PENERAPAN METODE *NAIVE BAYES* DALAM SENTIMEN ANALISIS TERHADAP PROGRAM MAKAN SIANG & SUSU GRATIS DI MEDIA *X*

SKRIPSI

Oleh:

JONATHAN DE AGUSTINO WULLA

20211000006

PROGRAM STUDI TEKNIK INFORMATIKA

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS BUDDHI DHARMA

TANGERANG

2025

PENERAPAN METODE *NAIVE BAYES* DALAM SENTIMEN ANALISIS TERHADAP PROGRAM MAKAN SIANG & SUSU GRATIS DI MEDIA *X*

SKRIPSI

Diajukan sebagai salah satu syarat untuk memperoleh gelar Sarjana pada

Program Studi Teknik Informatika Fakultas Sains dan Teknologi

Universitas Buddhi Dharma

Jenjang Pendidikan Strata 1

Oleh:

JONATHAN DE AGUSTINO WULLA

20211000006

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS BUDDHI DHARMA

TANGERANG

2025

LEMBAR PERSEMBAHAN

"When you feel like giving up, remember why you started."

Dengan mengucap puji syukur kepada Tuhan Yang Maha Esa, SKRIPSI ini kupersembahkan untuk:

1. Bapak Eusebius Bodi dan Ibu Marni Naibaho tercinta yang telah membesarkan aku dan selalu membimbing, mendukung, memotivasi, memberi apa yang terbaik bagiku serta selalu mendoakan aku untuk meraih kesuksesanku.
2. Kakak dan adik-adikku yang telah memberikan dukungan semangat serta dorongan yang senantiasa diberikan.
3. Bapak dosen pembimbing, Bapak Rino, M.Kom saya sampaikan rasa hormat dan terima kasih yang mendalam atas bimbingan, kesabaran, dan arahan yang begitu berarti selama proses penyusunan karya ini.
4. Nevada Livrida Ines terima kasih atas segala bentuk dukungan, semangat, dan perhatian yang tak pernah putus. Kehadiranmu menjadi salah satu sumber kekuatan terbesar dalam perjalanan ini.
5. Teman penulis Dicky Afandi Rajagukguk yang selalu menemani penulis dan banyak membantu dalam suka maupun duka selama 4 tahun perjalanan perkuliahan. Terima kasih atas kebersamaan, dukungan, dan semangat yang tak ternilai.

LEMBAR PERNYATAAN KEASLIAN SKRIPSI

Yang bertanda tangan di bawah ini.

NIM : 20211000006

Nama : Jonathan De Agustino Wulla

Jenjang Studi : Strata 1

Program Studi : Teknik Informatika

Peminatan : Data Science

Dengan ini saya menyatakan bahwa:

1. Skripsi ini adalah asli dan belum pernah diajukan untuk mendapat gelar akademik Sarjana atau kelengkapan studi, baik di Universitas Buddhi Dharma maupun di Pereguruan Tinggi lainnya.
2. Skripsi ini saya buat sendiri tanpa bantuan dari pihak lain, kecuali arahan dosen pembimbing.
3. Dalam Skripsi ini tidak terdapat karya atau pendapat yang telah ditulis atau dipublikasikan orang lain, kecuali secara tertulis dengan jelas dan dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan dicantumkan daftar pustaka.
4. Dalam Skripsi ini tidak terdapat pemalsuan (kebohongan), seperti buku, artikel, jurnal, data sekunder, pengolahan data, dan pemalsuan tanda tangan dosen atau Ketua Program Studi Universitas Buddhi Dharma yang dibuktikan dengan keasliannya
5. Lembar pernyataan ini saya buat dengan sesungguhnya, tanpa paksaan dan apabila dikemudian hari atau pada waktu lainnya terdapat penyimpangan dan ketidak benaran dalam pernyataan ini, saya bersedia menerima sanksi akademik berupa pencabutan gelar akademik yang telah saya peroleh karena Skripsi ini serta sanksi lai nnya sesuai dengan peraturan dan norma yang berlaku.

Tangerang, 05 Agustus 2025

Yang membuat pernyataan,

Jonathan De Agustino Wulla

20211000006

LEMBAR PERSETUJUAN PUBLIKASI KARYA ILMIAH

Yang bertanda tangan di bawah ini.

NIM : 20211000006

Nama : Jonathan De Agustino Wulla

Jenjang Studi : Strata 1

Program Studi : Teknik Informatika

Peminatan : Data Science

Dengan ini menyetujui untuk memberikan ijin kepada pihak Universitas Buddhi Dharma, Hak Bebas Royalti Non – Eksklusif (Non-exclusive Royalty-Free Right) atas karya ilmiah kami yang berjudul:PENERAPAN METODE NAIVE BAYES DALAM SENTIMEN ANALISIS TERHADAP PROGRAM MAKAN SIANG & SUSU GRATIS DI MEDIA X , beserta alat yang diperlukan (apabila ada). Dengan Hak Bebas Royalti Non – Eksklusif ini pihak Universitas Buddhi Dharma berhak menyimpan, mengalih-media atau format-kan, mengelolanya dalam pangkalan data (database), mendistribusikannya, dan menampilkan atau mempublikasikannya di internet atau media lain untuk kepentingan akademis tanpa perlu meminta ijin dari saya selama tetap mencantumkan nama saya sebagai penulis pertama atau pencipta karya ilmiah tersebut. Saya bersedia untuk menanggung secara pribadi, tanpa melibatkan pihak Universitas Buddhi Dharma, segala bentuk tuntutan hukum yang timbul atas pelanggaran Hak Cipta dalam karya ilmiah saya ini. Demikian pernyataan ini saya buat dengan sebenarnya.

Tangerang, 05 Agustus 2025

Yang membuat pernyataan,

Jonathan De Agustino Wulla

20211000006

LEMBAR PENGESAHAN PEMBIMBING

PENERAPAN METODE NAIVE BAYES DALAM SENTIMEN ANALISIS TERIIADAP PROGRAM MAKAN SIANG & SUSU GRATIS Dl MEDIA X

Dibuat Oleh:

NIM : 20211000006

Nama : Jonathan De Agustino Wulla

Telah disetujui untuk dipertahankan di hadapan Tim Penguji Ujian Komprehensif

Program Studi Teknik Informatika

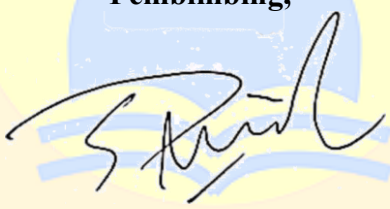
Peminatan Data Science

Tahun Akademik 2024/2025

Disahkan oleh,

Tangerang, 16 Juli 2025

Pembimbing,



Rino M.Kom

NUPTK. 5852763664130372

LEMBAR PENGESAHAN SKRIPSI

PENERAPAN METODE NAIVE BAYES DALAM SENTIMEN ANALISIS TERHADAP PROGRAM MAKAN SIANG & SUSU GRATIS DI MEDIA X

Dibuat Oleh:

NIM : 20211000006

Nama : Jonathan De Agustino Wulla

Telah disetujui untuk dipertahankan di hadapan Tim Penguji Ujian Komprehensif

Program Studi Teknik Informatika

Peminatan Data Science

Tahun Akademik 2024/2025

Disahkan oleh,

Tangerang, 05 Agustus 2025

Dekan,	**Ketua Program Studi,**
Dr. Yakub, M.Kom., M.M.	Hartana Wijaya, M.Kom.
NUPTK. 1836747648130172	NUPTK. 3844759660131192

LEMBAR PENGESAHAN TIM PENGUJI

Nama : Jonathan De Agustino Wulla

NIM : 20211000006

Fakultas : Sains dan Teknologi

Judul Skripsi : Penerapan Metode Naïve Bayes Dalam Sentimen Analisis Terhadap Program Makan Siang & Susu Gratis Di Media X

Dinyatakan LULUS setelah mempertahankan di depan Tim Penguji pada hari Selasa, 05 Agustus 2025.

