

**Hak Cipta Dilindungi Undang-Undang**

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

**KLASIFIKASI SENTIMEN MASYARAKAT DI TWITTER
PADA DATASET YANG KECIL DENGAN METODE NAIVE
BAYES CLASSIFIER**

TUGAS AKHIR

Disusun Sebagai Salah Satu Syarat
Untuk Memperoleh Gelar Sarjana Teknik
Pada Jurusan Teknik Informatika

Oleh

RIAN DELVY JURAI

NIM. 11950111735



UIN SUSKA RIAU

**FAKULTAS SAINS DAN TEKNOLOGI UNIVERSITAS ISLAM
NEGERI SULTAN SYARIF KASIM RIAU PEKANBARU**

2024

LEMBAR PERSETUJUAN

KLASIFIKASI SENTIMEN MASYARAKAT DI TWITTER PADA DATASET YANG KECIL DENGAN DENGAN METODE NAIVE BAYES CLASSIFIER

TUGAS AKHIR

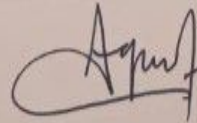
Oleh

Rian Delvy Juraidi

NIM. 11950111735

Telah diperiksa dan disetujui sebagai Laporan Tugas Akhir
di Pekanbaru, pada tanggal 05 Juli 2024

Pembimbing I,



SURYA AGUSTIAN, S.T., M.Kom.

NIP. 19760830 201101 1 003

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

LEMBAR PENGESAHAN

KLASIFIKASI SENTIMEN MASYARAKAT DI TWITTER PADA DATASET YANG KECIL DENGAN DENGAN METODE NAIVE BAYES CLASSIFIER

Oleh

Rian Delyv Juraidi

NIM. 11950111735

Telah dipertahankan di depan sidang dewan penguji
sebagai salah satu syarat untuk memperoleh gelar Sarjana Teknik
pada Universitas Islam Negeri Sultan Syarif Kasim Riau

Pekanbaru, 05 Juli 2024

Mengesahkan,

Ketua Jurusan,

IWAN ISKANDAR, S.T., M.T.

NIP. 19821216 201503 1 003

Dekan,

DR. BARTONO, M.Pd.
NIP. 19640301 1199203 1 003

DEWAN PENGUJI

Ketua : Muhammad Affandes, S.T., M.T.

Pembimbing I : Surya Agustian, S.T., M.Kom.

Penguji I : Dr. Rahmad Abdillah, S.T., M.T.

Penguji II : Yusra, S.T., M.T.

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

LEMBAR HAK ATAS KEKAYAAN INTELEKTUAL

Tugas Akhir yang tidak diterbitkan ini terdaftar dan tersedia di Perpustakaan Universitas Islam Negeri Sultan Syarif Kasim Riau adalah terbuka untuk umum dengan ketentuan bahwa hak cipta pada penulis. Referensi kepustakaan diperkenankan dicatat, tetapi pengutipan atau ringkasan hanya dapat dilakukan seizin penulis dan harus disertai dengan kebiasaan ilmiah untuk menyebutkan sumbernya.

Penggandaan atau penerbitan sebagian atau seluruh Tugas Akhir ini harus memperoleh izin dari Dekan Fakultas Sains dan Teknologi Universitas Islam Negeri Sultan Syarif Kasim Riau. Perpustakaan yang meminjamkan Tugas Akhir ini untuk anggotanya diharapkan untuk mengisi nama, tanda peminjaman dan tanggal pinjam.

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

SURAT PERNYATAAN

Saya yang bertanda tangan dibawah ini :

Nama : Rian Delvy Juraidi
Nim : 11950111735
Tempat/Tgl. Lahir : Melai, 26 Juli 2002
Fakultas/Pascasarjana : Sains dan Teknologi
Prodi : Teknik Informatika
Judul Skripsi : **Klasifikasi Sentimen Masyarakat Di Twitter Pada Dataset Yang Kecil Dengan Metode Naive Bayes Classifier**

Menyatakan dengan sebenar-benarnya bahwa :

1. Penulisan Skripsi dengan judul sebagaimana tersebut diatas adalah hasil pemikiran dan penelitian saya sendiri.
2. Semua kutipan pada karya tulis ini sudah disebutkan sumbernya.
3. Oleh karena itu skripsi saya ini, saya nyatakan bebas dari plagiat
4. Apabila dikemudian hari terbukti terdapat plagiat dalam penulisan skripsi saya tersebut, maka saya bersedia menerima sanksi sesuai peraturan perundang-undangan.

Demikian Surat Pernyataan ini saya buat dengan penuh kesadaran dan tanpa paksaan dari pihak manapun juga.

Pekanbaru, 11 Juli 2024
Yang membuat pernyataan



Rian Delvy Juraidi
NIM. 11950111735

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.



LEMBAR PERNYATAAN

Dengan ini saya menyatakan bahwa dalam Tugas Akhir ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar kesarjanaan di suatu Perguruan Tinggi, dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan didalam daftar pustaka.

Pekanbaru, 05 Juli 2024

Yang membuat pernyataan,

RIAN DELVY JURAIIDI

NIM. 11950111735

UIN SUSKA RIAU

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

LEMBAR PERSEMBAHAN

Bismillahirrahmirrahim

Alhamdulillah yaa Allah....

Bersyukur atas rezeki dan nikmat yang Allah berikan kepada penulis sehingga tugas akhir ini dapat diselesaikan. Terima kasih banyak yaa Allah.

Tak lupa pula shawat teruntut Baginda Rasulullah SAW. Dengan mengucapkan Allahumma Sholli'ala Muhammad wa'alaali Muhammad.

Tugas Akhir ini dipersembahkan kepada,

Orang tua yang selalu memberikan dukungan baik yang tampak dan tidak tampak, yang selalu memberika doa dan tuntunan sehingga penulis bisa menyelesaikan tugas akhir ini.

Terimakasih untuk semua teman-teman dan sahabat yang selalu memberikan doa, semangat, dukungan, dan bantuan sehingga penulis dapat menyelesaikan tugas akhir ini. Semoga Allah SWT. memberikan balasan yang setimpal.

UIN SUSKA RIAU

Hak Cipta Diindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

ABSTRAK

Studi ini menggunakan Algoritma Naive Bayes untuk menganalisis sentimen masyarakat terhadap Kaesang Pangarep sebagai Ketua Umum PSI di Twitter. Kaesang Pangarep menuai beragam tanggapan dari masyarakat: ada yang melihatnya sebagai langkah positif menuju reformasi, sementara yang lain skeptis terhadap potensi konflik kepentingan. Penelitian ini bertujuan mengukur opini publik dengan teknik klasifikasi sentimen, mencatat akurasi terbaik 60,35% dan F1 Score 52,07%. Hasil menunjukkan bahwa penggabungan data dari berbagai sumber seperti data Kaesang versi 1 dan 2, data terkait COVID, dan topik terbuka dapat signifikan meningkatkan akurasi model.

Kata kunci: Klasifikasi Sentimen, Naive Bayes, Kaesang Pengarep, PSI

Hak Cipta Diilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

ABSTRACT

This study utilizes the Naive Bayes algorithm to analyze public sentiment towards Kaesang Pangarep as the Chairman of the Indonesian Solidarity Party (PSI) on Twitter. Kaesang Pangarep's appointment has garnered diverse reactions from the public: some view it as a positive step towards policy reform, while others are skeptical about potential conflicts of interest. The research aims to gauge public opinion using sentiment classification techniques, achieving a best accuracy of 60.35% and an F1 Score of 52.07%. The results indicate that combining data from various sources such as Kaesang versions 1 and 2, COVID-related data, and open topics can significantly enhance model accuracy.

Keywords: Sentiment Classification, Naive Bayes, Kaesang Pengarep, PSI



KATA PENGANTAR

Assalammu 'alaikum wa rohmatullohi wa barokatuh.

Ahamdulillahi robbil'alamin, tak henti-hentinya kami ucapkan kehadiran Allah *Subhanahu wa ta'ala*, yang dengan rahmat dan hidayah-Nya kami mampu menyelesaikan Tugas Akhir ini dengan baik. Tidak lupa bershalawat kepada Nabi dan Rasul-Nya, Nabi Muhammad *Sholallohu 'alaihi wa salam*, yang telah membimbing kita sebagai umatnya menuju jalan kebaikan.

Tugas Akhir ini disusun sebagai salah satu syarat untuk mendapatkan gelar sarjana pada jurusan Teknik Informatika Universitas Islam Negeri Sultan Syarif Kasim Riau. Banyak sekali pihak yang telah membantu kami dalam penyusunan laporan ini, baik berupa bantuan materi ataupun berupa motivasi dan dukungan kepada kami. Semua itu tentu terlalu banyak bagi kami untuk membalasnya, namun pada kesempatan ini kami hanya dapat mengucapkan terima kasih kepada:

1. Bapak Prof. Dr. Khairunnas Rajab, M.Ag selaku Rektor Universitas Islam Negeri Sultan Syarif Kasim Riau.
2. Bapak Dr. Hartono, M.Pd. selaku Dekan Fakultas Sains dan Teknologi Universitas Islam Negeri Sultan Syarif Kasim Riau.
3. Bapak Iwan Iskandar, S.T., M.T. selaku Kepala Jurusan Teknik Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Sultan Syarif Kasim Riau.
4. Ibu Prof. Dr. Okfalisa, ST, M.Sc. selaku Dosen Pembimbing Akademik yang telah banyak membimbing dan membantu saya dalam Perkuliahan.
5. Bapak Surya Agustian, ST, M.Kom. selaku Dosen Pembimbing Tugas Akhir yang telah meluangkan banyak waktu dalam memberikan arahan, motivasi, kritik dan saran selama melakukan Tugas Akhir ini dapat terselesaikan.

Hak Cipta Diindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

6. Bapak Dr. Rahmad Abdillah, S.T., M.Kom selaku Dosen Penguji I yang telah memberikan masukan, motivasi dan kritik yang membangun hingga terselesaikannya Tugas Akhir ini dengan baik
7. Ibu Yusra, ST, MT selaku Dosen Penguji II yang telah memberikan masukan, motivasi dan kritik yang membangun hingga terselesaikannya Tugas Akhir ini dengan baik.
8. Bapak dan Ibu Dosen Teknik Informatika yang telah memberikan ilmu dan motivasinya.
9. Kedua orang tua dan keluarga, yang senantiasa memberikan dukungan, doa, motivasi, nasihat dan semangat sehingga penulis dapat menyelesaikan laporan ini.
10. “12110421918” selaku kekasih saya yang terus memberikan dukungan dengan tulus untuk berjuang menyelesaikan laporan ini hingga tuntas
11. Seluruh pihak yang belum penulis cantumkan, terima kasih atas dukungannya, baik material maupun spiritual.

