opini masyarakat cenderung negatif, mencerminkan ketidakpuasan terhadap calon tersebut.

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

### 4.2.5 Pengujian hasil analisa

#### a. Klasifikasi dan evaluasi



Setelah dilakukan *labelling* dan visualisasi data, langkah selanjutnya adalah klasifikasi dan evaluasi. Data yang telah diberi label kemudian dibagi menjadi dua bagian, yaitu data training dan data testing, untuk mempersiapkan model pembelajaran. Proses klasifikasi dan evaluasi dalam penelitian ini dilakukan menggunakan metode *Naive Bayes* dan *Support Vector Machine (SVM)*.

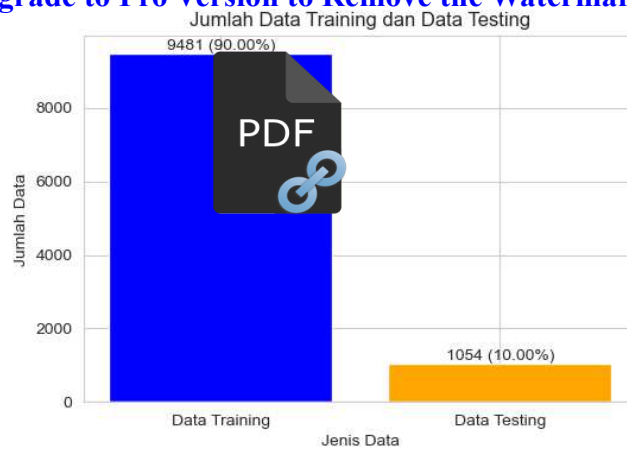
#### b. Pembagian Data *Training* dan Data *Testing*

Langkah pertama dalam proses klasifikasi menggunakan metode *Naive Bayes* dan *Support Vector Machine (SVM)* adalah membagi dataset menjadi dua bagian utama, yaitu data *training* dan data *testing*. Pembagian ini bertujuan untuk memastikan bahwa model dapat dilatih menggunakan data yang cukup dan kemudian diuji menggunakan data yang tidak digunakan saat pelatihan, sehingga dapat dievaluasi akurasi dan performa model secara efektif. Berikut adalah perbandingan jumlah pembagian data *training* dan data *testing*:

##### 1) Data training 90% dan data testing 10%

Dengan perbandingan 90% untuk data *training* dan 10% untuk data *testing* dari total 10.535 data, sebanyak 9.481 data digunakan untuk melatih model (data *training*), dan 1.054 data digunakan untuk menguji performa model (data *testing*). Pembagian data ini dapat dilihat pada gambar berikut ini.

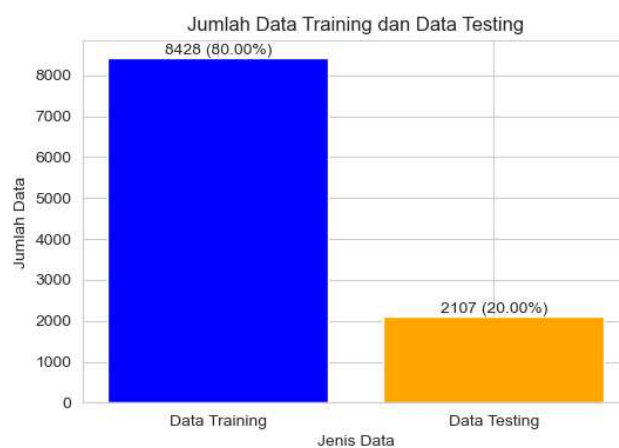
**Protected by PDF Anti-Copy Free**  
 (Upgrade to Pro Version to Remove the Watermark)



**Gambar 4.13** Data training 90% dan testing 10%

2) Data training 80% dan data testing 20%

Dengan perbandingan 80% untuk data *training* dan 20% untuk data *testing* dari total 10535 data, sebanyak 8.428 data digunakan untuk melatih model (data *training*), dan 2.107 data digunakan untuk menguji performa model (data *testing*). Pembagian data ini dapat dilihat pada gambar berikut ini.

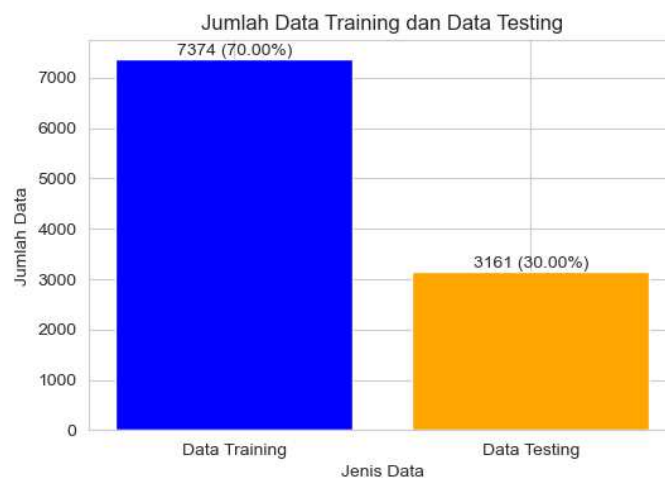


**Gambar 4.14** Data training 80% dan testing 20%

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)  
3) Data training 70% dan data testing 30%

Dengan perbandingan 70% untuk data *training* dan 30% untuk data *testing* dari total 10.535 data, sebanyak 7.374 data digunakan untuk melatih model (data *training*), dan 3.161 data digunakan untuk menguji performa model (data *testing*). Pembagian data ini dapat dilihat pada gambar berikut ini.



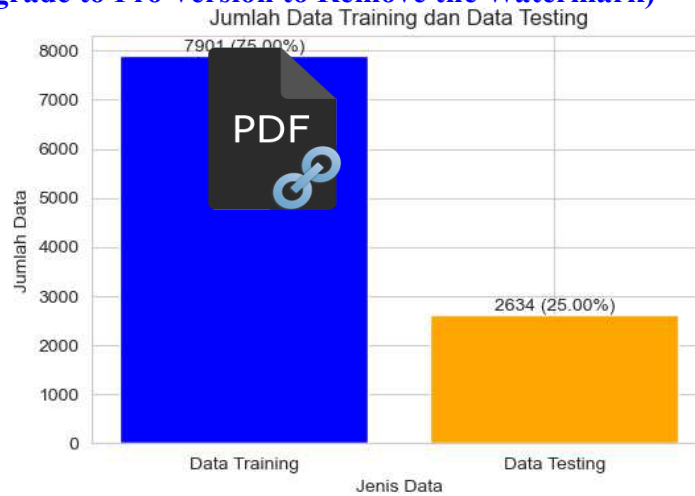
**Gambar 4.15** Data training 70% dan testing 30%

4) Data training 75% dan data testing 25%

Dengan perbandingan 75% untuk data *training* dan 25% untuk data *testing* dari total 10.535 data, sebanyak 7.901 data digunakan untuk melatih model (data *training*), dan 2.634 data digunakan untuk menguji performa model (data *testing*). Pembagian data ini dapat dilihat pada gambar berikut ini.

## Protected by PDF Anti-Copy Free

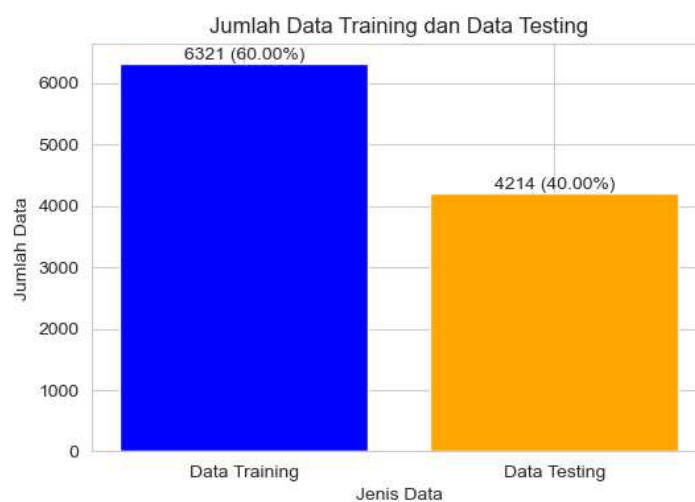
(Upgrade to Pro Version to Remove the Watermark)



**Gambar 4.16** Data training 75% dan testing 25%

5) Data training 60% dan data testing 40%

Dengan perbandingan 60% untuk data *training* dan 40% untuk data *testing* dari total 10.535 data, sebanyak 6.321 data digunakan untuk melatih model (data *training*), dan 4.214 data digunakan untuk menguji performa model (data *testing*). Pembagian data ini dapat dilihat pada gambar berikut ini.



**Gambar 4.17** Data training 60% dan testing 40%

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

### c. Klasifikasi dan Evaluasi Menggunakan Metode *Naive Bayes*

Metode *Naive Bayes* melakukan proses pelatihan dengan cara menghitung probabilitas muncul setiap fitur pada masing-masing kelas berdasarkan data training. Selama proses pelatihan, algoritma ini menggunakan asumsi independensi antar fitur untuk memprediksi kelas dari data baru. Setelah model dilatih, proses evaluasi dilakukan menggunakan data testing. Evaluasi ini memanfaatkan confusion matrix untuk menghitung metrik performa seperti *accuracy*, *precision*, *recall*, dan *F1-score*. Hasil dari proses klasifikasi pada data training dan data testing menggunakan metode *Naive Bayes* dapat dilihat pada gambar berikut.