Nama Penguji:		Tanda Tangan:
Ketua Sidang	: **Susanto Hariyanto, S.Kom.,M.Kom**	
	NUPTK. 2560764665130243	.....
Penguji I	: **Desiyanna Lasut, S.Kom., M.Kom**	
	NUPTK. 1534764665230253	.....
Penguji II	: **Rino, M.Kom**	
	NUPTK. 5852763664130372	.....

Mengetahui,

Dekan Fakultas Sains dan Teknologi

Dr. Yakub, M.Kom., M.M.

NUPTK. 1836747648130172

KATA PENGANTAR

Dengan mengucap Puji Syukur kepada Tuhan Yang Maha Esa, yang telah memberikan Rahmat dan karunia-Nya kepada penulis sehingga dapat menyusun dan menyelesaikan Skripsi ini dengan judul **Penerapan Metode Naive Bayes Dalam Sentimen Analisis Terhadap Program Makan Siang & Susu Gratis Di Media X**. Tujuan utama dari pembuatan Skripsi ini adalah sebagai salah satu syarat kelengkapan dalam menyelesaikan program pendidikan strata 1 program studi Teknik Informatika di Universitas Buddhi Dharma. Dalam penyusunan Skripsi ini Penulis banyak menerima bantuan dan dorongan baik moril maupun materil dari berbagai pihak, maka pada kesempatan ini penulis menyampaikan rasa terima kasih yang sebesar-besarnya kepada:

1. Ibu Dr. Limajatini, SE., M.M., B.K.P sebagai Rektor Universitas Buddhi Dharma.
2. Bapak Dr. Yakub, M.Kom., M.M. Dekan Fakultas Sains Dan Teknologi.
3. Bapak Hartana Wijaya, M.Kom sebagai Ketua Program Studi Teknik Informatika
4. Bapak Rino, M.Kom sebagai pembimbing yang telah membantu dan memberikan dukungan serta harapan untuk menyelesaikan penulisan Skripsi ini.
5. Orang tua dan keluarga yang selalu memberikan dukungan baik moril dan materil.
6. Teman-teman yang selalu membantu dan memberikan semangat.

Serta semua pihak yang terlalu banyak untuk disebutkan satu-persatu sehingga terwujudnya penulisan ini. Penulis menyadari bahwa penulisan Skripsi ini masih belum sempurna, untuk itu penulis mohon kritik dan saran yang bersifat membangun demi kesempurnaan penulisan di masa yang akan datang.

Akhir kata semoga Skripsi ini dapat berguna bagi penulis khususnya bagi para pembaca yang berminat pada umumnya

Tangerang, 05 Agustus 2025

Penulis

PENERAPAN METODE *NAIVE BAYES* DALAM SENTIMEN ANALISIS TERHADAP PROGRAM MAKAN SIANG & SUSU GRATIS DI *MEDIA X*

119 Halaman / 43 Tabel / 20 Gambar / 2 Lampiran

ABSTRAK

Selama kampanye pemilihan presiden Indonesia tahun 2024, Program Makan Siang dan Susu Gratis menjadi salah satu isu kebijakan utama yang mencuri perhatian publik, terutama di media sosial *X* (sebelumnya *Twitter*). Banyaknya opini publik yang beredar menciptakan tantangan tersendiri bagi para pembuat kebijakan dalam memahami persepsi masyarakat secara objektif dan berbasis data. Penelitian ini bertujuan untuk menganalisis sentimen publik terhadap program tersebut dengan menerapkan algoritma *Naïve Bayes* dan metode pembobotan *Term Frequency–Inverse Document Frequency* (*TF-IDF*). Sebanyak 964 *tweet* berbahasa Indonesia dikumpulkan secara otomatis menggunakan pemrograman di *Google Colab* melalui pustaka *tweet-harvest* dan autentikasi token. Data yang diperoleh kemudian diproses melalui tahapan praproses teks, meliputi pembersihan karakter, normalisasi huruf, tokenisasi, penghapusan *stopword*, dan *stemming* untuk menyiapkan data agar dapat dianalisis secara efektif. Selanjutnya, data diubah menjadi representasi numerik menggunakan *TF-IDF*, lalu diklasifikasikan ke dalam dua kategori sentimen, yaitu positif dan negatif. Proses klasifikasi dilakukan menggunakan algoritma *Naïve Bayes*, dengan skema evaluasi menggunakan *confusion matrix.* Hasil penelitian menunjukkan bahwa model mampu mencapai akurasi tinggi dalam mengidentifikasi arah sentimen publik. Temuan ini tidak hanya menggambarkan persepsi masyarakat terhadap program yang diusulkan, tetapi juga membuktikan efektivitas metode *machine learning* dalam menganalisis opini publik pada isu-isu sosial dan politik. Penelitian ini diharapkan dapat menjadi acuan dalam pengambilan keputusan berbasis data di masa mendatang.

Kata kunci: *Analisis Sentimen,Naïve Bayes,Program Makan Siang Dan Susu Gratis*

APPLICATION OF NAIVE BAYES METHOD IN SENTIMENT ANALYSIS OF FREE LUNCH & MILK PROGRAM IN MEDIA X

119 Pages / 43 Table / 20 images / 2 Libraries

ABSTRACT

During the 2024 Indonesian presidential election campaign, the Free Lunch and Milk Program became a major policy issue attracting public attention, particularly on social media platform X (formerly Twitter). The wide range of public opinions presented challenges for policymakers in understanding public perception objectively and based on data. This study aims to analyze public sentiment toward the program by applying the Naïve Bayes algorithm and the Term Frequency–Inverse Document Frequency (TF-IDF) weighting method. A total of 964 Indonesian-language tweets were automatically collected using programming in Google Colab through the tweet-harvest library and token authentication. The obtained data were then processed through text preprocessing stages, including character cleaning, case normalization, tokenization, stopword removal, and stemming to prepare the data for effective analysis. Next, the data were converted into numerical representation using TF-IDF, then classified into two sentiment categories: positive and negative. The classification process was carried out using the Naïve Bayes algorithm, with an evaluation scheme using a confusion matrix. The results showed that the model was able to achieve high accuracy in identifying the direction of public sentiment. These findings not only illustrate public perception of the proposed program but also demonstrate the effectiveness of machine learning methods in analyzing public opinion on social and political issues. This research is expected to serve as a reference for future data-driven decision-making.