Kami menyadari bahwa dalam penulisan laporan ini masih banyak kesalahan dan kekurangan, oleh karena itu kritik dan saran yang sifatnya membangun sangat kami harapkan untuk kesempurnaan laporan ini. Akhirnya kami berharap semoga laporan ini dapat memberikan sesuatu yang bermanfaat bagi siapa saja yang membacanya.

Wassalamu 'alaikum wa rohmatullohi wa barokatuh.

Pekanbaru, 05 Juli 2024

Penulis

UIN SUSKA RIAU

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

DAFTAR ISI

LEMBAR PERSETUJUAN.....	ii
LEMBAR PENGESAHAN.....	3
LEMBAR HAK ATAS KEKAYAAN INTELEKTUAL.....	4
LEMBAR PERNYATAAN.....	5
LEMBAR PERSEMBAHAN.....	6
ABSTRAK.....	7
ABSTRACT.....	8
KATA PENGANTAR.....	9
DAFTAR ISI.....	11
DAFTAR TABEL.....	14
DAFTAR GAMBAR.....	15
BAB I PENDAHULUAN.....	16
1.1 Latar Belakang.....	16
1.2 Rumusan Masalah.....	18
1.3 Batasan Masalah.....	18
1.4 Tujuan Penelitian.....	18
1.5 Manfaat Penelitian.....	19
BAB II KAJIAN PUSTAKA.....	20
2.1 Klasifikasi Sentimen.....	20
2.2. Text Preprocessing.....	22
2.1.1 Cleaning.....	23
2.1.2 Case Folding.....	23
2.1.3 Tokenizing.....	23

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

2.1.4	Normalisasi.....	23
2.1.5	Negation Handling	23
2.1.6	Stopword Removal	24
2.1.7	Stemming	24
2.2	Penggunaan Kamus Khusus dalam Preprocessing Teks	25
2.3	Pembobotan TF.IDF	25
2.3.1	Term Frequency (TF)	26
2.3.2	Inverse Document Frequency (IDF).....	26
2.4	Naïve Bayes Classifier	27
2.5	Confusion Matrix	29
2.6	Penelitian Terkait.....	31
BAB III METODOLOGI PENELITIAN.....		33
3.1	Pengumpulan Data.....	33
3.2	Pembagian Data	35
3.3	Preprocessing	35
3.3.1	Cleaning	35
3.3.2	Case Folding.....	36
3.3.3	Tokenizing.....	37
3.3.4	Normalisasi.....	37
3.3.5	Negation Handling	38
3.3.6	Stopword Removal	39
3.3.7	Stemming	39
3.4	Preprocessing Teks dengan Kamus Khusus	40
3.5	Skema preprocessing	41
3.6	Featuring Weighting (TF.IDF)	41

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

3.7	Klasifikasi Naive Bayes Classifier.....	42
3.8	Evaluasi	42
3.9	Kesimpulan dan Saran	43
BAB IV PEMBAHASAN		44
4.1	Tahap Text Preprocessing	44
4.2	Hasil Preprocessing dengan Kamus Khusus	46
4.3	Featuring Weighting (TF.IDF)	46
4.4	Klasifikasi Naive Bayes Classifier	47
4.5	Skema Percobaan.....	48
4.6	Proses Optimasi untuk mencari model optimal	50
4.7	Pengujian Data Uji.....	58
BAB V PENUTUP.....		61
5.1	Kesimpulan.....	61
5.2	Saran	61
DAFTAR PUSTAKA.....		62
DAFTAR RIWAYAT HIDUP.....		64

DAFTAR TABEL

Tabel 1. 1 Confusion Matrix	30
Tabel 1. 2 Penelitian Terkait	31
Tabel 2. 1 Dataset dalam penelitian	34
Tabel 2. 2 Contoh Cleaning	36
Tabel 2. 3 Contoh Case Folding	36
Tabel 2. 4 Contoh Tokenizing	37
Tabel 2. 5 Contoh Normalisasi	38
Tabel 2. 6 contoh negation handling	38
Tabel 2. 7 Contoh Stopword Removal	39
Tabel 2. 8 Contoh Stemming	40
Tabel 3. 1 <i>Text Preprocessing</i>	44
Tabel 3. 2 Skema Pembersihan Data	49
Tabel 3. 3 Baseline terhadap data kaesang v1	50
Tabel 3. 4 Proses pencarian model optimasi dengan variasi dataset Kaesang + Covid	51
Tabel 3. 5 Proses pencarian model optimasi dengan variasi dataset Kaesang + Opentopic	53
Tabel 3. 6 Proses pencarian model optimasi dengan dataset Kaesang + Covid + Opentopic	55
Tabel 3. 7 Hasil eksperimen pencarian model optimal dari berbagai skema dan kombinasi dataset training	57
Tabel 3. 8 Pengujian Data Uji	59
Tabel 3. 9 Hasil Leaderboard	59

Hak Cipta Diindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

DAFTAR GAMBAR

Gambar 3. 1 Metodologi Penelitian	33
Gambar 4. 1 TF.IDF.....	47



UIN SUSKA RIAU

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

Hak Cipta Diindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

BAB I

PENDAHULUAN

1.1 Latar Belakang

Melatih model dengan dataset kecil yang berisi label masih merupakan tantangan besar. Misalnya, saat memperluas analisis sentimen ke bahasa dan budaya baru, seringkali sulit untuk mendapatkan dataset berlabel yang lengkap dan menyeluruh (Gupta et al., 2018). Dalam situasi seperti ini, seringkali sulit untuk mendapatkan dataset berlabel yang lengkap dan menyeluruh. Dataset berlabel sangat penting untuk melatih model agar dapat memahami nuansa dan konteks dari teks dalam bahasa yang berbeda. Namun, mengumpulkan data semacam itu memerlukan banyak waktu, usaha, dan sumber daya. Selain itu, karena perbedaan budaya, cara orang mengekspresikan perasaan mereka dalam bahasa yang berbeda bisa sangat bervariasi, menambah kompleksitas dalam pembuatan dataset yang representatif. Oleh karena itu, menemukan cara yang efektif untuk melatih model dengan dataset kecil menjadi tantangan penting dalam upaya meningkatkan performa analisis.

Misalnya isu pengangkatan Kaesang Pangarep sebagai Ketua Umum PSI pada tahun 2023 yang menarik perhatian banyak orang. Sebagai anak Presiden yang populer, banyak yang mempertanyakan pengalaman politiknya. Orang-orang juga memperdebatkan alasan dia bergabung dengan PSI, apakah karena minat politik atau ada alasan lain. Di satu sisi, ada harapan bahwa elektabilitas PSI akan meningkat dengan Kaesang sebagai ketua. Di sisi lain, ada kekhawatiran bahwa partai ini mungkin kehilangan identitas kritisnya. Semua ini memicu berbagai reaksi dan diskusi di masyarakat (Adhyasta Dirgantara, 2023).

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:

- a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
- b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

Bagi PSI, mengetahui sentimen masyarakat mengenai hal ini sangat penting, untuk memperkirakan elektabilitas partai di pemilu mendatang. Bisa saja PSI meminta lembaga survey untuk mengukur sentimen terhadap isu pengangkatan tersebut. Bagi lembaga survey, tentu saja hasil analisis sentimen harus dapat diberikan segera, sehingga tidak akan cukup waktu untuk memberikan label yang cukup pada data sentimen untuk klasifikasi menggunakan machine learning.

Penelitian ini mensimulasikan problem data training yang terbatas untuk digunakan dalam klasifikasi sentimen menggunakan machine learning. Metode yang diusulkan adalah Naive Bayes (NB), yang mengembangkan model probabilistik yang sangat sederhana namun efektif untuk klasifikasi teks. Biasanya NB membutuhkan waktu pelatihan yang jauh lebih singkat dibandingkan dengan model lain seperti Support Vector Machines (SVM) (Narayanan et al., 2013)

Data training yang kecil dapat mengakibatkan model Naive Bayes memiliki performa yang kurang optimal. Hal ini disebabkan oleh keterbatasan data yang dapat digunakan untuk melatih model, yang dapat mengakibatkan estimasi probabilitas yang tidak akurat. Oleh karena itu, diperlukan strategi untuk mengatasi masalah ini dan meningkatkan akurasi klasifikasi.

Dalam beberapa kasus, dataset training yang tersedia mungkin terlalu kecil untuk menghasilkan model yang dapat dipercaya. Salah satu pendekatan untuk mengatasi ini adalah dengan memanfaatkan informasi dari dataset eksternal yang relevan, meskipun topiknya berbeda. Pendekatan ini memanfaatkan transfer learning atau teknik serupa untuk memperkaya representasi fitur yang digunakan oleh model Naive Bayes.

Penelitian ini akan menyelidiki performa yang dapat dicapai model Naive Bayes, di antara metode-metode machine learning lainnya dalam klasifikasi sentimen, dengan problem dataset yang kecil. Peningkatan

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

performa akan dilakukan dengan menguji tahap-tahap optimasi terhadap data training seperti penambahan data sentimen dengan kasus yang berbeda (data agregation), dan otimasi terhadap proses pembentukan fitur dalam NB.

1.2 Rumusan Masalah

1. Bagaimana mengoptimasi Metode Navie Bayes untuk data training yang kecil agar hasil klasifikasi lebih baik?
2. Bagaimana memanfaatkan dataset eksternal dengan topik yang berbeda untuk meningkatkan performa klasifikasi dari data training yang sedikit?