#### 1) Data Training

Gaussian Naive Bayes Accuracy: 0.9449715370018975					Multinomial Naive Bayes Accuracy: 0.8795066413662239				
Gaussian Naive Bayes Confusion Matrix:					Multinomial Naive Bayes Confusion Matrix:				
[[523 22 11]					[[547 6 3]				
[ 0 256 0]					[ 76 171 9]				
[ 0 25 217]]					[ 32 1 209]]				
Gaussian Naive Bayes Classification Report:					Multinomial Naive Bayes Classification Report:				
	precision	recall	f1-score	support		precision	recall	f1-score	support
Negatif	1.00	0.94	0.97	556	Negatif	0.84	0.98	0.90	556
Netral	0.84	1.00	0.92	256	Netral	0.96	0.67	0.79	256
Positif	0.95	0.90	0.92	242	Positif	0.95	0.86	0.90	242
accuracy			0.94	1054	accuracy			0.88	1054
macro avg	0.93	0.95	0.94	1054	macro avg	0.91	0.84	0.86	1054
weighted avg	0.95	0.94	0.95	1054	weighted avg	0.89	0.88	0.88	1054
Bernoulli Naive Bayes Accuracy: 0.8026565464895635									
Bernoulli Naive Bayes Confusion Matrix:									
[[541 15 0]									
[ 52 204 0]									
[ 84 57 101]]									
Bernoulli Naive Bayes Classification Report:									
	precision	recall	f1-score	support					
Negatif	0.80	0.97	0.88	556					
Netral	0.74	0.80	0.77	256					
Positif	1.00	0.42	0.59	242					
accuracy			0.80	1054					
macro avg	0.85	0.73	0.74	1054					
weighted avg	0.83	0.80	0.78	1054					

**Gambar 4.18** Hasil klasifikasi *Naive Bayes* dengan perbandingan 90:10

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

Gaussian Naive Bayes Accuracy: 0.8932130991931656

Gaussian Naive Bayes Confusion Matrix:  
 [[968 84 41]  
 [ 0 508 0]  
 [ 0 100 406]]

Gaussian Naive Bayes Classification Report:  

	precision	recall	f1-score	support
Negatif	1.00	0.86	0.92	1093
Netral	0.73	1.00	0.85	508
Positif	0.91	0.80	0.85	506
accuracy			0.89	2107
macro avg	0.88	0.90	0.88	2107
weighted avg	0.91	0.89	0.90	2107

Multinomial Naive Bayes Accuracy: 0.8448030374940674

Multinomial Naive Bayes Confusion Matrix:  
 [[1072 13 8]  
 [ 183 295 30]  
 [ 88 5 413]]

Multinomial Naive Bayes Classification Report:  

	precision	recall	f1-score	support
Negatif	0.80	0.98	0.88	1093
Netral	0.94	0.58	0.72	508
Positif	0.92	0.82	0.86	506
accuracy			0.84	2107
macro avg	0.89	0.79	0.82	2107
weighted avg	0.86	0.84	0.84	2107

Bernoulli Naive Bayes Accuracy: 0.8168011390602753

Bernoulli Naive Bayes Confusion Matrix:  
 [[989 104 0]  
 [ 84 420 4]  
 [ 79 115 312]]

Bernoulli Naive Bayes Classification Report:  

	precision	recall	f1-score	support
Negatif	0.86	0.90	0.88	1093
Netral	0.66	0.83	0.73	508
Positif	0.99	0.62	0.76	506
accuracy			0.82	2107
macro avg	0.83	0.78	0.79	2107
weighted avg	0.84	0.82	0.82	2107

**Gambar 4.19** Hasil klasifikasi *Naïve Bayes* dengan perbandingan 80:20

Gaussian Naive Bayes Accuracy: 0.8636507434356216

Gaussian Naive Bayes Confusion Matrix:  
 [[1395 179 57]  
 [ 0 769 0]  
 [ 0 195 566]]

Gaussian Naive Bayes Classification Report:  

	precision	recall	f1-score	support
Negatif	1.00	0.86	0.92	1631
Netral	0.67	1.00	0.80	769
Positif	0.91	0.74	0.82	761
accuracy			0.86	3161
macro avg	0.86	0.87	0.85	3161
weighted avg	0.90	0.86	0.87	3161

Multinomial Naive Bayes Accuracy: 0.8250553622271433

Multinomial Naive Bayes Confusion Matrix:  
 [[1592 30 9]  
 [ 315 401 53]  
 [ 139 7 615]]

Multinomial Naive Bayes Classification Report:  

	precision	recall	f1-score	support
Negatif	0.78	0.98	0.87	1631
Netral	0.92	0.52	0.66	769
Positif	0.91	0.81	0.86	761
accuracy			0.83	3161
macro avg	0.87	0.77	0.80	3161
weighted avg	0.84	0.83	0.81	3161

Bernoulli Naive Bayes Accuracy: 0.7889908256880734

Bernoulli Naive Bayes Confusion Matrix:  
 [[1380 249 2]  
 [ 113 648 8]  
 [ 95 200 466]]

Bernoulli Naive Bayes Classification Report:  

	precision	recall	f1-score	support
Negatif	0.87	0.85	0.86	1631
Netral	0.59	0.84	0.69	769
Positif	0.98	0.61	0.75	761
accuracy			0.79	3161
macro avg	0.81	0.77	0.77	3161
weighted avg	0.83	0.79	0.79	3161

**Gambar 4.20** Hasil klasifikasi *Naïve Bayes* dengan perbandingan 70:30

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

Gaussian Naive Bayes Accuracy: 0.8766135155656796					Multinomial Naive Bayes Accuracy: 0.834092634776006				
Gaussian Naive Bayes Confusion Matrix:					Multinomial Naive Bayes Confusion Matrix:				
[[1189 131 46] [ 0 639 0] [ 0 148 481]]					[[1340 19 7] [ 257 345 37] [ 110 7 512]]				
Gaussian Naive Bayes Classification Report:					Multinomial Naive Bayes Classification Report:				
	precision	recall	f1-score	support		precision	recall	f1-score	support
Negatif	1.00	0.87	0.93	1366	Negatif	0.79	0.98	0.87	1366
Netral	0.70	1.00	0.82	639	Netral	0.93	0.54	0.68	639
Positif	0.91	0.76	0.83	629	Positif	0.92	0.81	0.86	629
accuracy			0.88	2634	accuracy			0.83	2634
macro avg	0.87	0.88	0.86	2634	macro avg	0.88	0.78	0.81	2634
weighted avg	0.91	0.88	0.88	2634	weighted avg	0.85	0.83	0.82	2634

Bernoulli Naive Bayes Accuracy: 0.7927107061503417				
Bernoulli Naive Bayes Confusion Matrix:				
[[1184 182 0] [ 104 530 5] [ 86 169 374]]				
Bernoulli Naive Bayes Classification Report:				
	precision	recall	f1-score	support
Negatif	0.86	0.87	0.86	1366
Netral	0.60	0.83	0.70	639
Positif	0.99	0.59	0.74	629
accuracy			0.79	2634
macro avg	0.82	0.76	0.77	2634
weighted avg	0.83	0.79	0.79	2634

Gambar 4.21 Hasil klasifikasi *Naive Bayes* dengan perbandingan 75:25

Gaussian Naive Bayes Accuracy: 0.8433792121499762					Multinomial Naive Bayes Accuracy: 0.8189368770764119				
Gaussian Naive Bayes Confusion Matrix:					Multinomial Naive Bayes Confusion Matrix:				
[[1784 289 77] [ 0 1051 0] [ 0 294 719]]					[[2094 41 15] [ 420 558 73] [ 202 12 799]]				
Gaussian Naive Bayes Classification Report:					Multinomial Naive Bayes Classification Report:				
	precision	recall	f1-score	support		precision	recall	f1-score	support
Negatif	1.00	0.83	0.91	2150	Negatif	0.77	0.97	0.86	2150
Netral	0.64	1.00	0.78	1051	Netral	0.91	0.53	0.67	1051
Positif	0.90	0.71	0.79	1013	Positif	0.90	0.79	0.84	1013
accuracy			0.84	4214	accuracy			0.82	4214
macro avg	0.85	0.85	0.83	4214	macro avg	0.86	0.76	0.79	4214
weighted avg	0.89	0.84	0.85	4214	weighted avg	0.84	0.82	0.81	4214

Bernoulli Naive Bayes Accuracy: 0.760797342192691				
Bernoulli Naive Bayes Confusion Matrix:				
[[1710 427 13] [ 139 899 13] [ 106 310 597]]				
Bernoulli Naive Bayes Classification Report:				
	precision	recall	f1-score	support
Negatif	0.87	0.80	0.83	2150
Netral	0.55	0.86	0.67	1051
Positif	0.96	0.59	0.73	1013
accuracy			0.76	4214
macro avg	0.79	0.75	0.74	4214
weighted avg	0.81	0.76	0.77	4214

Gambar 4.22 Hasil klasifikasi *Naive Bayes* dengan perbandingan 60:40

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

Dari gambar di atas, diketahui bahwa hasil pengujian klasifikasi *Naive Bayes* dengan perbandingan data training 90% dan data testing 10% menunjukkan akurasi tertinggi dicapai oleh algoritma Gaussian sebesar 90,26%, diikuti oleh Multinomial sebesar 88%, dan Bernoulli sebesar 80,26%. Sementara itu, hasil pengujian klasifikasi *Naive Bayes* dengan perbandingan data training 80% dan data testing 20% menunjukkan akurasi tertinggi pada algoritma Gaussian sebesar 89,32%, diikuti oleh Multinomial sebesar 84,48%, dan Bernoulli sebesar 82%.

### d. Klasifikasi dan Evaluasi Metode *Support Vector Machine (SVM)*

Metode *Support Vector Machine (SVM)* melakukan pelatihan dengan cara mendeteksi pola atau karakteristik data pada tiap kelas yang terbentuk dari data training. Setelah proses pelatihan selesai, model dievaluasi menggunakan data testing. Evaluasi dilakukan dengan memanfaatkan confusion matrix untuk menghitung metrik performa, seperti *accuracy*, *precision*, *recall*, dan *F1-score*. Adapun hasil klasifikasi pada data training dan data testing menggunakan metode SVM dapat dilihat pada gambar di bawah ini.