Keyword: *Free Lunch and Milk Program,Naïve Baye,Sentiment Analysis*

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMPIRAN

BAB I

PENDAHULUAN

1.1 Latar Belakang

Saat ini, teknologi tumbuh dengan sangat cepat. Seiring berjalannya waktu, semakin mudah untuk mendapatkan informasi melalui berbagai portal dan platform media sosial. Dalam kebanyakan kasus, informasi yang tersedia di media sosial berupa teks dan disusun berdasarkan kontennya. Media sosial menjadi tempat orang berbagi pendapat mereka, baik dengan pujian, kritikan, ujaran kebencian, atau bahkan rumor yang dapat menyebabkan kontroversi. *Twitter*, sekarang dikenal sebagai *X*, adalah salah satu platform media sosial yang sering digunakan untuk berbicara. Survei yang dilakukan oleh Asosiasi Penyelenggara Jasa Internet Indonesia (APJII) menunjukkan bahwa jumlah pengguna internet di Indonesia diperkirakan akan mencapai 221.563.479 pada tahun 2024, naik sekitar 79,5% dari total populasi. Sementara itu, laporan dari We Are Social menunjukkan bahwa hingga Oktober 2023, terdapat sekitar 27,5 juta pengguna *Twitter* di Indonesia, menjadikannya negara dengan jumlah pengguna *Twitter* terbanyak keempat di dunia(Nursinggah et al., 2024).

Dalam Kamus Besar Bahasa Indonesia (KBBI), istilah warganet merujuk pada individu yang aktif menggunakan internet. Estu Suryowati menjelaskan bahwa istilah warganet atau netizen tidak ada sebelum munculnya media digital seperti internet, radio, dan televisi. Istilah warganet diambil dari gabungan kata "warga" dan "internet," yang ditulis dan diucapkan tanpa modifikasi, serta merupakan akronim dari "internet" dan "citizen." Saat ini, masyarakat modern sering disebut sebagai netizen atau warganet, yang menunjukkan keterlibatan mereka dengan media baru

berbasis internet. Netizen ini sering kali terlibat dalam memberikan komentar pada berbagai postingan berita di media sosial. Dalam konteks penelitian ini, netizen diartikan sebagai pengguna, pembaca, atau penonton media sosial yang menyampaikan pendapat dan perasaan mereka melalui komentar terhadap informasi yang dibagikan di platform tersebut(Saleh & Furbani, 2022).

X banyak berbicara tentang program makan siang gratis, yang diusulkan oleh salah satu pasangan calon untuk Pemilu Presiden Indonesia 2024. Program ini ditujukan untuk ibu hamil dan balita, serta anak-anak di TK, SD, SMP, dan sekolah menengah.Namun, kebijakan ini menimbulkan kontroversi dan memicu berbagai pendapat serta diskusi di media sosial, terutama di *X*. Kata kunci *Program Makan Siang Dan Susu Gratis* menjadi perbincangan hangat di kalangan masyarakat dan memunculkan beragam opini(Nursinggah et al., 2024).

Pada Februari 2024, rakyat Indonesia mengikuti pesta demokrasi lima tahunan di mana tiga pasangan calon berkompetisi untuk pemilihan presiden. Program Makan Siang dan Susu Gratis untuk lembaga pendidikan dan pesantren, serta bantuan nutrisi untuk bayi dan ibu hamil, dipromosikan oleh salah satu pasangan calon dengan nomor urut 02.Program ini dimaksudkan untuk menangani masalah *tengkes* atau *stunting*, yang merupakan masalah yang sangat mendasar di Indonesia. *Stunting* adalah kondisi yang disebabkan oleh kekurangan nutrisi jangka panjang selama masa pertumbuhan awal, yang dapat menyebabkan masalah fisik, psikologis, dan kognitif serta meningkatkan risiko penyakit metabolik.Data tahun 2022 menunjukkan bahwa angka *stunting* di Indonesia masih cukup tinggi, mencapai 21,6%.Sebagai solusi, program ini menyediakan makan siang gratis setiap hari bagi siswa prasekolah, SD, SMP, SMA, serta santri di pondok pesantren. Selain itu, ibu hamil dan anak balita

juga mendapat dukungan gizi untuk meningkatkan kesehatan mereka sekaligus membantu perekonomian keluarga. Kebijakan ini diharapkan dapat meningkatkan kesejahteraan masyarakat dan mengurangi angka *stunting*.Dengan anggaran sekitar 450 triliun rupiah per tahun, program diharapkan beroperasi secara merata pada tahun 2029.Targetnya adalah memberikan manfaat bagi sekitar 82,9 juta penerima di seluruh Indonesia(Saputra & Hasan, 2024).

Analisis sentimen adalah metode untuk menilai nada emosional dalam teks digital untuk menentukan apakah suatu pendapat bersifat positif, negatif, atau netral. Analisis sentimen adalah proses mempelajari pendapat, emosi, evaluasi, dan sikap masyarakat tentang berbagai hal, seperti barang, orang, organisasi, masalah, layanan, dan peristiwa.Selain itu, analisis sentimen sangat relevan untuk memahami persepsi publik terhadap suatu topik tertentu karena sumber informasi utama berasal dari media sosial, tempat pengguna secara aktif menyampaikan pandangan dan pendapat mereka. Oleh karena itu, analisis sentimen sangat terkait dengan masyarakat(Nursinggah et al., 2024).

Naïve Bayes Classifier adalah salah satu algoritma yang digunakan dalam *text mining* yang digunakan untuk memprediksi peluang di masa depan berdasarkan pengalaman sebelumnya. Algoritma ini bekerja dengan pendekatan probabilistik sederhana dan menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dalam dataset. Hal ini membuatnya cocok untuk analisis sentimen di media sosial, di mana data yang diperoleh mencakup berbagai topik, masalah, dan perbedaan penggunaan bahasa(Mauliza & Sipayung, 2024). Kecepatan dan kesederhanaan algoritma *Naïve Bayes* ideal dipilih untuk

menganalisis data dalam jumlah besar seperti yang dihasilkan oleh ulasan media sosial, dalam hal ini *Twitter*(Saputra & Hasan, 2024).

Banyak penelitian sebelumnya telah dilakukan terkait analisis sentimen dengan algoritma *Naive Bayes*. Salah satu penelitian membahas analisis sentimen terhadap program Kampus Merdeka; sistem berhasil mengklasifikasikan 272 opini sebagai sentimen positif dan 229 sebagai sentimen negatif, dengan tingkat akurasi 60%, *precision* 64%, *recall* 58%, dan skor F1-58%.Penelitian tambahan, yang menggunakan 2.007 data *Twitter*, menemukan 673 sentimen negatif, 668 sentimen positif, dan 665 sentimen netral. Hasil analisis sentimen yang dilakukan pada layanan pinjaman online di *Twitter* menggunakan 2.912 data menunjukkan akurasi sebesar 71%, ketepatan 73%, dan *recall* sebesar 89%. Selain itu, penelitian sentimen yang dilakukan menggunakan QRIS (*Quick Response Code Indonesia Standard*) menggunakan algoritma *Naive Bayes* menunjukkan bahwa dari 913 data, 65% adalah sentimen positif dan 35% adalah sentimen negatif. Hasil pengujian menunjukkan akurasi 99,89%, *precision* 99,83%, dan *recall* 99,68%. Studi lain yang menganalisis sentimen ulasan pengguna *e-commerce* dengan algoritma *Naive Bayes* menemukan akurasi 99,5%, *precision* 99,49%, dan *recall* 100%(Saputra & Hasan, 2024).