1.3 Batasan Masalah

1. Data training yang tersedia terdiri dari 300 sentimen, dengan 100 sentimen untuk setiap kelas. Ada 3 kelas: Negatif, Netral, dan Positif. Data training ini memiliki 2 versi, yang tersedia pada link GitHub dengan detail sebagaimana dijelaskan oleh (Agustian et al., 2024).
2. Data eksternal yang bisa digunakan mencakup data Covid dengan 800 sentimen dan data Opentopic dengan 15.000 sentimen, namun topik atau isu dari data Opentopic tidak spesifik.
3. Data testing yang digunakan mencakup 924 dengan label gold standard yang disimpan di server leaderboard

1.4 Tujuan Penelitian

Mengembangkan model klasifikasi sentimen untuk dataset tweet terbatas menggunakan metode Naïve Bayes yang dioptimalkan dengan TF x IDF, kemudian mengevaluasi peningkatan kinerja klasifikasi sentimen melalui penggunaan data eksternal sebagai tambahan pada data training yang terbatas.

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

1.5 Manfaat Penelitian

Penelitian ini memberikan wawasan penting tentang bagaimana masyarakat merespons Kaesang, Ketua Umum Partai Solidaritas Indonesia. Tujuan utamanya adalah untuk memperluas pengetahuan tentang penggunaan metode Naive Bayes Classifier dalam mengklasifikasikan sentimen di media sosial, khususnya Twitter. Dengan demikian, penelitian ini tidak hanya memberikan pemahaman yang lebih baik tentang pandangan publik, tetapi juga berpotensi untuk meningkatkan penggunaan algoritma klasifikasi dalam analisis sentimen di platform media sosial.



UIN SUSKA RIAU

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

BAB II

KAJIAN PUSTAKA

2.1 Klasifikasi Sentimen

Klasifikasi adalah langkah penting dalam membangun model dari suatu dataset. Tujuannya adalah untuk membuat prediksi kelas atau label suatu kasus berdasarkan model yang telah diperoleh dari data pelatihan (Zuhdi et al., 2019). Proses klasifikasi dimulai setelah tahap preprocessing selesai. Pertama, sejumlah tweet dipilih secara acak sesuai dengan jumlah yang ditentukan. Setelah proses klasifikasi selesai, hasilnya akan menampilkan tweet-tweet yang digunakan sebagai data uji beserta prediksi sentimennya. Setiap tweet akan dibandingkan dengan label aktual yang telah ditetapkan oleh tenaga ahli secara manual untuk memverifikasi ketepatan prediksi (Astari et al., 2020).

Penelitian ini bertujuan untuk mengklasifikasi sentimen terhadap isu Kaesang Pangarep sebagai Ketua Umum Partai Solidaritas Indonesia. Melalui klasifikasi sentimen ini, diharapkan dapat lebih memahami pandangan serta tanggapan masyarakat terhadap peran atau isu yang melibatkan Kaesang Pangarep dalam kepemimpinan partai tersebut. Data tweet yang dipilih secara acak akan dijadikan sampel untuk pengujian, dengan hasil prediksi sentimen memperlihatkan bagaimana respons masyarakat terhadap isu tersebut.

Naive Bayes adalah metode pembelajaran mesin yang menggunakan Teorema Bayes dengan asumsi independensi yang kuat antara fitur-fitur (Wahyuningsih & Utari, 2018). Metode ini umumnya digunakan dalam berbagai tugas klasifikasi karena efektifitasnya, terutama pada data yang memiliki struktur sederhana. Dalam algoritma Naive Bayes, Teorema Bayes digunakan untuk menggabungkan probabilitas sebelumnya (prior probability) dan probabilitas bersyarat (conditional probability) dalam suatu

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:

- a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
- b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

rumus, yang digunakan untuk menghitung probabilitas berbagai klasifikasi yang mungkin terjadi. Naive Bayes mengacu pada konsep probabilitas dan statistik yang diperkenalkan oleh Thomas Bayes, seorang ilmuwan Inggris, dan bekerja dengan melakukan prediksi probabilitas kejadian di masa depan berdasarkan pengalaman dari kejadian di masa sebelumnya (Darwis et al., 2021).

Penelitian ini menggunakan Naive Bayes Classifier dan menerapkan seleksi fitur dengan Information Gain. Studi ini bertujuan untuk menganalisis sentimen terhadap maskapai penerbangan. Hasil penelitian menunjukkan bahwa metode Naive Bayes Classifier mencapai tingkat akurasi sebesar 81% dalam tugas tersebut (Septianingrum et al., 2021).

Pengujian dilakukan dengan menggunakan confusion matrix melalui library SVM, menggunakan model yang sebelumnya telah dilatih dan diuji. Confusion matrix ini berukuran 3x3 dan memperlihatkan kelas aktual serta prediksi dari model. Hasil dari pengujian menunjukkan bahwa dalam klasifikasi sentimen, mayoritas kecenderungan adalah sentimen negatif, mencapai 77%. Sentimen netral mencakup 15% dari hasil klasifikasi, sementara sentimen positif hanya sebesar 8%. Data ini dihasilkan dengan membagi dataset menjadi dua bagian: 80% untuk pelatihan model dan 20% untuk pengujian model (Darwis et al., 2020).

Penelitian oleh Haga Simada Ginting, Kemas Muslim Lhaksmana, dan Danang Triantoro Murdiansyah berjudul “Klasifikasi Sentimen Terhadap Bakal Calon Gubernur Jawa Barat 2018 di Twitter Menggunakan Naive Bayes” mengungkapkan hasil analisis sentimen terhadap bakal calon gubernur. Berdasarkan penelitian ini, Deddy Mizwar mendapat respons positif tertinggi dengan persentase 34,3%, diikuti oleh Dedi Mulyadi dengan 32,88%, dan Ridwan Kamil dengan 23,36%. Data dikumpulkan dari 7 Desember 2017 hingga 14 Desember 2017 menggunakan metode Naive Bayes, yang memungkinkan klasifikasi sentimen dengan tingkat akurasi yang signifikan berdasarkan analisis tweet yang relevan (Ginting et al.,

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

2018).

Penelitian mengenai analisis sentimen terhadap tweet tentang Maxim di Twitter menggunakan pemrograman R dan algoritma K-Nearest Neighbors (KNN) dilakukan dengan mengumpulkan 1639 tweet berbahasa Indonesia secara acak melalui API Twitter di RStudio. Tahap awal melibatkan metode Lexicon Based untuk mengkategorikan sentimen sebagai positif, netral, atau negatif. Selanjutnya, klasifikasi dilakukan menggunakan algoritma KNN. Pengujian dilakukan dengan tiga skema: pertama, menggunakan 80% data latih dan 20% data uji; kedua, menggunakan 75% data latih dan 25% data uji; ketiga, menggunakan 70% data latih dan 30% data uji. Setiap skema diuji dengan nilai k dari 1 hingga 10. Hasil terbaik diperoleh pada skema pertama dengan 80% data latih dan 20% data uji, menggunakan $k=1$, menghasilkan akurasi sebesar 95,43% Text Preprocessing (Diwandanu & Wisudawati, 2023).

2.2. Text Preprocessing

Text preprocessing merupakan langkah awal dalam pemrosesan teks sebelum melakukan analisis atau tahap lebih lanjut dalam pemrosesan bahasa alami (Natural Language Processing atau NLP). Fungsinya adalah menyusun teks mentah menjadi bentuk yang lebih cocok untuk analisis atau pemodelan.

Text Preprocessing adalah metode dalam data mining yang mencakup transformasi data awal menjadi format yang lebih teratur dan dapat dimengerti. Data mentah sering kali memiliki kekurangan, ketidaksesuaian, dan potensi kesalahan yang signifikan (Andika et al., 2019).

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

Berikut adalah beberapa langkah umum dalam Text Preprocessing:

2.1.1 Cleaning

Pembersihan data adalah proses yang melibatkan evaluasi kualitas data dengan melakukan perubahan, modifikasi, atau penghapusan data yang dianggap tidak diperlukan, kurang lengkap, tidak akurat, atau memiliki format atau struktur file yang salah dalam basis data. Tujuannya adalah untuk menghasilkan data yang memiliki tingkat kualitas yang tinggi (Darwis et al., 2021).

2.1.2 Case Folding

Case Folding merupakan langkah untuk mengubah semua huruf dalam teks menjadi huruf kecil atau besar. Hal ini dilakukan agar tidak ada perbedaan berdasarkan kapitalisasi dalam penguraian teks. Misalnya, "Halo" dan "halo" akan dianggap sama setelah proses *Case Folding*.

2.1.3 Tokenizing

Tokenizing adalah tahap pemotongan string input berdasarkan tiap kata yang menyusunnya. Proses ini juga mencakup pembuangan beberapa karakter yang dianggap sebagai tanda baca (Andika et al., 2019). Token dapat berupa kata, frasa, atau karakter, tergantung pada tujuan analisis yang ingin dicapai.

2.1.4 Normalisasi

Proses normalisasi pada proses ini dilakukan perubahan kata yang tidak baku menjadi kata yang baku (Verawati & Audit, 2022). Contohnya, menggantikan "uda" dengan "sudah" atau "gimana" dengan "bagaimana".

2.1.5 Negation Handling

Tiap tweet yang mencakup kata-kata yang mengindikasikan negasi akan mengalami perubahan dalam nilai sentimennya. Kata-kata dengan sifat negasi seperti "bukan," "tidak," "enggak," "ga,"

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

"jangan," "nggak," "tak," dan "gak" akan mengalami perubahan dalam sentimennya jika mendahului kata-kata yang semula memiliki sentimen positif, dan sebaliknya (Rustiana & Rahayu, 2017a). Tujuan dari Negation Handling adalah untuk mendapatkan pemahaman yang lebih akurat dari teks yang berisi negasi. Saat negasi muncul, makna emosional dari kata-kata yang mengikuti negasi tersebut dapat mengalami perubahan yang signifikan. Contohnya, dalam kalimat "Saya tidak sudi memberikan bantuan," kata "tidak" mengubah perasaan dari positif menjadi negatif.

2.1.6 Stopword Removal

Pada langkah ini, kumpulan tweet yang telah melalui proses pembersihan akan menghapus karakter, tanda baca, dan kata-kata umum yang tidak memberikan makna atau informasi yang relevan (Rustiana & Rahayu, 2017a). seperti "dan", "atau", "di", dan sejenisnya.

2.1.7 Stemming

Stemming adalah tahap dalam pemrosesan teks yang bertujuan untuk mengubah kata-kata menjadi bentuk dasarnya sesuai dengan kaidah bahasa Indonesia yang benar. Misalnya, kata "berlari", "berlarian", dan "berlalu" akan diganti dengan "lari" setelah dilakukan *stemming*. Ini membantu mengurangi variasi kata dengan akar kata yang sama.