#### 1) Data Training

```

Confusion Matrix:
[[519  36  1]
 [ 61 163 32]
 [ 19  34 189]]
=====

Accuracy: 0.83
Accuracy: 82.64%
=====

              precision    recall  f1-score   support

Negatif         0.87         0.93         0.90         556
Netral           0.70         0.64         0.67         256
Positif          0.85         0.78         0.81         242

 accuracy          0.83         0.83         0.83        1054
 macro avg         0.81         0.78         0.79        1054
 weighted avg      0.82         0.83         0.82        1054

```

**Gambar 4.23** Hasil klasifikasi metode SVM untuk perbandingan 90:10


## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

```

Confusion Matrix:
[[1004  79  10]
 [ 128 309  71]
 [  38  67 401]]
=====
Accuracy: 0.81
Accuracy: 81.35%
=====

```



	precision	recall	f1-score	support
Negatif	0.86	0.92	0.89	1093
Netral	0.68	0.61	0.64	508
Positif	0.83	0.79	0.81	506
accuracy			0.81	2107
macro avg	0.79	0.77	0.78	2107
weighted avg	0.81	0.81	0.81	2107

**Gambar 4.24** Hasil klasifikasi metode SVM untuk perbandingan 80:20

```

Confusion Matrix:
[[1502 105  24]
 [ 204 463 102]
 [  58 104 599]]
=====
Accuracy: 0.81
Accuracy: 81.11%
=====

```

	precision	recall	f1-score	support
Negatif	0.85	0.92	0.88	1631
Netral	0.69	0.60	0.64	769
Positif	0.83	0.79	0.81	761
accuracy			0.81	3161
macro avg	0.79	0.77	0.78	3161
weighted avg	0.81	0.81	0.81	3161

**Gambar 4.25** Hasil klasifikasi metode SVM untuk perbandingan 70:30

```

Confusion Matrix:
[[1256  97  13]
 [ 166 387  86]
 [  50  90 489]]
=====
Accuracy: 0.81
Accuracy: 80.94%
=====

```

	precision	recall	f1-score	support
Negatif	0.85	0.92	0.89	1366
Netral	0.67	0.61	0.64	639
Positif	0.83	0.78	0.80	629
accuracy			0.81	2634
macro avg	0.79	0.77	0.78	2634
weighted avg	0.80	0.81	0.81	2634

**Gambar 4.26** Hasil klasifikasi metode SVM untuk perbandingan 75:25

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

```
Confusion Matrix:
[[1965 155 30]
 [ 284 619 148]
 [ 88 147 778]]
```

```
=====
Accuracy: 0.80
Accuracy: 79.78%
=====
```



	precision	recall	f1-score	support
Negatif	0.84	0.91	0.88	2150
Netral	0.67	0.59	0.63	1051
Positif	0.81	0.77	0.79	1013
accuracy			0.80	4214
macro avg	0.78	0.76	0.76	4214
weighted avg	0.79	0.80	0.79	4214

**Gambar 4.27** Hasil klasifikasi metode SVM untuk perbandingan 60:40

Dari gambar di atas, diketahui bahwa hasil pengujian klasifikasi menggunakan metode *Support Vector Machine (SVM)* dengan perbandingan data training 90% dan data testing 10% menghasilkan akurasi tertinggi sebesar 83%. Sementara itu, pada pengujian dengan perbandingan data training 80% dan data testing 20%, akurasi tertinggi yang diperoleh adalah 81,35%.

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

### 4.3 Pembahasan

Penelitian ini bertujuan untuk menganalisis elektabilitas Mochammad Ridwan Falaq di Pilkada DKI 2024 melalui data yang diambil dari media sosial YouTube. Langkah awal penelitian dimulai dengan pengumpulan data menggunakan *Jupyter Notebook*. Data dikumpulkan dalam periode Agustus hingga September 2024 dan berhasil mengumpulkan sebanyak 10.843 entri data. Setelah data terkumpul, dilakukan tahap *preprocessing* data, yang meliputi langkah-langkah:

- a) *Data cleaning*
- b) *Case folding*
- c) Tokenisasi
- d) Normalisasi
- e) *Stopword removal*
- f) *Stemming*

Setelah tahap *preprocessing* selesai, jumlah data yang bersih dan siap digunakan berkurang menjadi 10.534 entri data. Selanjutnya, dilakukan labelling data menggunakan kamus Inset (Indonesian Sentiment Lexicon). Hasil proses labelling menunjukkan distribusi sentimen sebagai berikut:

Sentimen negatif: 5.349 entri (50,35%)

Sentimen positif: 2.560 entri (24,10%)

Sentimen netral: 2.715 entri (25,56%)

Tahap berikutnya adalah visualisasi data, yang dilakukan untuk memahami distribusi dan pola sentimen. Visualisasi ini mencakup:

Wordcloud untuk sentimen positif dan negatif

Bar chart

Pie chart

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

Setelah visualisasi data selesai, dilakukan klasifikasi dan evaluasi menggunakan metode *Naive Bayes* dan *Support Vector Machine (SVM)*. Pada tahap klasifikasi digunakan metode *Naive Bayes*, dilakukan pengujian dengan memisahkan menjadi 90% data training dan 10% data testing. Hasil klasifikasi menunjukkan bahwa:

- 1) Algoritma Gaussian mencapai akurasi tertinggi sebesar 94,49%
- 2) Algoritma Multinomial mencapai akurasi sebesar 88%
- 3) Algoritma Bernoulli mencapai akurasi sebesar 80,26%

Sementara itu, pada klasifikasi menggunakan metode *Support Vector Machine (SVM)* dengan perbandingan data yang sama, akurasi tertinggi yang dicapai adalah sebesar 83%. Hasil ini menunjukkan bahwa metode *Naive Bayes* (terutama dengan algoritma Gaussian) memberikan performa klasifikasi yang lebih baik dibandingkan dengan metode *Support Vector Machine (SVM)* pada data yang digunakan dalam penelitian ini.



### **5.1 Kesimpulan**

Penelitian ini menganalisis elektabilitas Mochammad Ridwan Kamil pada Pilkada DKI 2024 melalui data media sosial YouTube, dengan hasil yang menunjukkan bahwa distribusi sentimen mayoritas masyarakat cenderung negatif, yakni 50,35%, diikuti oleh sentimen netral sebesar 25,56%, dan sentimen positif sebesar 24,10%. Hal ini mencerminkan persepsi publik yang lebih banyak tidak mendukung calon tersebut. Dalam hal kinerja metode klasifikasi, *Naive Bayes*, terutama dengan algoritma Gaussian, mencapai akurasi tertinggi sebesar 94,49%, menjadikannya lebih efektif dalam klasifikasi data dibandingkan dengan metode *Support Vector Machine (SVM)* yang hanya menghasilkan akurasi sebesar 83%. Hasil ini menunjukkan bahwa *Naive Bayes* lebih unggul dalam mengolah data sentimen terkait elektabilitas.

Selain itu, hasil penelitian mendukung prediksi peneliti bahwa tingginya sentimen negatif di media sosial YouTube menjadi indikator kuat yang menunjukkan kemungkinan kekalahan Mochammad Ridwan Kamil dalam Pilgub DKI 2024. Temuan ini menegaskan pentingnya pengelolaan persepsi publik di media sosial, yang dapat mempengaruhi citra dan elektabilitas calon pemimpin dalam kontestasi politik. Dengan demikian, penelitian ini memberikan wawasan mengenai dampak sentimen publik yang tercermin di media sosial terhadap hasil pemilihan, serta perlunya strategi komunikasi yang tepat untuk memitigasi sentimen negatif.

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

### 5.2 Saran

Untuk meningkatkan pengembangan pada penelitian ini di masa mendatang, ada beberapa hal yang dapat dilakukan, yaitu:



- 1) Memilih model klasifikasi atau algoritma yang berbeda
- 2) Memilih topik pembahasan yang berbeda, misalnya menganalisis sentimen terhadap calon pemimpin lain atau isu-isu politik tertentu, untuk memperluas cakupan penelitian
- 3) Menggunakan platform yang berbeda, seperti Twitter, Instagram, atau TikTok, yang mungkin memiliki demografi pengguna dan karakteristik data yang berbeda sehingga dapat memberikan wawasan yang lebih beragam
- 4) Memperluas periode pengumpulan data untuk mendapatkan lebih banyak entri yang mencerminkan perubahan sentimen masyarakat secara lebih dinamis



## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

- [9] V. K. S. Que, A. Iriani, and H. D. Purnomo, "Analisis Sentimen Transportasi Online Menggunakan Support Vector Machine Berbasis Particle Swarm Optimization," *J. Nas. T. Inf. dan Teknol. Inf.*, vol. 9, no. 2, pp. 162–170, 2020, doi: 10.22146/jnti.v9i2.102.
- [10] A. Liawati, R. Narasati, D. Solihudin, C. Lukman Rohmat, and S. Eka Permana, "Analisis Sentimen Komentar Politik Di Media Sosial X Dengan Pendekatan Deep Learning," *JATI (Jurnal Mhs. Tek. Inform.)*, vol. 7, no. 6, pp. 3557–3563, 2024, doi: 10.36040/jati.v7i6.8248.
- [11] T. Safitri, Y. Umaidah, and I. Maulana, "Analisis Sentimen Pengguna Twitter Terhadap Grup Musik BTS Menggunakan Algoritma Support Vector Machine," *J. Appl. Informatics Comput.*, vol. 7, no. 1, pp. 28–35, 2023, doi: 10.30871/jaic.v7i1.5039.
- [12] M. Hudha, E. Supriyati, and T. Listyorini, "Analisis Sentimen Pengguna Youtube Terhadap Tayangan #Matanajwamenantiterawan Dengan Metode Naïve Bayes Classifier," *JIKO (Jurnal Inform. dan Komputer)*, vol. 5, no. 1, pp. 1–6, 2022, doi: 10.33387/jiko.v5i1.3376.
- [13] F. Z. Ahmad, M. F. S. Arifandy, M. R. Caesarardhi, and N. A. Rakhmawati, "Bagaimana Masyarakat Menyikapi Pembelajaran Tatap Muka: Analisis Komentar Masyarakat pada Media Sosial Youtube Menggunakan Algoritma Deep Learning Sekuensial dan LDA," *J. Linguist. Komputasional*, vol. 4, no. 2, p. 40, 2021, doi: 10.26418/jlk.v4i2.57.
- [14] G. Sanjaya and K. M. Lhaksana, "Analisis Sentimen Komentar YouTube tentang Terpilihnya Menteri Kabinet Indonesia Maju Menggunakan Lexicon Based," vol. 7, no. 3, pp. 9698–9710, 2020.
- [15] A. Putri, "Pentingnya Data Cleaning Sebelum Visualisasi: Teknik Dan Tips," *Teknologipintar.org*, vol. 4, no. 5, pp. 2024–2025, 2024.
- [16] I. Iin, R. Supriatna, M. Mulyawan, and D. Rohman, "Penerapan Natural Language Processing Dalam Analisis Sentimen Cawapres 2024 Menggunakan Algoritma Naive Bayes," *JATI (Jurnal Mhs. Tek. Inform.)*, vol. 8, no. 1, pp. 1109–1115, 2024, doi: 10.36040/jati.v8i1.8572.