Data yang telah melewati tahap *preprocessing* harus diubah menjadi format numerik untuk mendukung analisis sentimen. Untuk melakukan konversi ini, metode pembobotan *TF-IDF* digunakan. Metode ini dikenal sebagai Term Frequency-Inverse Document Frequency (*TF-IDF*), yang menilai seberapa erat suatu kata (*term*) dengan dokumen dengan memberikan bobot pada setiap kata(Deolika & Taufiq Luthfi, 2021) memungkinkan *Naive Bayes* bekerja lebih optimal dalam

mengidentifikasi kata-kata kunci dalam ulasan. Kombinasi *TF-IDF* dan *Naive Bayes* telah banyak diakui sebagai pendekatan yang andal dalam penelitian analisis teks.

Studi ini menganalisis data dari rentang waktu 10 Februari 2024 sampai dengan 20 Maret 2025. Periode ini dipilih untuk mencakup dinamika opini publik menjelang dan setelah pemilihan presiden, serta selama pelaksanaan program makan siang gratis untuk memahami respons publik terhadap program tersebut di platform X. Data dikumpulkan melalui perintah *tweet-harvest* di lingkungan *Google Colab*, menggunakan token autentikasi dan parameter pencarian tertentu. Hasilnya disimpan dalam format .csv untuk keperluan analisis.Hasilnya diharapkan memberikan gambaran yang lebih jelas mengenai penerimaan masyarakat terhadap program makan siang gratis.Diharapkan analisis ini akan memberikan gambaran yang lebih jelas tentang bagaimana program tersebut diterima oleh masyarakat(Nursinggah et al., 2024).

1.2 Identifikasi Masalah

1. Belum diketahuinya persepsi publik secara kuantitatif terhadap Program Makan Siang & Susu Gratis yang diimplementasikan oleh pemerintah melalui media sosial atau platform berita daring.
2. Banyaknya opini yang beragam, baik positif dan negatif membuat sulit bagi pemerintah atau pemangku kebijakan untuk mengukur efektivitas dan penerimaan program ini.
3. Kurangnya analisis berbasis data terhadap opini publik, sehingga pengambilan keputusan masih cenderung subjektif dan tidak berbasis bukti yang kuat.

4. Perlu adanya metode analisis yang efektif, seperti *Naïve Bayes*, untuk mengklasifikasikan sentimen publik terhadap program ini secara otomatis dan akurat.

1.3 Ruang Lingkup Penelitian

Agar pembahasan lebih fokus berdasarkan perumusan masalah diatas dan sesuai dengan batasan kemampuan penulis sehingga dapat menyelesaikan permasalahan, maka berikut lingkup batasan masalah dalam penelitian:

1. Penelitian ini terbatas pada komentar masyarakat terkait Program Makan Siang & Susu Gratis Di Media X yang menjadi objek analisis sentimen.
2. Data yang digunakan untuk penelitian diambil berupa teks berbahasa Indonesia.
3. Jumlah data yang diambil sebanyak 964 ulasan yang berupa teks berbahasa Indonesia.
4. Dalam penelitian ini menggunakan algoritma *Naïve Bayes.*
5. Pembobotan kata menggunakan algoritma *TF-IDF.*
6. Pengolahan data analisis sentimen menggunakan Aplikasi pengolahan data *Rapidminer*
7. Klasifikasi sentiment dibagi menjadi dua kategori yaitu berupa positif, dan negatif.
8. Tidak dilakukan analisis lebih lanjut mengenai aspek spesifik dalam opini publik, seperti aspek ekonomi, kesehatan, atau pendidikan.

1.4 Tujuan dan Manfaat Penelitian

1.4.1. Tujuan Penelitian

1. Menganalisis sentimen publik terhadap Program Makan Siang & Susu Gratis berdasarkan data yang diperoleh dari Media X.
2. Menerapkan metode *Naïve Bayes* untuk mengklasifikasikan opini publik ke dalam kategori positif, negatif.
3. Merancang website untuk melakukan analisis sentimen secara otomatis menggunakan metode *Naïve Bayes* terhadap opini publik mengenai Program Makan Siang & Susu Gratis di Media *X*.

1.4.2. Manfaat Penelitian

1. Hasil analisis dapat digunakan untuk menilai efektivitas program dan memahami tanggapan masyarakat secara lebih objektif.
2. Memberikan transparansi mengenai bagaimana opini publik terhadap program tersebut dikaji secara ilmiah.
3. Menjadi referensi dalam penerapan metode *Naïve Bayes* pada analisis sentimen kebijakan publik.

1.5 Sistematika Penulisan

Penulisan tugas akhir ini disusun dalam beberapa bab dengan sistematika sebagai Berikut:

BAB I PENDAHULUAN

Bab ini berisi latar belakang masalah, rumusan masalah, tujuan penelitian, manfaat penelitian, serta ruang lingkup penelitian yang memberikan gambaran umum mengenai alasan dan fokus penelitian ini.

BAB II LANDASAN TEORI

Pada bab ini dibahas teori-teori yang relevan dengan penelitian, hasil penelitian terdahulu, serta landasan teoritis yang mendukung pembahasan penelitian. Bab ini juga menguraikan konsep-konsep yang digunakan sebagai dasar dalam analisis.

BAB III METODOLOGI PENELITIAN

Bab ini menjelaskan metode yang digunakan dalam penelitian, termasuk desain penelitian, objek dan subjek penelitian, teknik pengumpulan data, serta metode analisis data. Bab ini memberikan gambaran tentang bagaimana penelitian dilakukan.

BAB IV HASIL DAN PEMBAHASAN

Bab ini menyajikan hasil penelitian berdasarkan analisis data yang telah dilakukan. Selain itu, dalam bab ini juga terdapat pembahasan mengenai hasil eksperimen serta evaluasi kinerja metode yang digunakan.

BAB V SIMPULAN DAN SARAN

Bab ini berisi Simpulan dari penelitian yang telah dilakukan serta saran untuk pengembangan lebih lanjut.

BAB II

LANDASAN TEORI

2.1. Analisis Sentimen

Salah satu cabang *text mining* adalah analisis sentimen. Dataset yang digunakan untuk analisis ini dapat diperoleh dari berbagai sumber, seperti kolom ulasan produk di *e-commerce* yang menyampaikan perasaan dan pendapat mereka tentang produk yang mereka beli(Rahel Lina Simanjuntak et al., 2023).

Analisis sentimen memiliki manfaat yang luas dalam berbagai bidang, yang menjadi perhatian para pakar dan peneliti, terutama dalam konteks interaksi manusia-komputer. Selain itu, analisis ini juga memiliki relevansi yang signifikan dalam berbagai disiplin ilmu, termasuk sosiologi, pemasaran dan periklanan, psikologi, ekonomi, dan ilmu politik. Media sosial telah berkembang menjadi salah satu alat penting bagi masyarakat di era internet saat ini untuk berbagi pendapat dan perspektif mereka tentang berbagai topik. *E-commerce* adalah salah satu platform yang populer untuk tujuan ini.Melalui platform *e-commerce* ini, pengguna dapat dengan mudah menulis dan berbagi opini mereka mengenai produk atau layanan yang mereka temui. Fitur ulasan yang disediakan oleh *e-commerce* mempermudah proses ini, memungkinkan konsumen untuk memberikan *feedback* yang dapat memengaruhi persepsi orang lain mengenai produk yang dijual (Rahel Lina Simanjuntak et al., 2023).