Setelah menyelesaikan tahap Text Preprocessing, teks akan siap untuk ekstraksi fitur lebih lanjut, pelatihan dengan model pembelajaran mesin, atau analisis menggunakan algoritma pemrosesan bahasa alami lainnya. Pengolahan awal yang tepat dapat meningkatkan kualitas hasil analisis dan meningkatkan kinerja model dalam tugas-tugas seperti penilaian sentimen.

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:

- a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
- b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

2.2 Penggunaan Kamus Khusus dalam Preprocessing Teks

Penelitian telah menunjukkan bahwa penggunaan kamus sendiri dapat meningkatkan akurasi klasifikasi sentimen, terutama dalam menghadapi frasa slang dan negasi (Tiwari & Sinha, 2020). Pada penelitian ini, preprocessing teks dilakukan dengan menggunakan tiga kamus khusus yang dibuat sendiri. Kamus-kamus ini digunakan untuk memastikan konsistensi dan akurasi dalam proses normalisasi, handling negasi, dan penghapusan stopwords.

Kamus Normalisasi: Kamus ini berisi pasangan kata slang atau tidak baku dengan kata yang telah dinormalisasi atau baku. Contoh:

- ga → tidak
- gk → tidak
- ngga → tidak

Kamus Handling Negasi: Kamus ini berfungsi untuk menangani frasa yang mengandung negasi. Kata-kata negasi seperti "tidak", "bukan", dan sebagainya diidentifikasi dan diproses untuk memperbaiki analisis sentimen. Contoh:

- tidak → tidak_suka
- bukan → bukan_baik

Kamus Stopword: Kamus ini berisi daftar kata-kata umum yang tidak memiliki makna penting dalam analisis teks seperti "dan", "yang", "di", dll. Kata-kata ini dihapus dari teks untuk mengurangi noise dan meningkatkan kualitas fitur yang relevan.

2.3 Pembobotan TF.IDF

Pentingnya pembobotan TF.IDF terletak pada frekuensi kemunculan kata dalam suatu dokumen; semakin sering kata itu muncul dalam satu dokumen, semakin besar nilai kontribusinya. Namun, jika kata tersebut sering muncul dalam berbagai dokumen, nilai kontribusinya akan lebih kecil (Yutika et al., 2021). Metode Term Frequency Inverse Document Frequency (TF.IDF)

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:

- a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
- b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

merupakan suatu teknik untuk menilai sejauh mana sebuah istilah mewakili konten dalam suatu dokumen dengan memberikan bobot kepada setiap kata yang terdapat di dalamnya (Zhafira et al., 2021).

TF-IDF terdiri dari dua komponen yaitu:

2.3.1 Term Frequency (TF)

Pada awalnya, algoritma Term Frequency – Inverse Document Frequency (TF-IDF) dikembangkan untuk keperluan information retrieval, namun kini lebih sering digunakan untuk membandingkan dokumen (Vitandy et al., 2019). Term Frequency mengukur frekuensi kemunculan sebuah kata dalam dokumen. Bobotnya meningkat seiring dengan peningkatan frekuensi kata tersebut. Perhitungan Term Frequency dapat dilakukan menggunakan rumus pada persamaan 2.1 berikut:

$$TF(t, d) = \frac{\text{Jumlah kemunculan term } t \text{ dalam dokumen } d}{\text{Total jumlah term dalam dokumen } d} \quad (2.1)$$

2.3.2 Inverse Document Frequency (IDF)

Inverse Document Frequency (IDF) adalah perhitungan seberapa sering suatu kata muncul dalam semua dokumen berformat teks. Kata-kata yang umumnya muncul dalam semua dokumen memiliki nilai IDF yang lebih rendah dibandingkan dengan kata-kata yang jarang muncul (Nurdiansyah et al., 2021). *Inverse Document Frequency* dapat dihitung dengan menggunakan rumus dalam persamaan 2.2 berikut:

$$IDF(t, D) = \log \left(\frac{N}{1 + \text{Jumlah dokumen yang mengandung term } t} \right) \quad (2.2)$$

dimana N adalah total jumlah dokumen dalam korpus.

Setelah menghitung term frequency dan inverse document frequency, bobot kata-kata dalam dokumen diperoleh dengan mengalikan nilai TF dengan nilai IDF sebagaimana persamaan 2.3 di bawah ini.

$$TF-IDF(t, d, D) = TF(t, d) \times IDF(t, D) \quad (2.3)$$

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:

- a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
- b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

Dengan menggunakan metode TF.IDF, kata-kata umum dalam dokumen yang sering muncul di seluruh korpus akan memiliki bobot rendah, sementara kata-kata yang jarang muncul dalam dokumen tetapi hadir dalam jumlah terbatas dalam korpus akan memiliki bobot lebih tinggi. Pendekatan ini membantu dalam mengidentifikasi kata kunci yang paling relevan dan informatif dalam suatu dokumen atau kelompok dokumen.

2.4 Naïve Bayes Classifier

Metode klasifikasi Naive Bayes merupakan salah satu pendekatan yang populer dalam penggalian data, karena kemudahan penggunaannya, kecepatan pemrosesan, implementasi yang simpel, dan tingkat efektivitas yang tinggi. Konsep dasar dari Naive Bayes melibatkan prediksi peluang di masa depan berdasarkan pengalaman masa lalu (Zhafira et al., 2021). Naive Bayes Classifier menghitung probabilitas setiap fitur dalam dataset untuk setiap kelas yang mungkin. Dengan menerapkan teorema Bayes, algoritma ini mengevaluasi probabilitas kelas untuk data berdasarkan probabilitas fitur dalam dataset. Kelas dengan probabilitas tertinggi dipilih sebagai prediksi kelas untuk data tersebut.

Naive bayes menyatakan bahwa perkalian probabilitas suatu kelas C berdasarkan fitur x ($P(C|x)$) dengan probabilitas prior dari keseluruhan fiturnya ($P(x)$), sama dengan probabilitas fitur penyusun berdasarkan kelasnya $P(x|C)$ dikali dengan probabilitas prior dari kelasnya ($P(C)$). Sehingga untuk menentukan kelas dari suatu set fitur yang diketahui, dapat dihitung dari probabilitas fitur berdasarkan kelasnya dikali dengan probabilitas prior kelas yang dievaluasi, dibagi dengan probabilitas prior untuk keseluruhan fiturnya, atau dapat dituliskan dalam persamaan 2.4.

$$P(C|x) = \frac{P(x|C) \cdot P(C)}{P(x)} \quad (2.4)$$

Kelas diprediksi dengan membandingkan probabilitas posterior ($P(C|x)$) yang telah dihitung untuk masing-masing kelas. Data akan diklasifikasikan pada kelas dengan nilai probabilitas yang tertinggi, sehingga faktor denominator $P(x)$

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

dapat diabaikan. Sehingga persamaan 2.4 dapat disederhanakan menjadi persamaan 2.5 berikut.

$$P(C|x) \propto P(x|C) \cdot P(C) \quad (2.5)$$

Langkah awal menghitung prediksi dengan metode NB adalah dengan menghitung terlebih dahulu probabilitas prior untuk masing-masing kelas yang dievaluasi, yang dihitung berdasarkan persamaan 2.6.

$$P(C) = \frac{N_C}{N} \quad (2.6)$$

di mana N_C adalah Jumlah dokumen pada kelas C , dan N adalah Jumlah total dokumen yang ada pada data training.

Pada umumnya multinomial NB menggunakan fitur *bag of words* (daftar *vocabulary* v) dengan jumlah kemunculan kata (*word occurrence*) dari setiap kata di dalam dokumen (tweet). Maka untuk menghitung probabilitas setiap kata berdasarkan kelasnya, menggunakan persamaan 2.7. di bawah ini.

$$P(x|C_k) = \frac{\text{WordCount kata } (x) \text{ dalam dokumen kelas } (k) + 1}{\text{jumlah kata pada kelas } (k) + \text{jumlah semua kata } (v)} \quad (2.7)$$

Namun, pada penelitian ini, penulis menggunakan fitur TF.IDF, sehingga perhitungan probabilitas kata berdasarkan kelas menjadi berubah, yang dihitung menurut langkah-langkah berikut ini.

$$\hat{\theta}_{ci} = \frac{\alpha_i + \sum_{j:y_j \neq c} d_{ij}}{\alpha + \sum_{j:y_j \neq c} \sum_k d_{kj}} \quad (2.8)$$

di mana α_i adalah parameter smoothing untuk setiap kata seperti pada penerapan *word occurrence*, dan $\alpha = \sum_i \alpha_i$ untuk setiap kata pada urutan ke- i . Indeks j adalah menyatakan data tweet ke- j , dan indeks k menyatakan kelas ke- k (positif, netral atau negatif). Penjumlahan dengan sigma menyatakan bahwa yang dihitung adalah fitur d_{ij} dari seluruh dokumen j yang bukan berada pada kelas c yang dievaluasi.

Variabel d_{ij} berisi nilai dari TF.IDF dari kata ke- i pada dokumen j .

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

Sedangkan variabel d_{kj} adalah berisi nilai TF.IDF yang dievaluasi untuk seluruh kelas k yang ada ($k=3$ untuk positif, negatif dan netral dalam penelitian ini).

Untuk memudahkan komputasi (perkalian) terhadap banyak bilangan desimal yang lebih kecil dari 1, maka digunakan bentuk logaritmik sebagaimana persamaan (2.6). Skala logaritmik akan membuat angka desimal hasil perkalian menjadi lebih mudah dibaca, karena memiliki jumlah kata yang banyak dengan probabilitas masing-masing pada nilai di antara $\{0, 1\}$. Sehingga dengan mengganti dengan bentuk logaritmik, persamaan yang tadinya berbentuk perkalian, menjadi penjumlahan saja.

Dari penggunaan TF.IDF dalam persamaan 2.8, maka dalam bentuk logaritmik persamaan 2.7 dapat dituliskan kembali menjadi:

$$P(x|C_k) = \log \hat{\theta}_{ck} \quad (2.9)$$

Maka, kelas hasil prediksi dapat dihitung dengan persamaan 2.10 berikut ini.

$$P(C|x) \propto \prod_{i=1}^n P(x_i|C) \cdot P(C) \quad (2.10)$$

Penelitian ini menggunakan library `tfidf_vectorizer`¹ dari `sklearn` untuk menghitung probabilitas kata berdasarkan TF.IDF, dan library `multinomialNB`² untuk pembentukan model klasifikasinya.