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

- [17] N. P. G. Naraswati, R. Nooraeni, D. C. Rosmilda, D. Desinta, F. Khairi, and R. Damaiyanti, “Analisis Sentimen Publik dari Twitter Tentang Kebijakan Penanganan Covid-19 di Indonesia dengan Naive Bayes Classification,” *Sistemasi*, vol. 10, no. 1, pp. 1–10, 2021, doi: 10.32520/stmsi.v10i1.1179.
- [18] B. Hakim, “Analisa Sentimen Data Text Preprocessing Pada Data Mining Dengan Menggunakan Machine Learning,” *JBASE - J. Bus. Audit Inf. Syst.*, vol. 4, no. 2, pp. 16–22, 2021, doi: 10.30813/jbase.v4i2.3000.
- [19] M. Dimas Lutfiyanto, E. B. Setiawan, and S. Si, “Expansion Feature dengan Word2Vec untuk Analisis Sentimen pada Opini Politik di Twitter dengan Klasifikasi Support Vector Machine, Naïve Bayes, dan Random Forest,” *eProceedings Eng.*, vol. 8, no. 5, pp. 10399–10410, 2021.
- [20] Chely Aulia Misrun, E. Haerani, M. Fikry, and E. Budianita, “Analisis sentimen komentar youtube terhadap Anies Baswedan sebagai bakal calon presiden 2024 menggunakan metode naive bayes classifier,” *J. CoSciTech (Computer Sci. Inf. Technol.)*, vol. 4, no. 1, pp. 207–215, 2023, doi: 10.37859/coscitech.v4i1.4790.
- [21] P. Putra, A. M. H. Pardede, and S. Syahputra, “Analisis Metode K-Nearest Neighbour (Knn) Dalam Klasifikasi Data Iris Bunga,” *J. Tek. Inform. Kaputama*, vol. 6, no. 1, pp. 297–305, 2022.
- [22] V. Alpiana, A. Salam, F. Alzami, I. Rizqa, and D. Aqmala, “Analisis Topic-Modelling Menggunakan Latent Dirichlet Allocation (LDA) Pada Ulasan Sosial Media Youtube,” *J. Media Inform. Budidarma*, vol. 8, no. 1, p. 332, 2024, doi: 10.30865/mib.v8i1.7127.
- [23] E. Suharyanto and A. Zein, “Analisis Data Minat Calon Mahasiswa Universitas Pamulang Dengan Menggunakan Algoritma Naive Bayes Classifier,” *Sainstech J. Penelit. Dan ...*, pp. 70–76, 2022, [Online]. Available: <https://ejournal.istn.ac.id/index.php/sainstech/article/view/1434%0Ahttps://ejournal.istn.ac.id/index.php/sainstech/article/download/1434/943>

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

- [24] M. K. Insan, U. Hayati, and O. Nurdiawan, "ANALISIS SENTIMEN APLIKASI BRIMO PADA PERASAN PENGGUNA DI," vol. 7, no. 1, pp. 478–483, 2023.
- [25] Y. Amrozi, D. Yuliaty, M. H. Silo, N. Novianto, and R. Ramadhan, "Klasifikasi Jenis Buah Pisang Berdasarkan Citra Warna dengan Metode SVM," *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 11, no. 3, pp. 394–399, 2022, doi: 10.32736/sisfokom.v11i3.1502.
- [26] S. Rabbani, D. Safitri, N. Rahmadhani, A. A. F. Sani, and M. K. Anam, "Perbandingan Evaluasi Kernel SVM untuk Klasifikasi Sentimen dalam Analisis Kenaikan Harga BBM," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 3, no. 2, pp. 153–160, 2023, doi: 10.57152/malcom.v3i2.897.
- [27] D. T. Lukmana, S. Subanti, and Y. Susanti, "Analisis Sentimen Terhadap Calon Presiden 2019 Dengan Support Vector Machine Di Twitter," *Semin. Nas. Penelit. Pendidik. Mat. 2019 UMT*, no. 2002, pp. 154–160, 2019.
- [28] Wartumi, R. Kurniawan, and A. Y. Wijaya, "Analisis Data Sentimen Ulasan Pengguna Aplikasi Shopee di Google Play Store dengan Klasifikasi Algoritma Naïve Bayes," *J. Inform. dan Rekayasa Perangkat Lunak*, vol. 6, no. 1, pp. 164–170, 2024.
- [29] V. S. Ginting, Kusriani, and E. T. Luthfi, "Keterlambatan Pembayaran Uang Sekolah Menggunakan Python," vol. 4, no. 1, 2020.
- [30] Friska Aditia Indriyani, Ahmad Fauzi, and Sutan Faisal, "Analisis sentimen aplikasi tiktok menggunakan algoritma naïve bayes dan support vector machine," *TEKNOSAINS J. Sains, Teknol. dan Inform.*, vol. 10, no. 2, pp. 176–184, 2023, doi: 10.37373/tekno.v10i2.419.
- [31] Cucu Ika Agustyaningrum, Windu Gata, Ridan Nurfalalah, Ummu Radiyah, and Mawadatul Maulidiah, "Komparasi Algoritma Naive Bayes, Random Forest Dan Svm Untuk Memprediksi Niat Pembelanja Online," *J. Inform.*, vol. 20, no. 2, 2020.

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

- [32] R. M. Koretsky, "IMPLEMENTASI METODE DATA MINING K-MEANS CLUSTERING TERHADAP DATA PEMBAYARAN TRANSAKSI MENGGUNAKAN BAHASA PEMROGRAMAN PYTHON PADA CV DIGITAL DIMENSI," *IMPLEMENTASI Metod. DATA Min. K-MEANS Clust. TERHADAP DATA PEMBAYARAN TRANSAKSI MENGGUNAKAN Bhs. PEMROGRAMAN PYTHON PADA CV Digit. Dimens.*, vol. 8, no. 1, pp. 175–305, 2023, doi: 10.1201/b23421-3.
- [33] A. Handayani and I. Zufria, "Analisis Sentimen Terhadap Bakal Capres RI 2024 di Twitter Menggunakan Algoritma SVM," *J. Inf. Syst. Res.*, vol. 5, no. 1, pp. 53–63, 2023, doi: 10.47065/josh.v5i1.4379.
- [34] D. R. Berliana and B. Santoso, "Elektabilitas Ridwan Kamil Dan Anies Baswedan Dalam Simulasi Pilpres 2024 Di Twitter (Analisis Jaringan Media Sosial Dan Analisis Sentimen Pengguna Twitter Terhadap #Ridwankamil Dan #Aniesbaswedan)," *Mediakom J. Ilmu Komun.*, vol. 6, no. 2, pp. 150–162, 2022, doi: 10.35760/mkm.2022.v6i2.6962.
- [35] M. Azhari, Z. Situmorang, and R. Rosnelly, "Perbandingan Akurasi, Recall, dan Presisi Klasifikasi pada Algoritma C4.5, Random Forest, SVM dan Naive Bayes," *J. Media Inform. Budidarma*, vol. 5, no. 2, p. 640, 2021, doi: 10.30865/mib.v5i2.2937.
- [36] B. Imran, M. N. Karim, and N. I. Ningsih, "Klasifikasi Berita Hoax Terkait Pemilihan Umum Presiden Republik Indonesia Tahun 2024 Menggunakan Naïve Bayes Dan Svm," *Din. Rekayasa*, vol. 20, no. 1, pp. 1–9, 2024, doi: 10.20884/1.dinarek.2024.20.1.27.
- [37] P. P. Tfidfvectorizer and F. Panjaitan, "PERBANDINGAN PENGGUNAAN TFIDFVECTORIZER, COUNTVECTORIZER, DAN HASHINGVECTORIZER DENGAN OPTIMALISASI PARAMETER PADA MACHINE LEARNING UNTUK ANALISIS SENTIMEN PEMILU 2024," vol. 8, no. 4, pp. 7413–7419, 2024.
- [38] M. N. Adlini, A. H. Dinda, S. Yulinda, O. Chotimah, and S. J. Merliyana, "Metode Penelitian Kualitatif Studi Pustaka," *Edumaspul J. Pendidik.*, vol. 6, no. 1, pp. 974–980, 2022, doi: 10.33487/edumaspul.v6i1.3394.

## Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

- [39] R. Watrianthos, M. Giatman, W. Simatupang, R. Syafriyeti, and N. K. Daulay, “Analisis Sentimen Pembelajaran Campuran Menggunakan Twitter Data,” *J. Media Inform. Prima*, vol. 6, no. 1, p. 166, 2022, doi: 10.30865/mib.v6i1.3383
- [40] F. A. Larasati, D. E. Ratnawati, and B. T. Hanggara, “Analisis Sentimen Ulasan Aplikasi Dana dengan Metode Random Forest,” *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 6, no. 9, pp. 4305–4313, 2022.
- [41] C. Wulandari, L. Sunardi, and Hasbiana, “KLIK: Kajian Ilmiah Informatika dan Komputer Analisis Sentimen Aplikasi Spotify Pada Ulasan Pengguna di Google Play Store Menggunakan Metode Support Vector Machine,” *Media Online*, vol. 4, no. 5, pp. 2588–2595, 2024, doi: 10.30865/klik.v4i5.1762.
- [42] M. I. Ahmadi, D. Gustian, and F. Sembiring, “Analisis Sentiment Masyarakat terhadap Kasus Covid-19 pada Media Sosial Youtube dengan Metode Naive bayes,” *J. Sains Komput. Inform. (J-SAKTI)*, vol. 5, no. 2, pp. 807–814, 2021.
- [43] A. Z. Mubarak, H. Sibyan, and N. Hasanah, “Sentiment Analysis Kinerja Karyawan Di Oke Garden Menggunakan Metode Naïve Bayes Classifier,” vol. 3, no. 2, pp. 24–30, 2023, [Online]. Available: <https://journal.unisnu.ac.id/JISTER/>
- [44] F. F. Abdulloh and I. R. Pambudi, “Analisis Sentimen Pengguna Youtube Terhadap Program Vaksin Covid-19,” *CSRID (Computer Sci. Res. Its Dev. Journal)*, vol. 13, no. 3, p. 141, 2021, doi: 10.22303/csrid.13.3.2021.141-148.
- [45] Z. Y. Burnama, M. A. Rosid, and N. L. Azizah, “Analisis Sentimen Pada Komentar Youtube Dalam Turnamen MPL Season 13 Dengan Metode Ensemble Machine Learning Sentiment Analysis on YouTube Comments in MPL Season 13 Tournament Using Ensemble Machine Learning Method”.

**Protected by PDF Anti-Copy Free**  
(Upgrade to Pro Version to Remove the Watermark)



# LAMPIRAN

# Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

Lampiran acc judul

UNIVERSITAS BINA INSAN  
Jalan Jendral Besar ... Lubuk Kumpang Kec. Lubuklinggau Selatan I Kota Lubuklinggau Provinsi Sumatera Selatan

**PDF**

**Formulir Pengajuan Judul Skripsi  
Program Studi Informatika**

Nama : Daniel Adless Erich  
NIM : 2102020154  
Alamat : Jl.Perumahan Griya Asri Blok Q No.8  
No.Hp : 082184457663

Rumusan Masalah 1 : Bagaimana kinerja algoritma Naïve Bayes dan SVM dalam menganalisis sentimen publik terhadap elektabilitas Mochammad Ridwan Kamil pada Pilkada DKI Jakarta 2024?