2.2. Media Sosial

Media sosial adalah kumpulan aplikasi berbasis *Web* 2.0 yang memungkinkan pengguna membuat dan berbagi konten yang dibuat oleh pengguna. Selain itu,

media sosial juga dapat dianggap sebagai seperangkat alat untuk berkomunikasi dan bekerja sama yang memungkinkan berbagai jenis interaksi yang sebelumnya tidak dapat dilakukan oleh masyarakat umum (Pratidina & Mitha, 2023).

Secara sederhana, media sosial adalah tempat pengguna dapat berkolaborasi dan berinteraksi satu sama lain. Media sosial juga berfungsi sebagai perantara komunikasi bisnis berbasis internet yang memungkinkan penggunanya menampilkan diri, bekerja sama, berbagi informasi, dan berkomunikasi dengan orang lain, membentuk ikatan sosial virtual. Selain itu, media sosial mempermudah proses berbagi informasi antar pengguna. Media sosial sangat mudah diakses melalui perangkat seluler, memungkinkan pengguna mendapatkan informasi kapan saja dan di mana saja. Ini adalah keunggulan utama lainnya(Herdiyani et al., 2022).

2.3. Twitter

Generasi muda, terutama Generasi Y dan Generasi Z, merasakan peningkatan pemanfaatan teknologi informasi, terutama selama pandemi COVID-19. Lima platform media sosial yang paling banyak digunakan orang Indonesia adalah *YouTube, WhatsApp, Instagram, Facebook*, dan *Twitter*, menurut penelitian yang dilakukan *We Are Social dan Hootsuite* (2021). *Twitter* paling banyak digunakan oleh orang Indonesia karena fiturnya yang interaktif dan kemampuan untuk menyebarkan informasi dengan cepat (Khairunnisa & Alfaruqy, 2022).

Twitter, selain *WhatsApp, Instagram, dan Facebook*, telah menjadi platform media sosial yang sangat populer di mana banyak orang menggunakannya untuk berkomunikasi, berbagi ide, dan berbagi informasi.Menurut Gatra dalam Rosyida & Siroj (2021), salah satu platform media sosial yang paling populer di Indonesia

adalah *Twitter*. Indonesia adalah negara kelima dengan 19,5 juta pengguna *Twitter*, menurut data(Widyatnyana et al., 2023).

2.4. Program Makan Siang & Susu Gratis

Program makan siang gratis merupakan salah satu janji kampanye pasangan calon presiden nomor urut 2 pada Pemilu 2024. Tujuan dari program ini adalah untuk menghasilkan generasi yang cerdas, sehat, dan kompetitif dengan menyediakan makan siang dan susu secara gratis bagi siswa di sekolah dan pesantren, serta membantu ibu hamil dan anak balita mendapatkan gizi yang baik. Untuk mewujudkan Visi Indonesia Emas 2045, program ini juga diharapkan dapat meningkatkan produktivitas ekonomi(Sitanggang et al., 2024). . Diharapkan program makan siang dan susu gratis akan membantu mengatasi *stunting* dan meningkatkan kesehatan masyarakat Indonesia. Dengan anggaran tahunan sebesar 450 triliun rupiah, program diharapkan dimulai secara merata pada tahun 2029. Target program adalah untuk memberikan manfaat kepada 82,9 juta penerima, termasuk ibu hamil, anak balita, dan siswa sekolah (Saputra & Hasan, 2024).

2.5. Teks Preprocessing

Text preprocessing adalah tahapan dari proses awal terhadap teks yang bertujuan untuk mempersiapkan teks menjadi data yang akan diproses lebih lanjut. Tujuan dari tahapan ini adalah untuk mengubah data mentah yang tidak terstruktur menjadi format yang lebih teratur dan siap digunakan (Supriyanto et al., 2023). Adapun tahapan dalam *text preprocessing* diantaranya adalah sebagai berikut:

A. *Cleaning*

Cleaning adalah proses membersihkan teks dengan menghapus bagian yang tidak penting atau tidak penting dari teks yang diperoleh. Tujuan dari proses ini

adalah untuk meningkatkan kualitas data sehingga lebih siap untuk dianalisis.Dalam proses data *cleaning*, terdapat beberapa langkah yang dapat disesuaikan dengan kebutuhan pengolahan teks, diantaranya:Penghapusan *men tion*(@),Penghapusan karakter tunggal (*single character*), Penghapusan tanda baca(*punctuation removal*),Penghapusan *hashtag* (#) ,Penghapusan URL ,Penghapusan *whitespace*,Penghapusan angka (*number removal*) *Cleaning* membuat teks lebih bersih dan terstruktur, meningkatkan efektivitas analisis dan akurasi model yang digunakan dalam pemrosesan teks(Asriana et al., 2024).

B. *CaseFolding*

Salah satu tahap *pre-processing* data adalah *case folding*, yang bertujuan untuk menyeragamkan format teks dengan mengubah semua karakter menjadi huruf kecil. Ini dilakukan untuk menghilangkan perbedaan antara huruf besar dan kecil, sehingga teks menjadi lebih konsisten selama proses analisis. Sebagai contoh, kata "Analisis Sentimen" akan diubah menjadi "analisis sentimen". Ini menunjukkan bahwa hanya perubahan huruf kapital tidak membuat kata yang sama dianggap berbeda. Dalam proses pemrosesan bahasa natural (*NLP*), *case folding* sangat penting.karena banyak algoritma dan teknik pemrosesan teks bekerja lebih efektif ketika tidak membedakan antara huruf besar dan kecil. Dengan demikian, *case folding* dapat meningkatkan akurasi analisis dan model yang dibangun dari data tersebut(Indra Buana & Brahma Arianto, 2024).

C. *Tokenizing*

Sebagai contoh, token "saya suka belajar bahasa" akan dibagi menjadi empat kata: "saya", "suka", "belajar", dan "bahasa."Langkah ini sangat penting dalam

Natural Language Processing (*NLP*) karena membantu sistem komputer dalam memahami dan menganalisis teks dengan lebih efektif. Dengan menerapkan *tokenizing*, berbagai analisis lanjutan seperti klasifikasi teks, ekstraksi fitur, atau pembuatan model berbasis teks dapat dilakukan secara lebih akurat dan efisien (Indra Buana & Brahma Arianto, 2024).

D. *SpellingNormalization*

Normalisasi kata merupakan salah satu tahap dalam *pre-processing* teks yang bertujuan untuk mengubah kata-kata yang tidak baku menjadi bentuk standar. Proses ini dilakukan dengan menghilangkan karakter yang diulang secara berurutan agar kata lebih sesuai dengan kaidah bahasa yang benar, seperti mengubah kata *"okeee"* menjadi *"oke"* atau *"baikkk"* menjadi *"baik"*. Selain itu, normalisasi juga mencakup penghapusan karakter tunggal yang tidak memiliki makna signifikan, seperti *"y"* atau *"g"*, serta mengganti kata-kata *slang* dengan bentuk bakunya, misalnya *"g"* diubah menjadi *"nggak"* atau *"btw"* menjadi *"by the way"*. Dengan menerapkan normalisasi kata, teks menjadi lebih seragam dan lebih mudah diproses dalam analisis berbasis *Natural Language Processing* (*NLP*), sehingga dapat meningkatkan akurasi dalam memahami dan mengolah informasi yang terkandung di dalamnya(Asriana et al., 2024).