2.5 Confusion Matrix

Confusion matrix adalah tabel yang menampilkan jumlah data uji yang diklasifikasikan dengan benar dan yang diklasifikasikan dengan kesalahan oleh satu model klasifikasi (Normawati & Prayogi, 2021). *Confusion Matrix* menjelaskan jumlah prediksi yang benar dan salah yang dibuat oleh model, serta mengukur akurasi, presisi, recall, dan F1 Score dari model tersebut.

Berikut adalah komponen hasil klasifikasi yang dihitung untuk tiga kelas, yaitu Positif (Pos), Negatif (Neg), dan Netral (Net). "Aktual" mengacu pada hasil

¹ https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

² https://scikit-learn.org/stable/modules/naive_bayes.html

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

apotasi yang dilakukan manusia sebagai standar emas untuk pengujian, sementara "prediksi" adalah hasil yang diperoleh oleh sistem yang dikembangkan. "True" (T) menunjukkan bahwa hasil prediksi benar sesuai dengan nilai aktualnya, sementara "false" (F) menunjukkan bahwa hasil prediksi tidak sesuai dengan kelas yang seharusnya pada nilai aktual.

Tabel 1. 1 Confusion Matrix

		PREDIKSI		
		Positif	Negatif	Netral
AKTUAL	Positif	Tpos	FPosNeg	FPosNet
	Negatif	FNegPos	TNeg	FNegNet
	Netral	FNetPos	FNetNeg	TNet

Menggunakan *Confusion Matrix*, kita dapat menghitung beberapa evaluasi pentingseperti:

1. Accuracy (Akurasi): Prosentase prediksi yang benar dibandingkan dengan total data.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

2. Precision (Presisi): Proporsi kasus positif yang terprediksi dengan benar dibandingkan dengan total prediksi positif.

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

3. Recall (Sensitivity): Presentase kasus positif yang terprediksi dengan benar dibandingkan dengan total kasus positif aktual.

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

4. F1 score: Nilai rata-rata harmonik antara presisi dan recall.

$$F1\ Score = \frac{2 \times Recall \times Precision}{Recall + Precision} \quad (4)$$

Hak Cipta Dilindungi Undang-Undang

1. Diarangi mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

2.6 Penelitian Terkait

Pada penelitian ini, dilakukan studi literatur penelitian terkait sebagai bahan untuk referensi dan evaluasi dari penelitian yang dilakukan. Berikut adalah penelitian yang terkait:

Tabel 1. 2 Penelitian Terkait

No	Peneliti	Judul	Perbedaanya
1	Lingga Aji Andika, Pratiwi Amalia Nur Azizah, dan Respatiwan	Analisis Sentimen Masyarakat terhadap Hasil Quick Count Pemilihan Presiden Indonesia 2019 pada Media Sosial Twitter Menggunakan Metode Naive Bayes Classifier	Penelitian ini bertujuan untuk menemukan model Naive Bayes terbaik dan mengklasifikasikan sentimen. Hasilnya menunjukkan akurasi terbaik sebesar 82,90% dengan nilai signifikansi $\alpha = 0,05$. Klasifikasi yang didapat adalah 34,5% (471) tweet positif dan 65,5% (895) tweet negatif hasil quick count.
2	Nurman Satya Marga, Auliya Rahman Isnain, Debby Alita	Sentimen analisis tentang kebijakan pemerintah Terhadap kasus corona menggunakan metode naive Bayes	Penelitian menggunakan Naive Bayes untuk analisis sentimen tweet tentang kebijakan pemerintah terhadap COVID-19 menemukan bahwa 56,39% publik berpendapat positif dan 43,61% berpendapat negatif terhadap pemberlakuan sistem New Normal, berdasarkan 1823 tweet antara 6 Juli hingga 25 Juli 2020. Dalam pengujian menggunakan confusion matrix, Naive Bayes dengan ekstraksi fitur TF.IDF mencapai akurasi 81%, dengan precision 78%, recall 91%, dan F1-score 84%. Penggunaan N-Gram jenis Trigram meningkatkan akurasi menjadi 84%, dengan precision 84%, recall 86%, dan F1-score 85%, menunjukkan kemampuan baik algoritma ini dalam mengklasifikasikan tweet dengan frasa bahasa Indonesia yang panjang.
3	Amar P. Natasuwarna	Analisis sentimen keputusan pemindahan Ibukota negara menggunakan klasifikasi Naive bayes	Penelitian ini menunjukkan secara umum grafik berupa akurasi yang mengalami kenaikan nilai berbanding lurus dengan kenaikan data latih. Grafik akurasi secara konsisten mengalami kenaikan secara gradual bermula dari 87,00% pada rasio 50:50 menjadi 92,00% pada rasio 90:10, diperoleh rata-rata akurasi menjadi 89,86%. Oleh sebab itu, rasio 90:10 dan rasio 70:30. sering digunakan pada

Hak Cipta Dilindungi Undang-Undang

1. Diarangi mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

4	Sunneng Sandino Berutu	Text Mining dan Klasifikasi Sentimen Berbasis Naïve Bayes Pada Opini Masyarakat terhadap Makanan Tradisional	<p>penelitian untuk mendapatkan nilai akurasi.</p> <p>Penelitian ini melibatkan pengumpulan data dari internet, pembersihan data untuk menghilangkan noise, penyaringan data untuk memilih yang relevan, penerjemahan data ke dalam bahasa yang digunakan, pembagian data menjadi bagian-bagian yang lebih kecil, dan pengembangan model untuk mengklasifikasikan teks menggunakan Naïve Bayes. Hasil analisis menunjukkan bahwa makanan gudeg memiliki persentase sentimen positif tertinggi sebesar 57,9%. Sebaliknya, makanan rendang memiliki persentase sentimen negatif tertinggi sebesar 21,9%. Sedangkan, makanan hamburger memiliki persentase tertinggi untuk sentimen netral. Model klasifikasi yang dikembangkan menggunakan dataset hamburger mencapai nilai akurasi tertinggi sebesar 0,72, dengan presisi 0,72 dan recall 0,68 dalam memprediksi sentimen data.</p>
5	Eni Tri Handayani , Ari Sulistiyawati	Analisis Sentimen Respon Masyarakat Terhadap Kabar Harian COVID-19 pada Twitter Kementerian Kesehatan Dengan Metode Klasifikasi Naive Bayes	<p>Penelitian ini menghasilkan sentimen masyarakat pengguna Twitter terkait respon terhadap kabar harian COVID-19 yang disampaikan oleh akun resmi Kementerian Kesehatan Republik Indonesia. Diperoleh presentase kelas sentimen negatif sebesar 77%. Hasil pengujian menunjukkan tingkat akurasi sebesar 78%, dengan precision 92%, recall 85%, dan F1-Score 88%.</p>

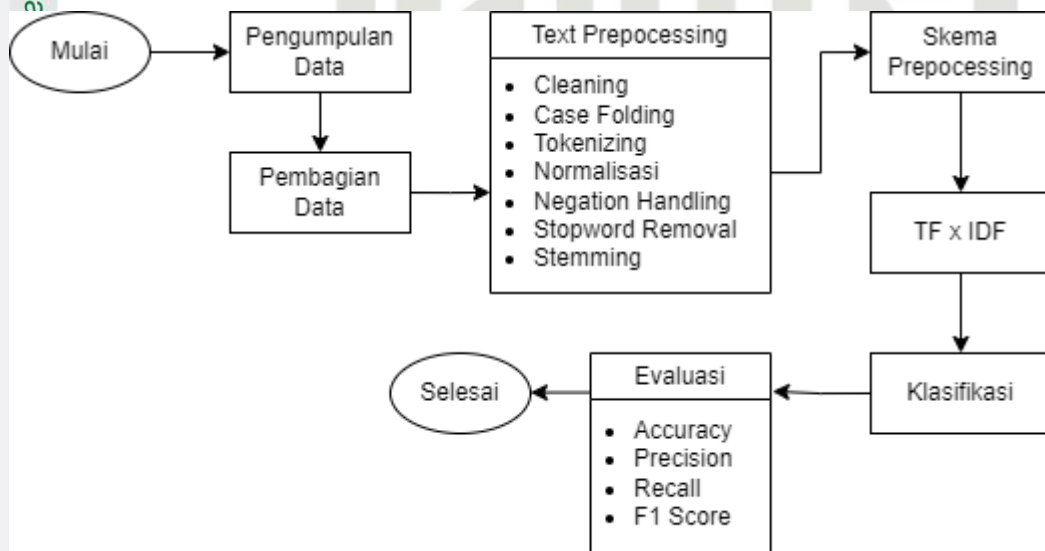
Dari hasil penelitian terkait dapat disimpulkan bahwa naive bayes telah terbukti efektif dalam mengklasifikasikan sentimen dari tweet terkait kebijakan pemerintah dan respons masyarakat terhadap COVID-19. Hasil penelitian menunjukkan bahwa sebagian besar masyarakat menunjukkan sentimen negatif terhadap kebijakan seperti pemberlakuan sistem New Normal. Analisis juga mengungkap variasi dalam sentimen terhadap makanan tertentu seperti gudeg, rendang, dan hamburger. Metode ekstraksi fitur seperti TF.IDF dan N-Gram memberikan kontribusi dalam meningkatkan akurasi model, memungkinkan identifikasi yang lebih baik terhadap nuansa dan konteks dalam teks bahasa Indonesia.

Hak Cipta Dilindungi Undang-Undang

1. Diarangi mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

BAB III METODOLOGI PENELITIAN

Metodologi penelitian merujuk pada penjelasan mengenai proses langkah-langkah yang direncanakan untuk menjawab pertanyaan penelitian dan menghasilkan informasi yang akurat sesuai dengan tujuan penelitian. Rangkaian langkah-langkah penelitian yang digunakan dalam penelitian ini pada Gambar 1 di bawah ini.