Judul 1 : Komparasi Algoritma Naïve Bayes dan Support Vector Machine (SVM) Dalam Menganalisis Elektabilitas Mochammad Ridwan Kamil Pada Pilkada DKI Jakarta 2024 Melalui data Media Sosial YouTube

Rumusan Masalah 2 : Bagaimana efektivitas algoritma Logistic Regression dalam mengklasifikasikan email spam?

Judul 2 : Optimasi Klasifikasi Email Spam Menggunakan Algoritma Logistic Regression

Rumusan Masalah 3 : Bagaimana perbandingan akurasi antara algoritma Naive Bayes dan Logistic Regression dalam mengklasifikasikan SMS spam?

Judul 3 : Perbandingan Algoritma Naive Bayes dan Logistic Regression dalam Klasifikasi SMS Spam

**Diusulkan Judul Nomor :1(satu)/ 2(Dua)/ 3(Tiga)\***

Lubuklinggau, September 2024  
Mahasiswa yang mengusulkan,  
*(Signature)*  
**(Daniel Adless Erich)**

Menyetujui Dosen Pembimbing,  
Pembimbing 1 (Lukman Sunardi, M. Kom) *(Signature)*  
Pembimbing 2 (Nelly Khairani Daulay, M. Kom) *(Signature)* 25/9/2024

**Mengesahkan,**  
Dekan Fakultas Ilmu Teknik  
*(Signature)*  
**(Dr. Rudi Kurniawan, St., M. Kom)**


**Mengetahui,**  
Ketua Program Studi,  
*(Signature)*  
**(Budi Santoso, M. Kom)**

0733-4553932 (Rektorat Universitas Bina Insan)  
0733-3280300 (Pascasarjana)  
0812-1826-6228 (Marketing UNIVBI)  
0852-3151-5800 (Admin UNIVBI)  
Admin@univbinainsan.ac.id  
univbinainsan.ac.id - pasca.univbinainsan.ac.id

# Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

Lampiran Bimbingan Proposal P1

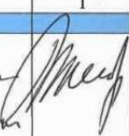





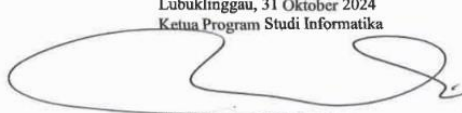
**PDF**

UNIVERSITAS BINA INSAN  
FAKULTAS ILMU TEKNIK  
Lubuklinggau, Kalimantan Selatan

### LEMBAR PEMBIMBINGAN PROPOSAL SKRIPSI

Nama : Daniel Adless Erich  
 Nim : 2102020154  
 Program Studi : Informatika  
 Pembimbing 1 : Lukman Sunardi, M. Kom  
 Pembimbing 2 : Nelly Khairani Daulay, M. Kom  
 Judul : Komparasi Algoritma Naive Bayes dan Support Vector Machine (SVM) Dalam Menganalisis Elektabilitas Mochammad Ridwan Kamil Pada Pilkada DKI Jakarta 2024 Melalui data Media Sosial YouTube

NO	TANGGAL	TOPIK	KOMENTAR PEMBIMBING	TANDA TANGAN PEMBIMBING	
				1	2
1.	11/2024	Kajian pustaka	Perbaiki format penulisan tambahkan teori NIS dan sumber tambahkan sumber referensi		
2.	211/2024	Metode penelitian	Perbaiki penulisan sumber " " " " " " " " " " " " Perbaiki daftar pustaka		
3.	411/2024	Draft proposal	Langkah Draft Proposal 1. cover 2. kata pengantar 3. Daftar isi, tabel, gambar, lampiran 4. "		
4.	G-4-2024	ACC	ACC Lanjut Sampul		



Lubuklinggau, 31 Oktober 2024  
 Ketua Program Studi Informatika  
  
 (Budi Santoso, M. Kom)

0733-4553332 (Rektorat Universitas)    0612-1826-6228 (Marketing UNIVBI)  
 0733-9280300 (Bina Insan)    0652-3151-5800 (Admin UNIVBI)  
 0733-3280200 (Pascasarjana)    Admin@univbinainstan.ac.id    univbinainstan.ac.id - pasca.univbinainstan.ac.id

# Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

Lampiran Bimbingan Proposal P2

UNIVERSITAS  
**BINA INSAN**  
AKULTAS ILMU TEKNIK

Jalan Jendral Sudirman No. 100, Lubuklinggau Selatan 1 Kota Lubuklinggau Provinsi Sumatera Selatan





DIKIRAN DWI TUNGGAL PALEMBANG

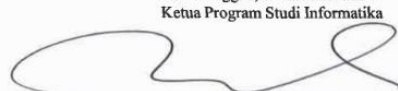
**SITAS BINA INSAN**

AKULTAS ILMU TEKNIK

### LEMBAR BIMBINGAN PROPOSAL SKRIPSI

Nama : Daniel Adless Erich  
 Nim : 2102020154  
 Program Studi : Informatika  
 Pembimbing 1 : Lukman Sunardi, M. Kom  
 Pembimbing 2 : Nelly Khairani Daulay, M. Kom  
 Judul : Komparasi Algoritma Naïve Bayes dan Support Vector Machine (SVM) Dalam Menganalisis Elektabilitas Mochammad Ridwan Kamil Pada Pilkada DKI Jakarta 2024 Melalui data Media Sosial YouTube

NO	TANGGAL	TOPIK	KOMENTAR PEMBIMBING	TANDA TANGAN PEMBIMBING	
				1	2
1	2/10/2024	Judul	Judul disesuaikan Tema dan metode yg digunakan dgn permasalahan		
2	7/10/2024	Kajian pustaka	- perbanyak kajian literatur - Setiap rumus beri nomor urut.		
3	12/10/2024	Metode penelitian	Perkuat lagi metode metode penelitian yg digunakan		
4	15/10/2024	Metode pengisian dan penulisan	Selalu gunakan tabel model terdapat untuk setiap tabel yg ada, jika ada rumus serupa maka nomor urutnya mengikuti yg pertama - tabel yang berisi angka sebaiknya buat rata-tengah		

Lubuklinggau, 31 Oktober 2024  
 Ketua Program Studi Informatika  
  
 (Budi Santoso, M. Kom)

0733-4553332 (Rektorat Universitas)

0733-3280300 (Bina Insan)

0733-3280200 (Pasosarjana)

0612-1826-6326 (Marketing UNIVBI)

0852-3151-5800 (Admin UNIVBI)

Admin@univbininsan.ac.id

univbininsan.ac.id - pasca.univbininsan.ac.id

**Protected by PDF Anti-Copy Free**  
 (Upgrade to Pro Version to Remove the Watermark)



**LEMBAR PROPOSAL SKRIPSI**

Nama : Daniel Adless E...  
 Nim : 2102020154  
 Program Studi : Informatika  
 Pembimbing 1 : Lukman Sunardi, M. Kom  
 Pembimbing 2 : Nelly Khairani Daulay, M. Kom  
 Judul : Komparasi Algoritma Naïve Bayes dan Support Vector Machine (SVM) Dalam Menganalisis Elektabilitas Mochammad Ridwan Kamil Pada Pilkada DKI Jakarta 2024 Melalui data Media Sosial YouTube

NO	TANGGAL	TOPIK	KOMENTAR PEMBIMBING	TANDA TANGAN PEMBIMBING	
				1	2
5.	30/10/2024	Daftar Pustaka	- Penulisan Daftar pustaka dibuat rata kiri kanan		
6.	31/10/2024	Ace	Ace sudah Kau Lemat Pembimbing I		

Lubuklinggau, 31 Oktober 2024  
 Ketua Program Studi Informatika

(Budi Santoso, M. Kom)

0733-4553332 (Rektorat Universitas)    0812-1826-6328 (Marketing UNIVBI)  
 0733-3280300 (Bina Insan)    0852-3151-5800 (Admin UNIVBI)  
 0733-3280200 (Pascasarjana)    Admin@univbinainson.ac.id    univbinainson.ac.id - pasia.univbinainson.ac.id

# Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

Channel YouTube Podcast Close the Door (Deddy Corbuzier)



# Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

## Channel YouTube Najwa Shihab

The screenshot displays the YouTube channel page for Najwa Shihab. At the top, there is a navigation bar with the YouTube logo and a search bar containing 'mata najwa'. Below this is the channel banner, which features the channel name 'najwa shihab' and the 'narasi' logo. The channel profile picture shows Najwa Shihab. The channel information includes the handle '@NajwaShihab', 10.2 million subscribers, and 2.6 thousand videos. A 'Disubscribe' button is visible. Below the channel information, there are tabs for 'Beranda', 'Video', 'Shorts', 'Live', 'Podcast', 'Playlist', and 'Komunitas'. The 'Video' tab is selected, showing a list of video uploads. The first video is 'MUSYAWARAH KE CALON MENTERI PARLIMEN' with a duration of 43:20. The second video is 'MUSYAWARAH KE GRAND PRIX SEASON SINGAPORE 2024' with a duration of 2:34. The third video is 'LULUK - LUKMAN: Strategi Peranteng' with a duration of 55:48. The fourth video is 'Jenis Bertahan KHOFIFAH - EMIL' with a duration of 54:04. A 'Subscription' list is visible on the left side of the page, including tvOneNews, KOMPASTV, METRO TV, and BellatorMMA.

# Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

Lampiran Perbaikan Seminar Proposal Skripsi



## LEMBAR PERBAIKAN SEMINAR PROPOSAL SKRIPSI

Nama Mahasiswa : Daniel Adress Enich  
NIM : 1101010154  
Jenjang Pendidikan : Strata 1 (S1)  
Fakultas : Ilmu Teknik  
Program Studi : Informatika  
Konsentrasi :  
Judul : Komparasi algoritma naive bayes dan support vector machine  
Dalam menganalisis (akrabilitas) muhammad Kidwan Famiil  
Pada Pilkada DKI 2014 melalui Data media Sosial Youtube

No	Dosen Penguji	Komentar Perbaikan	Tanda Tangan Ujian	Tanda Tangan Revisi
1	Lueman Sunardi, M. Kom	Alat dan bahan diturangi		
2	Alelly Khairani Dawlay, M. Kom	Waktu dan tempat penelitian diganti pada bagian tempat penelitian		
3	Dr. Muhammad Akbar, ST, M. IT	Perkuat latarbelakang alasan memilih Youtube dan media sum, kembangkan masalah ditambah, alat dan bahan diturangi Serta tempat penelitian diganti		

Lubuklinggau, 25 November 2024  
Ketua Program Studi Informatika

(Budi Santoso M. Kom)

# Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

## Lampiran Bimbingan Skripsi P1



**Nama** : Daniel Adless Erich  
**Nim** : 2102020154  
**Program Studi** : Informatika  
**Pembimbing 1** : Lukman Sunardi, M. Kom  
**Pembimbing 2** : Nelly Khairani Daulay, M. Kom  
**Judul** : Komparasi Algoritma Naïve Bayes dan Support Vector Machine (SVM) Dalam Menganalisis Elektabilitas Mochammad Ridwan Kamil Pada Pilkada DKI Jakarta 2024 Melalui data Media Sosial YouTube

NO	TANGGAL	TOPIK	KOMENTAR PEMBIMBING	TANDA TANGAN PEMBIMBING	
				1	2
	3/1 2025		Perbaiki Sesuai format Skripsi. " Penulisan gambar " Penulisan tabel		
	4/1 2025		Perbaiki bab 3 " bab 4 " Perbaiki Kesimpulan dan saran		
	7/1 2025		Lengkapi Draft Skripsi		
	8/1 2025		cece sebelum Ujian Skripsi		

Lubuklinggau, 02 Januari 2025  
Ketua Program Studi Informatika

(Budi Santoso, M. Kom)

# Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

Lampiran Bimbingan Skripsi P2



Nama : Daniel Adless Erich  
Nim : 2102020154  
Program Studi : Informatika  
Pembimbing 1 : Lukman Sunardi, M. Kom  
Pembimbing 2 : Nelly Khairani Daulay, M. Kom  
Judul : Komparasi Algoritma Naïve Bayes dan Support Vector Machine (SVM) Dalam Menganalisis Elektabilitas Mochammad Ridwan Kamil Pada Pilkada DKI Jakarta 2024 Melalui data Media Sosial YouTube

NO	TANGGAL	TOPIK	KOMENTAR PEMBIMBING	TANDA TANGAN PEMBIMBING	
				1	2
1.	5-12-2024	BAB I	-Tambahkan literatur dan Alasan memilih metode		
2.	7-12-2024	BAB II	Perbaiki penulisan kerangka berpikir, setiap gambar dan tabel nomor gambar/ tabel tersebut kaitkan dengan keterangan		
3.	10-12-2024	BAB III	Perbaiki tata letak dan penulisan pengolahan data, tabel gunakan model tabel terbuka.		
4.	12-12-2024	Bab IV	Perbaiki penulisan dan tata letak dari HASIL		
5.	15-12-2024	Bab V	esimpulan buat dalam satu paragraf saja		
6.	18-12-2024		ACC Bab I spt 5 lampir penting I		

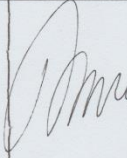
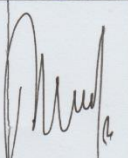


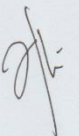
Lubuklinggau, 02 Januari 2025  
Ketua Program Studi Informatika

(Budi Santoso, M. Kom)


UNIVERSITAS BINA INSAN  
FAKULTAS ILMU TEKNIK

**LEMBAR PERBAIKAN UJIAN SKRIPSI**

Nama Mahasiswa : Daniel Adess Ench  
NIM : 2102020154  
Jenjang Pendidikan : Strata 1 (S1)  
Fakultas : Ilmu Teknik  
Program Studi : Informatika  
Konsentrasi :  
Judul : Komparasi Algoritma Naive Bayes dan Support vector machine (Svm) dalam menganalisis elektabilitas Mochammad Ridwan Kamil pada Pilkada DKI 2024 melalui Data media Sosial Youtube

No	Dosen Penguji	Komentar Perbaikan	Tanda Tangan Ujian	Tanda Tangan Revisi
1	Lukman Sunardi			
2	Nelly Khairani Dary			
3	Harna Octavia L.W			

Lubuklinggau, 20 Januari 2025  
Ketua Program Studi Informatika

  
Budi Santoso, M.Kom

**Protected by PDF Anti-Copy Free**  
(Upgrade to Pro Version to Remove the Watermark)

Lampiran Jurnal



ESCAF<sup>4th</sup> 2025  
p-ISSN : 2962-7710  
e-ISSN : 3021-8594

Daniel Adless Erich, Lukman Sunardi, Nelly Khairani Daulay

**KOMPARASI ALGORITMA *NAIVE BAYES* DAN  
*SUPPORT VECTOR MACHINE (SVM)* DALAM MENGANALISIS  
ELEKTABILITAS MOCHAMMAD RIDWAN KAMIL PADA PILKADA DKI  
2024 MELALUI DATA MEDIA SOSIAL YOUTUBE**

Daniel Adless Erich<sup>1\*</sup>, Lukman Sunardi<sup>2</sup>, Nelly Khairani Daulay<sup>3</sup>

<sup>1</sup>Program Studi Informatika, Universitas Bina Insan, Lubuklinggau  
e-mail: <sup>1</sup>2102010154@mhs.univbinainsan.ac.id, <sup>2</sup>lukmanmmci@gmail.com,  
<sup>3</sup>nellykhairanilestari@gmail.com  
<sup>\*</sup>Corresponding Author

**Abstrak**

Penelitian ini menganalisis sentimen publik terhadap Mochammad Ridwan Kamil dalam Pilkada DKI 2024 menggunakan data *YouTube*. Data dikumpulkan melalui *web scraping* pada Agustus–September 2024 dengan total 10.843 entri. Proses preprocessing mencakup data *cleaning*, *casefolding*, tokenisasi, normalisasi, *stopword removal*, dan *stemming*. Sentimen diklasifikasikan menggunakan algoritma *Naive Bayes* dan *Support Vector Machine (SVM)*. Hasil analisis menunjukkan bahwa opini negatif mendominasi sebesar 50,35%, diikuti oleh sentimen netral 25,56% dan positif 24,10%. Model *Gaussian Naive Bayes* mencapai akurasi tertinggi sebesar 94,49%, mengungguli SVM yang hanya memperoleh 83%. Temuan ini mengindikasikan bahwa dominasi sentimen negatif dapat berdampak pada elektabilitas Ridwan Kamil dalam Pilkada DKI 2024. Penelitian ini menyoroti pentingnya strategi komunikasi politik berbasis media sosial dalam mengelola persepsi publik. Hasilnya dapat menjadi acuan bagi tim kampanye dalam merancang strategi yang lebih efektif untuk meningkatkan citra kandidat.

**Kata kunci** : Analisis Sentimen, *Naive Bayes*, *Support Vector Machine*, Media Sosial, Pilkada DKI 2024.

**Abstract**

*This study analyzed public sentiment against Mochammad Ridwan Kamil in the 2024 DKI Regional Election using YouTube data. Data was collected via web scraping in August–September 2024 with a total of 10,843 entries. The preprocessing process includes data cleaning, casefolding, tokenization, normalization, stopword removal, and stemming. Sentiments are classified using the Naive Bayes and Support Vector Machine (SVM) algorithms. The results of the analysis showed that negative opinion dominated by 50.35%, followed by neutral sentiment of 25.56% and positive 24.10%. The Gaussian Naive Bayes model achieved the highest accuracy of 94.49%, outperforming the SVM which only received 83%. This finding indicates that the dominance of negative sentiment can have an impact on Ridwan Kamil's electability in the 2024 DKI Regional Head Election. This study highlights the importance of social media-based political communication strategies in managing public perception. The results can be a reference for the campaign team in devising more effective strategies to improve the candidate's image.*

**Keywords:** Sentiment Analysis, Naive Bayes, Vector Machine Support, Social Media, 2024 DKI Regional Election.

## I. PENDAHULUAN

Tahun 2024 menjadi titik yang signifikan bagi masyarakat Indonesia, di mana berbagai pemilihan umum, termasuk pemilihan kepala daerah (Pilkada), akan dilangsungkan secara serentak. Pemilihan ini tidak hanya memberikan kesempatan kepada masyarakat untuk menentukan pemimpin yang layak, tetapi juga menjadi panggung bagi para kandidat untuk menunjukkan visi dan misi mereka. Dalam dunia politik, elektabilitas juga merupakan perbincangan hangat yang menjadi fokus sentral masyarakat secara insidental, memberikan pengaruh kepada individu atau pihak tertentu. Dengan kemudahan akses terhadap informasi terkini dari berbagai sumber digital, masyarakat semakin sering memberikan tanggapan yang diartikan sebagai *feedback* untuk tokoh-tokoh atau pihak tertentu[1].

Kemajuan teknologi dalam kehidupan modern menuntut masyarakat untuk selalu bergerak cepat dan instan dalam memenuhi segala aspek kebutuhan sehari-hari. Era revolusi industri 4.0 yang memanfaatkan teknologi internet telah berdampak pada transformasi digitalisasi di hampir segala aspek kehidupan, termasuk dalam penyampaian pesan dari individu ke individu maupun dari individu ke khalayak luas. Muncul berbagai macam platform media sosial yang diciptakan untuk mewadahi kebutuhan tersebut, salah satunya adalah *YouTube*, yang berfungsi sebagai media berbagi video untuk penyampaian informasi maupun hiburan[2].

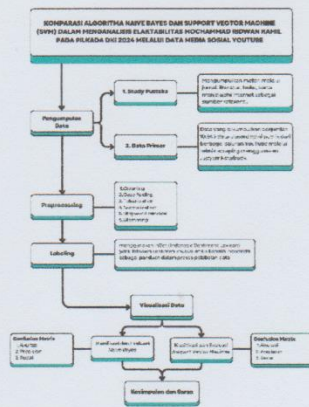
Berdasarkan laporan dari *We Are Social dan Hootsuite* pada bulan Januari 2021, aplikasi media sosial yang populer digunakan di Indonesia adalah *YouTube* yaitu 93,8 persen dari total keseluruhan pengguna internet. Kemudian untuk *Instagram* yaitu 86,6 persen, *Facebook* yaitu 85,5 persen dan *TikTok* 38,7

persen[3]. Dalam konteks ini, menilai elektabilitas Mochammad Ridwan Kamil melalui data dari komentar di media sosial *YouTube* menjadi sangat relevan.