E. *Filtering*

Salah satu langkah penting dalam *pre-processing* data teks adalah *filtering* atau penghilangan kata-kata berhenti (*stopwords*). Kata-kata berhenti ini adalah kata-kata umum yang sering muncul dalam teks, seperti "dan", "atau", "di", dan sebagainya, tetapi tidak memberikan informasi yang signifikan untuk analisis.

Pada tahap *filtering*, kata-kata berhenti ini diidentifikasi dan dihapus dari teks agar analisis lebih terfokus pada kata-kata kunci yang lebih informatif. Penghapusan *stop words* dapat mengurangi kompleksitas data, mempercepat proses analisis, dan meningkatkan akurasi model dalam berbagai aplikasi *Natural Language Processing* (*NLP*) seperti analisis sentimen, klasifikasi teks, dan pencarian informasi(Indra Buana & Brahma Arianto, 2024).

F. *Stemming*

Salah satu tahap *pre-processing* data teks adalah *stemming*, atau normalisasi. Tujuannya adalah untuk mengubah kata menjadi bentuk dasarnya atau akar kata. Dalam proses ini, imbuhan seperti awalan (*prefix*), sisipan (*infix*), dan akhiran (*suffix*) dihilangkan, sehingga hanya kata dasar yang memiliki makna inti yang tersedia.

Sebagai contoh, kata "berlari", "lari", dan "lari-lari" akan dikonversi menjadi bentuk dasar "lari". Dengan demikian, berbagai variasi kata yang memiliki akar yang sama dapat dikonsolidasikan, sehingga mempermudah analisis teks. *Stemming* berperan penting dalam berbagai aplikasi *Natural Language Processing* (*NLP*) seperti klasifikasi teks, analisis sentimen, dan pengelompokan dokumen, karena membantu dalam mengurangi kompleksitas data dan meningkatkan akurasi model analisis teks(Indra Buana & Brahma Arianto, 2024).

2.6. TF-IDF

Term Frequency Inverse Document Frequency (*TF-IDF*) adalah algoritma yang digunakan untuk membatasi hubungan antara kata atau istilah dengan ulasan umum yang ada.(Sera et al., 2023). Metode Frekuensi Kata-I dalam *TF-IDF* terdiri

dari tiga langkah. Pertama, *Frequency of Words (TF)* dihitung untuk mengetahui seberapa sering kata muncul dalam sebuah dokumen. Kemudian, *Frequency of Documents (DF)* dihitung untuk mengetahui berapa banyak dokumen yang mengandung kata tersebut. Terakhir, *Frequency of Inverse Documents (IDF)* dihitung untuk mengurangi bobot kata jika kata tersebut muncul dengan sering dalam banyak dokumen (Gerald Halim & Noer, 2023). Setiap kata dalam data teks dapat diubah menjadi nilai numerik oleh pembobotan *TF-IDF* (Junaedi et al., 2024), sehingga dapat digunakan dalam berbagai algoritma pemrosesan teks seperti *Naïve Bayes.*Persamaan *TF-IDF* dirumuskan sebagai berikut:

$$\mathrm{TF} - \mathrm{IDF}(t,d) = \mathrm{TF}(t,d) \times \mathrm{IDF}(t)$$

Di mana:

TF(t,d) = menunjukkan frekuensi kemunculan term *t* dalam dokumen *d*.

IDF(t) = adalah *Inverse Document Frequency*, yang dihitung menggunakan rumus:

$$Log\,\frac{n}{nt}$$

dengan *N* sebagai jumlah total dokumen dalam koleksi dan *nt* sebagai jumlah dokumen yang mengandung *term t*.

2.7. Naïve Bayes

Selanjutnya, algoritma yang digunakan dalam analisis sentimen adalah *Naïve Bayes Classifier. Naïve Bayes* adalah metode pembelajaran mesin yang mengandalkan perhitungan probabilitas dan statistik dan pertama kali diperkenalkan oleh ilmuwan asal Inggris Thomas Bayes(Rahel Lina Simanjuntak et al., 2023).

Metode ini dapat digunakan untuk klasifikasi yang menggunakan pendekatan statistik dan probabilitas. Dengan menggunakan data atau pengalaman sebelumnya, *Naive Bayes* memprediksi peluang di masa depan (Mauliza & Sipayung, 2024).

Didasarkan pada Teorema Bayes, metode klasifikasi *Naïve Bayes* digunakan untuk memprediksi kemungkinan suatu kelas berdasarkan data historis dengan asumsi bahwa setiap fitur dalam data bersifat independen satu sama lain (asumsi naif). Algoritma *Naïve Bayes* menghitung kemungkinan suatu kelas dengan menggabungkan frekuensi kemunculan fitur dalam dataset yang diberikan(Christian & Fenriana, 2023). Pendekatan ini memungkinkan model untuk melakukan prediksi dengan cepat dan efisien pada dataset yang sangat besar. Karena kemudahan implementasinya dan kemampuan untuk menangani data berukuran besar, *Naive Bayes* sering digunakan dalam berbagai aplikasi seperti analisis sentimen, klasifikasi teks, deteksi spam, dan sistem rekomendasi.Secara sistematis Teorema Bayes dirumuskan sebagai:

$$P(A|B) = \big(P(B|A) * P(A)\big)/P(B)$$

Di mana:

B = data dengan kelas yang belum diketahui.

A = hipotesis bahwa data termasuk dalam suatu kelas tertentu.

P(A|B) = *posterior probability*, yaitu probabilitas hipotesis A berdasarkan kondisi data B.

P(A) = *prior probability*, yaitu probabilitas awal dari hipotesis A sebelum mempertimbangkan data B.

P(B|A) = probabilitas bahwa data B terjadi jika hipotesis A benar.

P(B) =probabilitas bahwa data B terjadi secara keseluruhan.

2.8. Klasifikasi

Salah satu jenis model dalam data mining, klasifikasi digunakan untuk memberi label kategori pada data, seperti "aman" atau "berisiko" untuk data keuangan, "ya" atau "tidak" untuk data penjualan, atau "pengobatan A", "pengobatan B", atau "pengobatan C" untuk data medis. Proses penerapan klasifikasi dimulai dengan membuat model data yang lebih rinci, disebut data latih. Setelah model tersebut dibuat melalui proses pelatihan, model tersebut dapat digunakan untuk mengklasifikasikan data baru dalam proses yang disebut proses uji(Rahel Lina Simanjuntak et al., 2023)

2.9. Data

Data merupakan deskripsi mengenai suatu objek, kejadian, aktivitas, atau transaksi yang secara terpisah belum memiliki makna yang jelas bagi pengguna. Sebagai contoh, deretan angka seperti 6.30, 27, 6.32, 28, 6.34, 27 tidak memiliki arti yang spesifik tanpa adanya konteks yang menjelaskan makna angka-angka tersebut. Data dapat berbentuk nilai yang terformat, teks, citra, audio, maupun video(Taufik et al., 2022).Data di bagi menjadi beberapa jenis yaitu:

A. Data Format

   Data yang memiliki format tertentu, seperti data yang merepresentasikan tanggal, waktu, atau nilai mata uang. Contohnya, "10 Maret 2024" sebagai data tanggal atau "Rp 50.000" sebagai data yang menunjukkan nilai mata uang.