Gambar 3. 1 Metodologi Penelitian

3.1 Pengumpulan Data

Penelitian ini menggunakan empat dataset sentimen dari Twitter, yaitu sentimen terhadap pengangkatan Kaesang sebagai Ketua Umum PSI Data Kaesang v1 (KS) Kaesang v2 (KS1), sentimen terhadap program vaksin COVID-19 (CV), dan sentimen terhadap topik lain (OP).(s4gustian, 2024)

Dataset Kaesang terbagi menjadi dua set data, yaitu Data Kaesang v1 dan Kaesang v2, yang masing-masing berisi 300 tweet. Peneliti bebas

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

memilih salah satu atau bahkan kedua dataset tersebut. Selain itu, peneliti juga dapat memperbanyak jumlah sampel data dengan menggunakan dataset lain seperti Data Covid dan Opentopic.

Dataset Kaesang dikumpulkan menggunakan proses scrapping. Sejumlah 2309 tweet yang terkait dengan kata kunci "Kaesang PSI" berhasil dikumpulkan dalam rentang waktu antara tanggal 25 September 2023 hingga 3 Oktober 2023. Selanjutnya, data tersebut diberi label positif, netral, atau negatif melalui proses crowdsourcing. Setiap tweet dilabel oleh 4 orang anotator, dan label ditentukan berdasarkan mayoritas suara. Apabila ada tweet dengan label yang berbeda dan tanpa mayoritas yang dominan, tweet tersebut dihapus dan dianggap tidak valid.. Sedangkan untuk data uji, terdapat 924 data, dengan label gold standard-nya disimpan di server leaderboard.³

Dataset COVID dan dataset OpenTopic menggunakan data dari penelitian sebelumnya oleh Sahbuddin & Agustian (2022), Kusairi & Agustian (2022), dan Ash Shiddicky & Surya Agustian (2022), yang terdiri dari 8000 tweet dengan label positif, negatif, dan netral. Data COVID dan OpenTopic ini digunakan untuk menambah data pelatihan bagi dataset Kaesang. Model paling optimal dari hasil penelitian ini diterapkan pada data uji tersebut, dan hasil prediksinya dikirimkan ke sistem leaderboard untuk mendapatkan skor evaluasi.

Tabel 2. 1 Dataset dalam penelitian

Dataset	Jumlah Sentimen
Data Kaesang v1	300
Data Kaesang v2	300
Data Covid	8000
Data Opentopic	15000

³Agustian.

Hak Cipta Dilindungi Undang-Undang

1. Diarangi mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

3.2 Pembagian Data

Pada tahap pembagian data, data yang tersedia untuk pelatihan terdiri dari 300 data Kaesang dan 8000 data eksternal terkait Covid, dengan label positif, negatif, dan netral. Sementara itu, data uji berjumlah 924 data, yang label gold standard-nya disimpan di server leaderboard (Agustian et al., 2024). Model yang paling optimal dari hasil penelitian ini diterapkan pada data uji tersebut, dan hasil prediksinya diunggah ke sistem leaderboard untuk mendapatkan skor evaluasi.

3.3 Preprocessing

Sebelum data dimanfaatkan dalam langkah-langkah berikutnya, dilakukan tahap pra-pemrosesan untuk mengubah data mentah menjadi format yang lebih dimengerti oleh sistem dengan tujuan meningkatkan hasil dari proses klasifikasi sentimen. Tahap ini mencakup normalisasi, penanganan negasi dan penghapusan stopword menggunakan kamus kustom yang telah saya buat, yang krusial dalam konteks text preprocessing dengan dataset terbatas sebanyak 300 data. Metode Manual Annotation dan Koreksi yang saya terapkan memungkinkan saya untuk secara langsung meninjau dan memperbaiki setiap entri dalam kamus dengan cermat, memastikan akurasi dan relevansi kamus sesuai dengan konteks dataset. Dengan fokus yang teliti terhadap detail dan kualitas kamus, pendekatan ini memastikan alat text preprocessing yang dikembangkan memberikan nilai tambah yang signifikan dalam analisis data, meningkatkan akurasi proses text mining, serta mendukung interpretasi yang lebih mendalam terhadap hasil analisis yang dihasilkan.

3.3.1 Cleaning

Proses cleaning teks yang saya lakukan menggunakan beberapa langkah dengan regex dan penanganan karakter unicode. Pertama, menggunakan regex untuk mendeteksi dan menghapus

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

tautan (URL) dari teks. Selanjutnya, regex digunakan untuk mengganti mention (misalnya @username) dengan tag "USER" dan menggabungkan tag "USER" yang berulang menjadi satu "USER" saja. Kemudian, menggunakan regex untuk menghapus semua angka dari teks. Setelah itu, teks di-encode dan di-decode untuk menangani karakter unicode, dan jika terjadi kesalahan, teks tetap seperti semula.

Tabel 2. 2 Contoh Cleaning

Sebelum Cleaning	Sesudah Cleaning
@psi_id @kaesangp Asli ini re-marketing @psi_id ke ibu-ibu dan wanita bagus banget... Lgsg salfok sama baju imutnya kaesang. Ini kena banget dan politik jd adeeemmmmm banget... Ngga ada kalimat kasar, vulgar, caci maki, dsb... Mantap...	USER USER Asli ini re-marketing USER ke ibu-ibu dan wanita bagus banget... Lgsg salfok sama baju imutnya kaesang. Ini kena banget dan politik jd adeeemmmmm banget... Ngga ada kalimat kasar, vulgar, caci maki, dsb... Mantap...

3.3.2 Case Folding

Case Folding merupakan proses di mana semua huruf dalam teks diubah menjadi huruf kecil. Tujuannya adalah untuk menghilangkan perbedaan yang mungkin muncul berdasarkan kapitalisasi dalam interpretasi teks.

Tabel 2. 3 Contoh Case Folding

Sebelum Case Folding	Sesudah Case Folding
USER USER Asli ini re-marketing USER ke ibu-ibu dan wanita bagus banget... Lgsg salfok sama baju imutnya kaesang. Ini kena banget dan politik jd adeeemmmmm banget... Ngga ada kalimat kasar, vulgar, caci maki, dsb... Mantap...	user user asli ini re-marketing user ke ibu-ibu dan wanita bagus banget... lgsg salfok sama baju imutnya kaesang. ini kena banget dan politik jd adeeemmmmm banget... ngga ada kalimat kasar, vulgar, caci maki, dsb... mantap...

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

3.3.3 Tokenizing

Tokenisasi adalah proses di mana teks data tweet dibagi menjadi bagian-bagian kecil yang disebut token. Dalam implementasi saya, saya menggunakan `word_tokenize` dari NLTK untuk memecah teks menjadi token-token (kata-kata).

Tabel 2. 4 Contoh Tokenizing

Sebelum Tokenizing	Sesudah Tokenizing
user user asli ini re-marketing user ke ibu-ibu dan wanita bagus banget... lgsg salfok sama baju imutnya kaesang. ini kena banget dan politik jd adeeemmmm banget... ngga ada kalimat kasar, vulgar, caci maki, dsb... mantap...	['user', 'user', 'asli', 'ini', 're-marketing', 'user', 'ke', 'ibu-ibu', 'dan', 'wanita', 'bagus', 'banget', 'lgsg', 'salfok', 'sama', 'baju', 'imutnya', 'kaesang', 'ini', 'kena', 'banget', 'dan', 'politik', 'jd', 'adeeemmmm', 'banget', 'ngga', 'ada', 'kalimat', 'kasar', 'vulgar', 'caci', 'maki', 'dsb', 'mantap',]

3.3.4 Normalisasi

Dalam proses normalisasi teks saya menggunakan kamus sendiri (normalisasi_dict), saya mulai dengan memuat kamus yang berisi pasangan kata kunci dan nilai untuk mengganti kata kunci dalam teks. Selanjutnya, setiap token dalam teks diperiksa terhadap kamus normalisasi untuk penggantian dengan nilai yang sesuai. Ini membantu menstandarisasi kata-kata yang serupa namun berbeda, meningkatkan konsistensi dalam analisis seperti analisis sentimen atau klasifikasi teks. Dengan kamus normalisasi yang saya buat, saya memiliki kontrol atas cara kata-kata dalam teks diubah ke dalam bentuk standar yang telah ditetapkan sebelumnya untuk tujuan analisis.

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

Tabel 2. 5 Contoh Normalisasi

Sebelum Normalisasi	Sesudah Normalisasi
['user', 'user', 'asli', 'ini', 're-marketing', 'user', 'ke', 'ibu-ibu', 'dan', 'wanita', 'bagus', 'banget', 'lgsg', 'salfok', 'sama', 'baju', 'imutnya', 'kaesang', 'ini', 'kena', 'banget', 'dan', 'politik', 'jd', 'adeemmmmm', 'banget', 'ngga', 'ada', 'kalimat', 'kasar', 'vulgar', 'caci', 'maki', 'dsb', 'mantap',]	['user', 'user', 'asli', 'ini', 'remarketing', 'user', 'ke', 'ibu-ibu', 'dan', 'wanita', 'bagus', 'sangat', 'langsung', 'salah fokus', 'dengan', 'baju', 'imutnya', 'kaesang', 'ini', 'terkena', 'sangat', 'dan', 'politik', 'jadi', 'adem', 'sangat', 'tidak', 'ada', 'kalimat', 'kasar', 'vulgar', 'caci', 'maki', 'dan sebagainya', 'mantap',]

3.5 Negation Handling

Dalam penanganan negasi pada teks, saya menggunakan kamus negasi sendiri (`negation_dict`). Kamus ini berisi kata-kata seperti "tidak", "bukan", "tak", yang menandakan negasi. Ketika kata-kata ini muncul dalam teks, saya mengikuti langkah-langkah khusus untuk mengubah kata-kata yang mengikutinya. Misalnya, kata "suka" setelah "tidak" dapat diubah menjadi "suka_negasi". Hal ini membantu meningkatkan keakuratan dalam analisis teks, terutama dalam konteks analisis sentimen atau klasifikasi teks. Dengan menggunakan kamus negasi yang saya buat sendiri, saya dapat mengontrol bagaimana teks diinterpretasikan setelah munculnya kata-kata negasi, sehingga meningkatkan konsistensi dan keakuratan hasil analisis.