Di sinilah teknologi machine learning, seperti algoritma *Naive Bayes* dan *Support Vector Machine (SVM)*, berperan. Algoritma ini memungkinkan klasifikasi sentimen masyarakat, baik positif maupun negatif, dilakukan secara otomatis dan efisien. *Naive Bayes* bekerja berdasarkan probabilitas dan terkenal dengan kesederhanaannya dalam menganalisis teks, sementara *Support Vector Machine (SVM)* mencari garis pemisah terbaik (*hyperplane*) yang memisahkan data sentimen menjadi kategori yang berbeda, seperti positif, negatif dan netral. Dengan kedua algoritma ini, analisis terhadap ribuan komentar di *YouTube* bisa diselesaikan secara lebih efektif. Penelitian ini tidak hanya penting untuk mengetahui elektabilitas Ridwan Kamil, tetapi juga memberikan wawasan mendalam mengenai pendekatan yang paling efektif dalam menganalisis sentimen publik di media sosial. Oleh karena itu, penelitian ini bertujuan untuk menemukan algoritma mana yang paling akurat dan efisien dalam menganalisis data sosial yang besar dan kompleks.

## II. METODOLOGI PENELITIAN

Penelitian ini menggunakan literatur mentar YouTube terkait Pilgub DKI Jakarta 2024 dari saluran Najwa Shihab dan Deddy Corbuzier. Data dikumpulkan melalui scraping selama periode Agustus–September 2024. Data ini diproses menggunakan Algoritma Naïve Bayes dan SVM.



Gambar 1. Kerangka Berfikir

## III. HASIL DAN PEMBAHASAN

### 1. Pengumpulan Data

Data yang digunakan dalam penelitian ini adalah data primer yang diambil dari ulasan pada saluran YouTube Najwa Shihab dan Deddy Corbuzier *Close the Door*. Pengambilan data dilakukan dengan menggunakan *library Python* untuk melakukan *scraping*, yang dilakukan melalui *Jupyter Notebook*. Data yang dikumpulkan mencakup atribut seperti *publishedAt*, *authorDisplayName*, *text*, dan *likeCount* dari komentar pengguna di YouTube. Total data yang terkumpul sebanyak 10.843 ribu data ulasan pengguna terkait Pilgub DKI 2024 pada kedua saluran YouTube ini, yang mencakup periode dari bulan Agustus hingga September 2024.

```
# scrape video: channel id: UC0P9P...
df = pd.DataFrame(columns=['id', 'publishedAt', 'authorDisplayName', 'text', 'likeCount'])

df

publishedAt  authorDisplayName  text
0  2024-08-20T14:14:52Z  @nawashihab1  Pahami ya bu Bu bu...
1  2024-08-20T13:42:16Z  @nawashihab1  ya baw...
2  2024-08-20T13:42:16Z  @nawashihab1  Maaf...
3  2024-08-20T13:31:16Z  @nawashihab1  ...
4  2024-08-20T13:31:16Z  @nawashihab1  ...
```

Gambar 2. Hasil Scraping

### 2. Text Processing

*Preprocessing* merupakan langkah krusial dalam analisis data yang bertujuan untuk menyiapkan data sebelum melakukan analisis lebih lanjut. Tahapan ini merupakan proses ekstraksi teks, dari data tidak terstruktur diubah menjadi data yang terstruktur agar bisa diolah lebih lanjut untuk proses klasifikasi. Secara umum, terdapat enam tahapan dalam *preprocessing* yang akan dilaksanakan yaitu:

#### a. Cleaning

Proses identifikasi, koreksi, dan penghapusan ketidaksesuaian atau anomali dalam dataset untuk meningkatkan kualitas data, memastikan keakuratan analisis, dan memungkinkan visualisasi data yang informatif [4].

```
publishedAt  authorDisplayName  text
0  2024-08-20T14:14:52Z  @nawashihab1  Pahami ya bu Bu bu...
1  2024-08-20T13:42:16Z  @nawashihab1  ya baw...
2  2024-08-20T13:42:16Z  @nawashihab1  Maaf...
3  2024-08-20T13:31:16Z  @nawashihab1  ...
4  2024-08-20T13:31:16Z  @nawashihab1  ...
```

Gambar 3. Hasil Proses Cleaning

#### b. Case Folding

*case folding* merupakan proses pemerataan huruf dari huruf kapital menjadi huruf kecil atau sebaliknya [16].

```
publishedAt  authorDisplayName  text
0  2024-08-20T14:14:52Z  @nawashihab1  Pahami ya bu Bu bu...
1  2024-08-20T13:42:16Z  @nawashihab1  ya baw...
2  2024-08-20T13:42:16Z  @nawashihab1  Maaf...
3  2024-08-20T13:31:16Z  @nawashihab1  ...
4  2024-08-20T13:31:16Z  @nawashihab1  ...
```

Gambar 4. Hasil Proses case folding

c. Tokenization

Tokenisasi adalah proses memecah konten untuk memecah konten menjadi kata-kata, istilah, simbol, dan elemen bermakna lainnya.

Gambar 5. Hasil Proses Tokenization

d. Normalisasi

Normalisasi mengubah semua bentuk kata yang diberikan menjadi satu bentuk yang baku[5].

Gambar 6. Hasil Proses Normalisasi

e. Stopword Removal

Teknik menghapus kata-kata yang umum digunakan dan kata yang tidak memiliki arti khusus seperti kata ganti, preposisi, dan konjungsi[6].

Gambar 7. Hasil Proses Stopword Removal

f. Stemming

Bertujuan untuk menghilangkan imbuhan awalan, akhiran dan sisipan untuk kembali dalam bentuk kata dasar sesuai KBBI[7].

Gambar 8. Hasil Proses Stemming

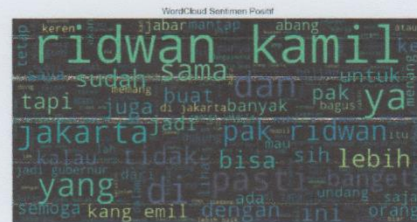
3. Labeling

Labelling merupakan proses pemberian label sentimen pada data teks berdasarkan analisis kata-kata yang ada. Dalam penelitian ini, labelling dilakukan menggunakan Inset (Indonesian Sentiment Lexicon)

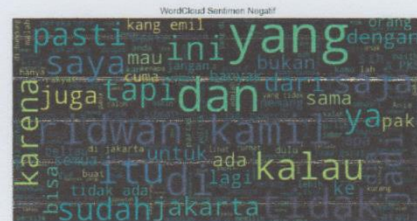
Gambar 9. Hasil Labeling

4. Visualisasi

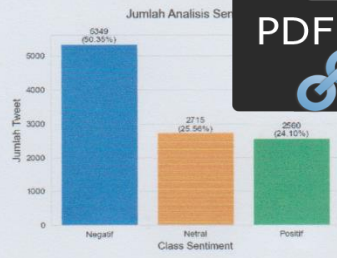
Visualisasi adalah proses penyajian data dalam bentuk grafik atau diagram untuk memudahkan pemahaman dan analisis. Dalam analisis teks, visualisasi digunakan untuk menggambarkan hasil dari proses seperti distribusi sentimen, frekuensi kata, atau hubungan antar variabel. Teknik visualisasi yang umum digunakan termasuk wordcloud, bar chart, dan pie chart, yang dapat membantu menampilkan pola atau tren yang ada dalam data. Berikut hasil visualisasi:



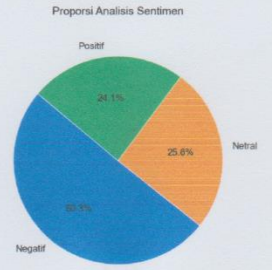
Gambar 10. Hasil wordcloud positif



Gambar 11. Hasil wordcloud negatif



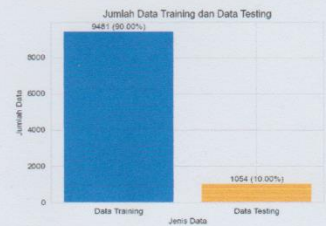
Gambar 12. Hasil visualisasi bar chart



Gambar 11. Hasil Visualisasi pie chart

5. Splitting Data

Langkah pertama dalam proses klasifikasi menggunakan metode Naive Bayes dan Support Vector Machine (SVM) adalah membagi dataset menjadi dua bagian utama, yaitu data training dan data testing. Dengan perbandingan 90% untuk data training dan 10% untuk data testing.



Gambar 12. Hasil Splitting Data

6. Evaluasi Model

Evaluasi ini memanfaatkan confusion matrix untuk menghitung metrik performa seperti accuracy, precision, recall, dan F1-score. Hasil dari proses klasifikasi pada

data training dan data testing menggunakan metode Naive Bayes dan Svm.

7. Confusion Matrix

Confusion matrix adalah alat evaluasi yang digunakan untuk mengukur kinerja model klasifikasi dalam machine learning. Matriks ini berbentuk tabel yang memperlihatkan jumlah prediksi benar dan salah dari suatu model klasifikasi, dengan membandingkan hasil prediksi model terhadap nilai aktual yang sebenarnya.

```

Gaussian Naive Bayes Accuracy: 0.9440715370618975
Gaussian Naive Bayes Confusion Matrix:
[[523 22 11]
 [ 8 256  8]
 [  0 25 217]]

Gaussian Naive Bayes Classification Report:
precision  recall  f1-score  support
Negatif    1.00    0.94    0.97    556
Netral     0.84    1.00    0.92    256
Positif    0.95    0.90    0.92    242

accuracy    0.93
macro avg   0.93    0.95    0.94    1054
weighted avg 0.95    0.94    0.95    1054

Multinomial Naive Bayes Accuracy: 0.8795060413662239
Multinomial Naive Bayes Confusion Matrix:
[[347  4  3]
 [ 74 174  8]
 [ 32  1 306]]

Multinomial Naive Bayes Classification Report:
precision  recall  f1-score  support
Negatif    0.84    0.98    0.90    556
Netral     0.96    0.87    0.92    256
Positif    0.95    0.86    0.90    242

accuracy    0.91
macro avg   0.91    0.94    0.90    1054
weighted avg 0.89    0.89    0.89    1054

Bernoulli Naive Bayes Accuracy: 0.803656544895635
Bernoulli Naive Bayes Confusion Matrix:
[[81 18  8]
 [ 83 206  8]
 [ 64 57 301]]

Bernoulli Naive Bayes Classification Report:
precision  recall  f1-score  support
Negatif    0.89    0.97    0.88    556
Netral     0.74    0.80    0.77    256
Positif    1.00    0.41    0.59    242

accuracy    0.85
macro avg   0.85    0.73    0.74    1054
weighted avg 0.83    0.80    0.78    1054
    
```

Gambar 13. Confusion Matrix Naive Bayes

```

Confusion Matrix:
[[519 36 11]
 [ 61 183 32]
 [ 25 34 189]]

Accuracy: 0.83
Accuracy: 81.94%

precision  recall  f1-score  support
Negatif    0.87    0.93    0.90    556
Netral     0.79    0.64    0.67    256
Positif    0.85    0.78    0.81    242

accuracy    0.81
macro avg   0.81    0.78    0.79    1054
weighted avg 0.82    0.83    0.82    1054
    
```

Gambar 13. Confusion Matrix Svm



Berdasarkan hasil pengklasifikasi Naive Bayes dan *Support Vector Machine* (SVM) dengan perbandingan data 90% dan data testing 10% menunjukkan bahwa akurasi tertinggi dicapai oleh algoritma *Gaussian Naive Bayes* sebesar 94,49%, diikuti oleh *Multinomial Naive Bayes* sebesar 88%, dan *Bernoulli Naive Bayes* sebesar 80,26%. Sementara itu metode *Support Vector Machine* (SVM) menghasilkan akurasi tertinggi sebesar 83%.