B. Teks

Data berupa rangkaian huruf, angka, dan simbol-simbol tertentu yang memiliki struktur, namun tidak bergantung pada masing-masing karakter secara individual. Contoh dari data teks antara lain artikel berita, laporan ilmiah, atau percakapan dalam platform komunikasi digital.

C. Citra

Data dalam bentuk gambar atau visual, seperti grafik, foto, hasil pencitraan medis (rontgen), atau ilustrasi lainnya.

D. Audio

Data yang berbentuk suara, seperti rekaman percakapan, musik, suara hewan, atau bunyi alam seperti hujan dan angin.

E. Video

Data yang terdiri dari kumpulan gambar bergerak yang dapat disertai dengan suara. Contohnya adalah rekaman seminar, film dokumenter, atau video edukasi.

2.10. Informasi

Menurut Burch dan Grundnitski (Taufik et al., 2022), Proses *transformasi* data menjadi informasi yang kemudian digunakan untuk pengambilan keputusan dan menghasilkan data baru, yang membentuk siklus berkelanjutan.

Perbedaan utama antara data dan informasi terletak pada makna yang dikandungnya. Data merupakan sekumpulan fakta mentah yang belum memiliki arti, sedangkan informasi memiliki makna yang dapat dipahami oleh penerima. Makna dalam informasi menjadi aspek krusial, karena memungkinkan penerima untuk

menginterpretasikan, menarik kesimpulan, dan menggunakannya sebagai dasar dalam pengambilan keputusan(Taufik et al., 2022).

2.11. Crawling

Crawling adalah proses mengumpulkan data dari halaman web dan kemudian menyimpannya secara offline. Penelitian ini menggunakan kata kunci #Pelaku untuk melakukan crawling di platform media sosial *Twitter*. *API Twitter* digunakan untuk mengumpulkan data. *Twitter, username*, waktu pemostingan, dan informasi lainnya yang relevan dengan penelitian dimasukkan ke dalam data yang dikumpulkan.

2.12. RapidMiner

RapidMiner adalah perangkat lunak *open-source* yang digunakan untuk analisis data seperti *data mining*, *text mining*, dan analisis prediktif. Perangkat lunak ini memiliki banyak alat yang membantu pengguna menerapkan algoritma pembelajaran mesin. seperti *Naïve Bayes*, untuk berbagai keperluan analisis. Dalam konteks analisis sentimen, *RapidMiner* memungkinkan peneliti untuk mengakses dan menerapkan metode *Naïve Bayes* secara efisien pada data dari media sosial, seperti *Twitter*. Dengan demikian, *RapidMiner* dapat digunakan untuk mengolah dan menganalisis sentimen masyarakat terhadap suatu topik tertentu, misalnya program makan siang dan susu gratis(Saputra & Hasan, 2024).

2.13. Aplikasi Berbasis *Website*

Program komputer yang memanfaatkan peramban web dan teknologi internet untuk melakukan berbagai tugas dikenal sebagai aplikasi berbasis web.

Aplikasi ini menggunakan dua skrip sisi *server*, seperti *ASP* atau *PHP*, untuk menangani penyimpanan dan pengambilan data, dan skrip sisi klien, seperti *JavaScript* dan *HTML*, untuk menampilkan dan mengawasi interaksi pengguna. Metode ini memungkinkan pengguna berinteraksi secara langsung dengan berbagai fitur situs web, seperti formulir online, kolom komentar, dan sistem manajemen konten (*CMS*). Metode ini memungkinkan komunikasi yang lebih dinamis antara pengguna dan pemilik situs *web* (Suryawinata, 2019).

2.14. *Natural Language Processing* (*NLP)*

Pemrosesan bahasa alami *(NLP*) adalah bidang ilmu komputer di mana pemrosesan bahasa alami digunakan untuk mengatasi masalah yang dihadapi sistem komputer dalam pengenalan bahasa percakapan sehari-hari. Analisis sentimen, yang bertujuan untuk memahami opini atau perasaan dari teks yang tidak terstruktur, seperti unggahan di media sosial, seperti *Twitter,* adalah salah satu aplikasi *natural linguistics*.

Dalam *NLP*, terdapat dua teknik utama yang digunakan untuk memahami bahasa, yaitu analisis sintaksis dan analisis semantik. Analisis *sintaksis* berfokus pada struktur tata bahasa untuk memastikan bahwa suatu kalimat sesuai dengan aturan *linguistik*, sedangkan analisis *semantik* bertujuan untuk menafsirkan makna dari suatu kalimat agar dapat dipahami secara lebih mendalam oleh sistem komputer(Asriana et al., 2024).

2.15. Machine Learning

Kecerdasan buatan (AI) telah menjadi salah satu pendorong utama kemajuan teknologi dalam era digital yang semakin terhubung dan penuh data. Pemahaman bahasa, pengambilan keputusan, dan pemecahan masalah adalah contoh tugas-tugas

yang biasanya membutuhkan kecerdasan manusia, dan kecerdasan buatan adalah fokus *AI*. Metode *AI* bervariasi. mulai dari logika simbolis hingga pembelajaran mesin dan pembelajaran mendalam. Tujuan utama dari kecerdasan buatan adalah menciptakan sistem yang dapat memahami informasi, belajar dari pengalaman, merancang strategi, serta beradaptasi dengan lingkungan secara mandiri, sehingga dapat meningkatkan efisiensi dan efektivitas dalam berbagai bidang, termasuk bisnis, kesehatan, dan industri(Pipin et al., 2024).

Menurut buku *An Introduction to Statistical Learning* karya Gareth James, Daniela Witten, Trevor Hastie, dan Robert Tibshirani (2013), *Machine Learning* didefinisikan sebagai proses di mana komputer belajar dari data tanpa diberikan instruksi yang jelas. Hal ini menunjukkan bahwa sistem *Machine Learning* mampu mengenali pola dan membuat generalisasi sendiri dalam menyelesaikan suatu permasalahan tanpa harus diberikan aturan atau instruksi eksplisit oleh program. Dengan kata lain, sistem ini secara otomatis menyesuaikan diri berdasarkan data yang diproses, sehingga dapat digunakan untuk berbagai aplikasi, seperti prediksi, klasifikasi, dan pengambilan keputusan berbasis data. Perbedaan mendasar antara *Machine Learning* dan pemrograman konvensional terletak pada cara sistem menangani data dan mengambil keputusan. Dalam pemrograman konvensional, pemrogram harus menentukan secara eksplisit aturan dan instruksi yang mengatur bagaimana sistem merespons setiap kemungkinan masukan. Dengan kata lain, setiap skenario harus diprogram secara manual agar sistem dapat berfungsi dengan baik. Sementara itu, dalam *Machine Learning*, sistem tidak memerlukan aturan yang ditentukan secara eksplisit oleh pemrogram. Sebaliknya, sistem belajar dari data yang diberikan, mengenali pola, dan membuat generalisasi sendiri untuk menghasilkan prediksi atau keputusan. Pendekatan ini memungkinkan *Machine*

Learning untuk menangani berbagai masalah yang kompleks dan dinamis tanpa harus menuliskan aturan untuk setiap kemungkinan yang ada(Sumanto & Heryana, n.d.).