Tabel 2. 6 contoh negation handling

Sebelum Negation Handling	Sesudah Negation Handling
['user', 'user', 'asli', 'ini', 'remarketing', 'user', 'ke', 'ibu-ibu', 'dan', 'wanita', 'bagus', 'sangat', 'langsung', 'salah fokus', 'dengan', 'baju', 'imutnya', 'kaesang', 'ini', 'terkena', 'sangat', 'dan', 'politik', 'jadi', 'adem', 'sangat', 'tidak', 'ada', 'kalimat', 'kasar', 'vulgar', 'caci', 'maki', 'dan sebagainya', 'mantap',]	['user', 'user', 'asli', 'ini', 'remarketing', 'user', 'ke', 'ibu-ibu', 'dan', 'wanita', 'bagus', 'sangat', 'langsung', 'salah fokus', 'dengan', 'baju', 'imutnya', 'kaesang', 'ini', 'terkena', 'sangat', 'dan', 'politik', 'jadi', 'adem', 'sangat', 'tidak', 'ada', 'kalimat', 'kasar', 'vulgar', 'caci', 'maki', 'dan sebagainya', 'mantap',]

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

3.3.6 Stopword Removal

Pada langkah ini, tweet yang sudah dibersihkan akan dihilangkan karakter, tanda baca, dan kata-kata umum yang tidak relevan atau informatif (Rustiana & Rahayu, 2017b). Dalam penghapusan stopword pada teks, saya menggunakan kamus stopword sendiri (`stopwords_set`). Kamus ini digunakan untuk mengenali kata-kata umum seperti "dan", "atau", "yang", yang dianggap tidak memberikan nilai signifikan dalam analisis teks. Prosesnya sederhana: setiap kata dalam teks diperiksa terhadap kamus stopword, dan jika kata tersebut terdaftar sebagai stopword, kata tersebut dihapus dari teks. Langkah ini membantu membersihkan teks dari kata-kata yang tidak relevan sehingga memungkinkan fokus pada kata-kata kunci yang lebih bermakna dalam analisis teks. Dengan kamus stopword yang saya buat sendiri, saya memiliki kendali atas proses penghapusan stopword sesuai dengan kebutuhan dan tujuan analisis yang sedang dilakukan.

Tabel 2. 7 Contoh Stopword Removal

Sebelum Stopword Removal	Sesudah Stopword Removal
['user', 'user', 'asli', 'ini', 'remarketing', 'user', 'ke', 'ibu-ibu', 'dan', 'wanita', 'bagus', 'sangat', 'langsung', 'salah fokus', 'dengan', 'baju', 'imutnya', 'kaesang', 'ini', 'terkena', 'sangat', 'dan', 'politik', 'jadi', 'adem', 'sangat', 'tidak', 'ada', 'kalimat', 'kasar', 'vulgar', 'caci', 'maki', 'dan sebagainya', 'mantap',]	['user', 'user', 'asli', 'remarketing', 'user', 'ibu-ibu', 'wanita', 'bagus', 'sangat', 'langsung', 'salah fokus', 'baju', 'imutnya', 'kaesang', 'terkena', 'sangat', 'politik', 'adem', 'sangat', 'tidak', 'ada', 'kalimat', 'kasar', 'vulgar', 'caci', 'maki', 'dan sebagainya', 'mantap']

3.3.7 Stemming

Stemming adalah proses mengubah kata menjadi bentuk dasarnya dengan menghapus imbuhan atau afiksasi pada kata dalam dokumen, atau mengubah kata kerja menjadi kata benda. Sebagai contoh, kata "dihilangkan" setelah menghapus imbuhan di- dan -kan menjadi "hilang" (Wardani, 2015). Dalam melakukan stemming pada teks, saya

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

menggunakan algoritma `PorterStemmer` dari library `nltk`. Stemming adalah proses untuk mengubah kata-kata dalam teks menjadi bentuk dasar atau akar kata dengan menghapus akhiran kata seperti "-s", "-es", "-ed", "-ing", dan sejenisnya. Contohnya, kata-kata seperti "imutnya" disederhanakan menjadi "imut". Dengan menggunakan `PorterStemmer`, saya memperbaiki konsistensi representasi kata-kata dalam teks, meskipun bentuk dasarnya tidak selalu sesuai dengan kata aslinya secara gramatikal atau semantis. Stemming adalah langkah penting dalam pre-processing teks untuk tujuan seperti information retrieval dan analisis teks.

Tabel 2. 8 Contoh Stemming

Sebelum Stemming	Sesudah Stemming
['user', 'user', 'asli', 'remarketing', 'user', 'ibu-ibu', 'wanita', 'bagus', 'sangat', 'langsung', 'salah fokus', 'baju', 'imutnya', 'kaesang', 'terkena', 'sangat', 'politik', 'adem', 'sangat', 'tidak', 'ada', 'kalimat', 'kasar', 'vulgar', 'caci', 'maki', 'dan sebagainya', 'mantap']	['user', 'user', 'asli', 'remarketing', 'user', 'ibu-ibu', 'wanita', 'bagus', 'sangat', 'langsung', 'salah fokus', 'baju', 'imut', 'kaesang', 'kena', 'sangat', 'politik', 'adem', 'sangat', 'tidak', 'ada', 'kalimat', 'kasar', 'vulgar', 'caci', 'maki', 'dan sebagainya', 'mantap']

3.1 Preprocessing Teks dengan Kamus Khusus

Pada penelitian ini, preprocessing dilakukan dengan menggunakan tiga kamus khusus yang dibuat sendiri, yaitu: kamus normalisasi, kamus handling negasi, dan kamus stopword. Langkah-langkah preprocessing yang dilakukan adalah sebagai berikut:

- Normalisasi Teks: Menggunakan kamus normalisasi, setiap kata dalam teks yang tidak baku diubah menjadi kata baku.
- Handling Negasi: Mengidentifikasi kata negasi dalam teks dan mengubah frasa setelah kata negasi sesuai dengan kamus handling negasi.
- Penghapusan Stopword: Menghapus kata-kata yang termasuk dalam kamus stopword dari teks.

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

3.5 Skema preprocessing

Skema pada text preprocessing adalah serangkaian aturan atau langkah-langkah yang ditetapkan untuk mengelola teks mentah sebelum dijalankan melalui proses klasifikasi lebih lanjut. Setiap kombinasi dari skema di atas menawarkan pendekatan yang berbeda dalam persiapan teks untuk klasifikasi. menerapkan skema yang sesuai, proses text preprocessing dapat meningkatkan akurasi dan relevansi hasil analisis yang dihasilkan.

3.6 Featuring Weighting (TF.IDF)

TF.IDF adalah teknik pembobotan yang banyak digunakan dan terkenal dengan tingkat akurasi dan recall yang tinggi, terutama dalam bidang information retrieval. Keunggulan TF.IDF terletak pada pengukurannya terhadap frekuensi kemunculan kata dalam sebuah dokumen; semakin sering kata muncul, semakin besar kontribusinya. Sebaliknya, jika kata tersebut muncul di banyak dokumen, kontribusinya akan menurun. TF.IDF terdiri dari dua komponen utama: Term Frequency (TF) dan Inverse Document Frequency (IDF) (Yutika et al., 2021).

Pada tahap ini, setiap dokumen dalam korpus diubah menjadi representasi vektor menggunakan metode TF.IDF dengan menggunakan library sklearn tfidf_vectorizer.

TF.IDF Vectorizer berfungsi dengan memberikan bobot pada setiap kata dalam dokumen berdasarkan seberapa sering kata tersebut muncul dalam dokumen tertentu dan di seluruh korpus dokumen. Teknik ini sangat berguna dalam mengidentifikasi kata kunci yang paling relevan dan informatif dalam sebuah dokumen atau kumpulan dokumen. TF.IDF Vectorizer menghitung bobot TF.IDF dengan mengalikan nilai Term Frequency dengan Inverse Document Frequency, sehingga menghasilkan bobot yang lebih tinggi untuk kata-kata yang jarang muncul dalam korpus tetapi sering muncul dalam dokumen tertentu

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

3.7 Klasifikasi Naive Bayes Classifier

Naive Bayes Classifier adalah model klasifikasi yang sederhana dan efektif, terutama untuk klasifikasi teks. Model ini mendasarkan diri pada dasar Bayesian Network, di mana semua atribut dianggap independen dengan nilai kelas variabel. Kelebihan dari Naive Bayes Classifier mencakup kesederhanaannya, kecepatannya, dan tingkat akurasi yang tinggi (Indrayuni, 2019).

Pada tahap ini, kita mengajari model Naive Bayes Classifier dengan menggunakan data latih. Model ini menghitung seberapa sering setiap kelas (misalnya, perasaan atau kategori) muncul sebelumnya dan seberapa sering setiap kata muncul dalam setiap kelas. Setelah model terlatih, kita menggunakan model ini untuk memprediksi kelas pada data uji. Data uji juga mengalami persiapan yang sama dengan data latih, yaitu melalui proses pembobotan TF.IDF.

Setelah mendapatkan bobot TF.IDF, langkah selanjutnya adalah menggunakan Naive Bayes Classifier untuk klasifikasi sentimen. Berikut langkah-langkah implementasinya:

- Memanggil model Naive bayes untuk klasifikasi, yaitu multinomialNB, dan disimpan dalam variabel model.
- Melakukan training model NB ini terhadap data input tweet yang telah diubah menjadi vektor TF.IDF.
- Melakukan klasifikasi data testing yang telah diubah menjadi vektor TF.IDF menggunakan model NB hasil proses training di atas.

3.8 Evaluasi

Mengevaluasi model klasifikasi menggunakan confusion matrix adalah langkah penting dalam memahami seberapa baik kinerja model, terutama dalam hal prediksi untuk kelas-kelas yang berbeda. Matriks kebingungan adalah tabel yang digunakan untuk menggambarkan kinerja

model klasifikasi. Matriks ini sangat berguna untuk klasifikasi biner, namun dapat diperluas untuk klasifikasi multi-kelas juga. Matriks ini membandingkan nilai target aktual dengan nilai yang diprediksi oleh model, hasil klasifikasi yang dihitung untuk tiga kelas (Positif, Negatif, Netral) adalah sebagai berikut: "Aktual" merujuk pada nilai yang ditetapkan oleh manusia sebagai standar emas, sedangkan "prediksi" adalah hasil dari model. "True" (T) menunjukkan prediksi yang benar sesuai dengan nilai aktual, sedangkan "False" (F) menunjukkan prediksi yang salah.