#### IV. KESIMPULAN

Penelitian ini menganalisis elektabilitas Mochammad Ridwan Kamil pada Pilkada DKI 2024 melalui data media sosial YouTube, dengan hasil yang menunjukkan bahwa distribusi sentimen mayoritas masyarakat cenderung negatif, yakni 50,35%, diikuti oleh sentimen netral sebesar 25,56%, dan sentimen positif sebesar 24,10%. Hal ini mencerminkan persepsi publik yang lebih banyak tidak mendukung calon tersebut. Dalam hal kinerja metode klasifikasi, *Naive Bayes*, terutama dengan algoritma *Gaussian Naive Bayes*, mencapai akurasi tertinggi sebesar 94,49%, menjadikannya lebih efektif dalam klasifikasi data dibandingkan dengan metode *Support Vector Machine* (SVM) yang hanya menghasilkan akurasi sebesar 83%. Hasil ini menunjukkan bahwa *Naive Bayes* lebih unggul dalam mengolah data sentimen terkait elektabilitas. Selain itu, hasil penelitian mendukung prediksi peneliti bahwa tingginya sentimen negatif di media sosial YouTube menjadi indikator kuat yang menunjukkan kemungkinan kekalahan Mochammad Ridwan Kamil dalam Pilgub DKI 2024. Temuan ini menegaskan pentingnya pengelolaan persepsi publik di media sosial, yang dapat mempengaruhi citra dan elektabilitas calon pemimpin dalam kontestasi politik.

#### V. SARAN

Untuk meningkatkan hasil pengembangan pada penelitian ini di masa mendatang, ada beberapa saran yang dapat dilakukan, yaitu:

- 1) Memilih model klasifikasi atau algoritma yang berbeda
- 2) Memilih topik pembahasan yang berbeda, misalnya menganalisis sentimen terhadap calon pemimpin lain atau isu-isu politik tertentu, untuk memperluas cakupan penelitian
- 3) Menggunakan platform yang berbeda, seperti Twitter, Instagram, atau TikTok, yang mungkin memiliki demografi pengguna dan karakteristik data yang berbeda sehingga dapat memberikan wawasan yang lebih beragam
- 4) Memperluas periode pengumpulan data untuk mendapatkan lebih banyak entri yang mencerminkan perubahan sentimen masyarakat secara lebih dinamis

#### VI. DAFTAR PUSTAKA

- [1] D. Sandi, E. Utami, and K. Kusnawi, "Analisis Sentimen Publik Terhadap Elektabilitas Ganjar Pranowo di Tahun Politik 2024 di Twitter dengan Algoritma KNN dan Naive Bayes," *J. Media Inform. Budidarma*, vol. 7, no. 3, p. 1097, 2023, doi: 10.30865/mib.v7i3.6298.
- [2] R. S. Ainul Wildan, R. Adam Rajagede, and R. Rahmadi, "Analisis Sentimen Politik Berdasarkan Big Data dari Media Sosial Youtube: Sebuah Tinjauan Literatur," *Pros. Autom.*, vol. 2, no. 1, 2021.
- [3] Y. Fitriani, "Pemanfaatan Media Sosial Sebagai Media Penyajian Konten Edukasi Atau Pembelajaran Digital," *J. Inf. Syst. Applied. Manag. Account. Res.*, vol. 5, no. 4, pp. 1006-1013, 2021, doi: 10.52362/jisamar.v5i4.609.
- [4] A. Putri, "Pentingnya Data Cleaning Sebelum Visualisasi: Teknik Dan Tips," *Teknologipintar.org*, vol. 4,

**Protected by PDF Anti-Copy Free**  
(Upgrade to Pro Version to Remove the Watermark)



ESCAF 4<sup>th</sup> 2025  
p-ISSN :2962-7710  
e-ISSN: 3021-8594



Daniel Adless Erich, Nelly Khairani Dauly

- no. 5, pp. 2014-2014.
- [5] B. Hakim, "Analisa Sentimen Data Text Preprocessing Pada Data Mining Dengan Menggunakan Machine Learning," *JBASE - J. Bus. Audit Inf. Syst.*, vol. 4, no. 2, pp. 16-22, 2021, doi: 10.30813/jbase.v4i2.3000.
- [6] M. Dimas Lutfiyanto, E. B. Setiawan, and S. Si, "Expansion Feature dengan Word2Vec untuk Analisis Sentimen pada Opini Politik di Twitter dengan Klasifikasi Support Vector Machine, Naïve Bayes, dan Random Forest," *eProceedings Eng.*, vol. 8, no. 5, pp. 10399-10410, 2021.
- [7] Chely Aulia Misrun, E. Haerani, M. Fikry, and E. Budianita, "Analisis sentimen komentar youtube terhadap Anies Baswedan sebagai bakal calon presiden 2024 menggunakan metode naive bayes classifier," *J. CoSciTech (Computer Sci. Inf. Technol.)*, vol. 4, no. 1, pp. 207-215, 2023, doi: 10.37859/coscitech.v4i1.4790.

# Protected by PDF Anti-Copy Free

(Upgrade to Pro Version to Remove the Watermark)

Lampiran Bebas Plagiasi



**BAGI PENELITIAN DAN PENGABDIAN KEPADA MASYARAKAT**  
**UNIVERSITAS BINA INSAN**  
Jendral Besar HM. Soeharto KM.13 Kelurahan Lubuk Kupang Kecamatan Lubuklinggau Selatan I  
Kota Lubuklinggau Provinsi Sumatera Selatan

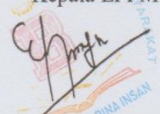
### SURAT KETERANGAN BEBAS PLAGIASI

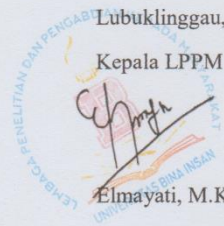
Menerangkan bahwa mahasiswa :

Nama : Daniel Adless Erich  
NIM : 2102020154  
Fakultas : Fakultas Ilmu Teknik  
Program Studi : Informatika

Memiliki jurnal dengan Judul **“komparasi Algoritma Naive Bayes Dan Support Vector Machine (SVM) Dalam Menganalisis Elektabilitas Mochammad Ridwan Kamil Pada Pilkada DKI 2024 Melalui Data Media Sosial Youtube”** Telah diterbitkan pada Prosiding : ESCAF (*Economic, Social, Computer, Agriculture and Fisheries*) 4<sup>th</sup> 2025, sehingga dinyatakan memenuhi standar bebas plagiasi dari Universitas Bina Insan.

Demikian surat keterangan ini disampaikan dengan sebenarnya untuk dapat dipergunakan sebagaimana mestinya.

Lubuklinggau, 21 Februari 2025  
Kepala LPPM  
  
Elmayati, M.Kom



☎ 0733-452218 (Rektorat Universitas)    📞 0821-8420-2909 (Rektorat Universitas)  
0733-3280300 (Bina Insan)            📞 0822-8084-5222 (Bina Insan)  
0733-3280200 (Pascasarjana)        📞 0823-7474-0255 (Pascasarjana)    ✉ Lppm@univbinaisan.ac.id    🌐 Lppm.univbinaisan.ac.id

## Protected by PDF Anti Copy Free

(Upgrade to Pro Version to Remove the Watermark)

Nama : Daniel Adless Erich  
 NIM : 2102020154  
 Konsentrasi : Analisis Data  
 Program Studi : Informatika  
 Judul : Analisis Algoritma Naive Bayes Support Vector Machine (SVM) dalam Menganalisis Elektabilitas Mochammad Ridwan Kamil pada Pilkada DKI 2024 melalui Data Media Sosial YouTube

Dosen Pembimbing I : Lukman Sunardi, M. Kom  
 Dosen Pembimbing II : Nelly Khairani Daulay, M. Kom  
 Tanggal Ujian Skripsi : 23 Januari 2025

**Point Check :**

1. SAMPUL SKRIPSI	
2. HALAMAN JUDUL SKRIPSI	
3. HALAMAN PENGESAHAN PEMBIMBING SKRIPSI	
4. HALAMAN PENGESAHAN KOMISI PENGUJI SKRIPSI	
5. SURAT PERNYATAAN	
6. ABSTRAK (BAHASA INDONESIA)	
7. ABSTRACT (BAHASA INGGRIS)	
8. MOTTO DAN HALAMAN PERSEMBAHAN	
9. KATA PENGANTAR	
10. DAFTAR ISI	
11. DAFTAR TABEL	
12. DAFTAR GAMBAR	
13. DAFTAR LAMPIRAN	
14. DAFTAR RIWAYAT HIDUP	
15. ISI SKRIPSI (BAB I S/D BAB V)	
16. DAFTAR PUSTAKA	
17. LAMPIRAN - LAMPIRAN	
➤ SK. PEMBIMBING DAN PENGUJI (PROPOSAL, HASIL, SKRIPSI)	
➤ SURAT KETERANGAN TELAH MELAKSANAKAN RISET	
➤ FORMULIR PERBAIKAN UJIAN SKRIPSI	
➤ FORMULIR PERBAIKAN SEMINAR PROPOSAL	
➤ FORMULIR BIMBINGAN SKRIPSI	
➤ FORMULIR BIMBINGAN PROPOSAL	
➤ PLAGIARISM SCAN REPORT (TURNITIN)	
➤ JURNAL (TEMPLATE ADA DI LPPM)	
➤ LISTING PROGRAM/HASIL WAWANCARA/KUISIONER DIGUNAKAN	
➤ LAIN-LAIN YANG DIPERLUKAN	

Dengan ini dinyatakan layak untuk di jilid sesuai dengan format yang berlaku dilingkungan Program Studi Informatika Universitas Bina Insan Lubuklinggau.

Pemeriksa Kelayakan,  
 Ketua Prodi Budi Santoso, M. Kom

.....