3.9 Kesimpulan dan Saran

Di tahap ini, akan disimpulkan hasil dari penelitian ini dan memberikan saran untuk penelitian berikutnya yang akan dilakukan oleh peneliti lain.

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

BAB V

PENUTUP

5.1 Kesimpulan

Berdasarkan tahapan penelitian dan pengujian dengan metode Naïve Bayes Classifier, dapat disimpulkan bahwa metode ini dapat digunakan dalam proses klasifikasi sentimen masyarakat di Twitter terhadap Kaesang Pangarep sebagai ketua umum Partai Solidaritas Indonesia. Dari 4 pengujian yang berbeda, diperoleh model terbaik dengan nilai akurasi tertinggi yaitu 60,35% dan F1 Score sebesar 52,07% pada perbandingan data training dan testing 80:20 menggunakan penggabungan data Kaesang v1, data Kaesang v2, data covid dan data open topic. Pengujian menunjukkan bahwa menggabungkan data tersebut menghasilkan menghasilkan akurasi sebesar 60,35% dan F1 Score sebesar 52,07%, namun perlu diperhatikan bahwa jika jumlah data dikurangi, nilai F1 Score dan akurasi cenderung lebih rendah. Saran untuk penelitian berikutnya adalah menggunakan data yang lebih banyak dan bervariasi untuk hasil yang lebih dapat diandalkan.

5.2 Saran

Saran untuk penelitian ini adalah mengembangkan optimasi lain yang dapat dilakukan dengan menggunakan Naive Bayes, seperti dengan memanfaatkan fitur bag-of-words dan word embeddings.

Hak Cipta Dilindungi Undang-Undang

1. Diarangi mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:

- a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
- b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

DAFTAR PUSTAKA

- Ahyasta Dirgantara, D. M. (2023). *Kaesang Pangarep Resmi Jadi Ketua Umum PSI*. Kompas. <https://nasional.kompas.com/read/2023/09/25/19504431/kaesang-pangarep-resmi-jadi-ketua-umum-psi>
- Ahustian, S., Syah, M. I., Fatiara, N., & Abdillah, R. (2024). New Directions in Text Classification Research: Maximizing The Performance of Sentiment Classification from Limited Data. In *ArXiv*. <https://arxiv.org/abs/2407.05627>
- Andika, L. A., Azizah, P. A. N., & Respatiwan, R. (2019). Analisis Sentimen Masyarakat terhadap Hasil Quick Count Pemilihan Presiden Indonesia 2019 pada Media Sosial Twitter Menggunakan Metode Naive Bayes Classifier. *Indonesian Journal of Applied Statistics*, 2(1), 34. <https://doi.org/10.13057/ijas.v2i1.29998>
- Astari, N. M. A. J., Dewa Gede Hendra Divayana, & Gede Indrawan. (2020). Analisis Sentimen Dokumen Twitter Mengenai Dampak Virus Corona Menggunakan Metode Naive Bayes Classifier. *Jurnal Sistem Dan Informatika (JSI)*, 15(1), 27–29. <https://doi.org/10.30864/jsi.v15i1.332>
- Darwis, D., Shintya Pratiwi, E., Ferico, A., & Pasaribu, O. (2020). PENERAPAN ALGORITMA SVM UNTUK ANALISIS SENTIMEN PADA DATA TWITTER KOMISI PEMBERANTASAN KORUPSI REPUBLIK INDONESIA. In *Jurnal Ilmiah Edutic* (Vol. 7, Issue 1).
- Darwis, D., Siskawati, N., & Abidin, Z. (2021). Penerapan Algoritma Naive Bayes Untuk Analisis Sentimen Review Data Twitter Bmkg Nasional. *Jurnal Tekno Kompak*, 15(1), 131. <https://doi.org/10.33365/jtk.v15i1.744>
- Diwandanu, M. T., & Wisudawati, L. M. (2023). Analisis Sentimen Terhadap Twit Maxim Pada Twitter Menggunakan R Programming Dan K Nearest Neighbors. *Jurnal Ilmiah Informatika Komputer*, 28(1), 1–16. <https://doi.org/10.35760/ik.2023.v28i1.7909>
- Ginting, H. S., Lhaksmana, K. M., & Murdiansyah, D. T. (2018). Klasifikasi Sentimen Terhadap Bakal Calon Gubernur Jawa Barat 2018 di Twitter. *E-Proceeding of Engineering*, 5(1), 1793–1802.
- Gupta, R., Sahu, S., Espy-Wilson, C., & Narayanan, S. (2018). Semi-Supervised and Transfer Learning Approaches for Low Resource Sentiment Classification. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2018-April*, 5109–5113. <https://doi.org/10.1109/ICASSP.2018.8461414>
- Indrayuni, E. (2019). Klasifikasi Text Mining Review Produk Kosmetik Untuk Teks Bahasa Indonesia Menggunakan Algoritma Naive Bayes. *Jurnal Khatulistiwa Informatika*, 7(1), 29–36. <https://doi.org/10.31294/jki.v7i1.1>
- Narayanan, V., Arora, I., & Bhatia, A. (2013). Fast and accurate sentiment classification using an enhanced Naive Bayes model. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8206 LNCS, 194–201. https://doi.org/10.1007/978-3-642-41278-3_24
- Normawati, D., & Prayogi, S. A. (2021). Implementasi Naive Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter. *Jurnal Sains Komputer & Informatika (J-SAKTI)*, 5(2), 697–711.
- Nurdiansyah, Y., Rahman, F., & Pandunata, P. (2021). Analisis Sentimen Opini Publik Terhadap Undang-Undang Cipta Kerja pada Twitter Menggunakan Metode Naive Bayes Classifier. *Prosiding Seminar Nasional Sains Teknologi Dan Inovasi Indonesia (SENASTINDO)*, 3(November), 201–212.

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:

- a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
- b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

- <https://doi.org/10.54706/senastindo.v3.2021.158>
- Rustiana, D., & Rahayu, N. (2017a). Analisis Sentimen Pasar Otomotif Mobil: Tweet Twitter Menggunakan Naïve Bayes. *Simetris: Jurnal Teknik Mesin, Elektro Dan Ilmu Komputer*, 8(1), 113–120. <https://doi.org/10.24176/simet.v8i1.841>
- Rustiana, D., & Rahayu, N. (2017b). Analisis Sentimen Pasar Otomotif Mobil: Tweet Twitter Menggunakan Naïve Bayes. *Simetris: Jurnal Teknik Mesin, Elektro Dan Ilmu Komputer*, 8(1), 113–120. <https://doi.org/10.24176/simet.v8i1.841>
- s4gustian. (2024). *Dataset for Sentiment Classification Task in Bahasa Indonesia*. Github. https://github.com/s4gustian/Small_DataSet_Sentiment_Classification
- Septianingrum, F., Jaman, J. H., & Enri, U. (2021). Analisis Sentimen Pada Isu Vaksin Covid-19 di Indonesia dengan Metode Naive Bayes Classifier. *Jurnal Media Informatika Budidarma*, 5(4), 1431. <https://doi.org/10.30865/mib.v5i4.3260>
- Tswari, S., & Sinha, A. (2020). Sentiment Analysis of Facebook Data using Machine Learning. *International Journal of Innovative Research in Applied Sciences and Engineering*, 4(4), 735–742. <https://doi.org/10.29027/ijirase.v4.i4.2020.735-742>
- Verawati, I., & Audit, B. S. (2022). Algoritma Naïve Bayes Classifier Untuk Analisis Sentiment Pengguna Twitter Terhadap Provider By.u. *Jurnal Media Informatika Budidarma*, 6(3), 1411. <https://doi.org/10.30865/mib.v6i3.4132>
- Vitandy, S. W. U., Supianto, A. A., & Bachtiar, F. A. (2019). Analisis Sentimen Evaluasi Kinerja Dosen menggunakan Term Frequency- Inverse Document Frequency dan Naïve Bayes Classifier. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, 3(6), 6080–6088. <https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/5645>
- Wahyuningsih, S., & Utari, D. R. (2018). Perbandingan Metode K-Nearest Neighbor , Naive Bayes dan Decision Tree untuk Prediksi Kelayakan Pemberian Kredit. *Konferensi Nasional Sistem Informasi 2018 STMIK Atma Luhur Pangkalpinang, 8 – 9 Maret 2018*, 619–623.
- Wardani, S. (2015). Analisis Sentimen Data Presiden Jokowi Dengan Preprocessing Normalisasi Dan Stemming Menggunakan Metode Naive Bayes Dan Svm. *Jurnal Dinamika Informatika*, 5(November), 1–13.
- Yatika, C. H., Adiwijaya, A., & Faraby, S. Al. (2021). Analisis Sentimen Berbasis Aspek pada Review Female Daily Menggunakan TF.IDF dan Naïve Bayes. *Jurnal Media Informatika Budidarma*, 5(2), 422. <https://doi.org/10.30865/mib.v5i2.2845>
- Zafira, D. F., Rahayudi, B., & Indriati, I. (2021). Analisis Sentimen Kebijakan Kampus Merdeka Menggunakan Naive Bayes dan Pembobotan TF.IDF Berdasarkan Komentar pada Youtube. *Jurnal Sistem Informasi, Teknologi Informasi, Dan Edukasi Sistem Informasi*, 2(1), 55–63. <https://doi.org/10.25126/justsi.v2i1.24>
- Zuhdi, A. M., Utami, E., & Raharjo, S. (2019). *Abdul Malik Zuhdi 1* , *Ema Utami 2*) , *Suwanto Raharjo 3*) 3. 5, 1–7.

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
 - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.



DAFTAR RIWAYAT HIDUP

Nama : Rian Delvy Juraidi
 Jenis Kelamin : Laki-Laki
 Tempat, Tanggal Lahir : Melai, 26 Juli 2002
 Status : Belum Menikah
 Warga Negara : Indonesia
 Golongan darah : A
 Anak ke : 1
 Jumlah Saudara : 3
 Alamat : Jl. Parit pisang, Desa Melai Kecamatan Rangsang Barat, Kabupaten Kepulauan Meranti, Riau
 Email : 11950111735@students.uin-suska.ac.id

Riwayat Pendidikan

2007-2013 : SDN Negeri 8 Melai
 2013-2016 : MTS Al-Mutahidah Melai
 2016-2019 : SMK Negeri 1 Tebing Tinggi