2.16. Teks Mining

Text mining adalah proses berbasis pengetahuan yang melibatkan interaksi pengguna dengan sekumpulan dokumen dalam jangka waktu tertentu melalui berbagai metode analisis. *Text mining* berbeda dengan *data mining*, yang biasanya berfokus pada data terstruktur dalam basis data, dan berfokus pada data tekstual yang tidak terstruktur, seperti kumpulan artikel atau dokumen. Tujuan utama dari *text mining* adalah untuk mengekstraksi informasi berharga dari sumber data dengan mengidentifikasi serta mengeksplorasi pola-pola penting. Oleh karena itu, meskipun memiliki tujuan yang serupa, *text mining* dan *data mining* memiliki perbedaan dalam sumber data yang digunakan.Meskipun demikian, kedua metode ini memiliki kemiripan dalam arsitektur tingkat tinggi, karena sama-sama bergantung pada beberapa tahapan utama, seperti preprocessing data, penerapan algoritma untuk menemukan pola, serta penggunaan visualisasi data guna meningkatkan pemahaman terhadap hasil analisis(Findawati et al., 2020).

Text mining adalah cabang khusus dari *data mining* yang berfokus pada eksplorasi data dalam bentuk teks. Sumber data utama *text mining* berasal dari dokumen, dengan tujuan utama untuk menemukan kata-kata kunci yang dapat mewakili isi dokumen sehingga memungkinkan analisis keterkaitan antar dokumen.Dalam praktiknya, *text mining* dapat menganalisis, mengelompokkan, serta menentukan kesamaan antar dokumen guna memahami hubungan antara dokumen dengan variabel lain.

Saat ini, *text mining* banyak diterapkan dalam berbagai bidang, seperti penyaringan spam, analisis sentimen, pengukuran preferensi pelanggan, peringkasan dokumen, serta pengelompokan topik penelitian.Menurut Abbot (2013), *text mining* memiliki beberapa kategori utama, yaitu:

1. *Search and Information Retrieval:* Menyimpan serta mengambil kembali dokumen teks, termasuk dalam sistem mesin pencari dan pencocokan kata kunci.
2. *Document Clustering*: Mengelompokkan dan mengategorikan istilah, paragraf, atau dokumen dengan metode clustering.
3. *Document Classification*: Pengelompokan dokumen berdasarkan kategori tertentu menggunakan teknik klasifikasi dokumen.
4. *Web Mining*: Penerapan data dan text mining pada internet yang berfokus pada hubungan antar situs web dalam skala besar.
5. *Information Extraction*: Mengidentifikasi serta mengekstraksi informasi yang relevan dari teks.
6. *Natural Language Processing* (NLP): Pemrosesan bahasa alami untuk menginterpretasikan teks dalam bentuk yang lebih terstruktur.
7. *Concept Extraction*: Pengelompokan kata atau frasa dalam suatu kelompok yang memiliki makna serupa.

2.17. Confusion Matrix

Confusion Matrix merupakan metode evaluasi yang sederhana namun banyak digunakan dalam permasalahan klasifikasi, khususnya klasifikasi biner. Dalam jurnal

yang ditulis oleh (Riehl et al., 2023), dijelaskan bahwa *confusion matrix* terdiri dari empat elemen utama yaitu *True Positive (TP), True Negative (TN), False Positive (FP),* dan *False Negative (FN).* Keempat elemen ini mewakili hasil prediksi model dibandingkan dengan label sebenarnya dan menjadi dasar untuk menghitung berbagai metrik evaluasi kinerja model klasifikasi(Riehl et al., 2023).

2.18. Penelitian Yang Relevan

2.18.1. *Sentiment Analysis Methods, Applications, and Challenges: A Systematic Literature Review* Oleh Chen et al. (2024)

Chen et al. (2024), dalam jurnal berjudul *"Sentiment Analysis Methods, Applications, and Challenges: A Systematic Literature Review"* yang diterbitkan pada *Journal of King Saud University - Computer and Information Sciences*, menyajikan kajian sistematis terhadap berbagai pendekatan, aplikasi, dan tantangan dalam bidang analisis sentimen. Penelitian ini menelusuri perkembangan metode analisis sentimen terkini, mulai dari pendekatan berbasis *leksikon* dan *machine learning* konvensional hingga model *deep learning* modern seperti *Transformer*, termasuk *BERT* dan *GPT*.

Studi ini mengulas lebih dari 300 publikasi yang relevan, dengan fokus utama pada lima aspek penting, yaitu teknik ekstraksi fitur, metode klasifikasi, evaluasi performa model, permasalahan data seperti ketidakseimbangan label, serta penerapan praktis di berbagai sektor seperti keuangan, layanan pelanggan, dan pemerintahan digital (*e-government*). Salah satu kontribusi penting dari penelitian ini adalah identifikasi terhadap celah penelitian yang masih belum banyak dijelajahi, misalnya keterbatasan model dalam menangani data

multibahasa serta rendahnya kemampuan generalisasi model terhadap data di dunia nyata.

Penelitian ini juga menyoroti tantangan dalam aspek replikasi studi, di mana mayoritas penelitian sebelumnya tidak menyertakan kode sumber atau dataset terbuka, sehingga menyulitkan proses verifikasi dan pengembangan lebih lanjut. Selain itu, studi ini mendorong pengembangan *riset* yang menggabungkan *deep learning* dengan pendekatan *explainable AI* guna meningkatkan transparansi serta kepercayaan terhadap sistem klasifikasi otomatis dalam analisis sentimen.

2.18.2. Komparasi Algoritma *Naïve Bayes* dan *Logistic Regression* untuk Analisis Sentimen *Metaverse* Oleh Ramadhani dan Suryono (2024)

Ramadhani dan Suryono (2024), dalam jurnal berjudul *"Komparasi Algoritma Naïve Bayes dan Logistic Regression untuk Analisis Sentimen Metaverse"*, membahas analisis opini masyarakat Indonesia terhadap perkembangan teknologi *metaverse* melalui data komentar di media sosial X. Penelitian ini bertujuan untuk membandingkan performa algoritma *Naïve Bayes* dan *Logistic Regression* dalam klasifikasi sentimen, serta mengatasi permasalahan ketidakseimbangan data menggunakan metode *SMOTE*.

Penelitian ini menggunakan 6.728 data *tweet* berbahasa Indonesia yang dikumpulkan dengan teknik *crawling* menggunakan *library Harvest* pada Python. Tahapan *text preprocessing* meliputi *cleansing*, *case folding*, *tokenizing*, *stopword removal*, *stemming*, dan *labeling* dengan bantuan *VADER*

lexicon. Pembobotan kata dilakukan dengan *TF-IDF,* kemudian data dilatih dengan dua algoritma klasifikasi: *Naïve Bayes* dan *Logistic Regression.*

Sebelum optimasi, *Logistic Regression* menunjukkan akurasi sebesar 91% dan *Naïve Bayes* sebesar 90%. Namun, nilai *precision, recall, dan F1-score* masih rendah, khususnya pada sentimen negatif. Setelah dilakukan optimasi dengan *SMOTE*, performa kedua algoritma meningkat signifikan. *Logistic Regression* mencapai akurasi 95%, *precision* 94%, *recall* 93%, dan *F1-score* 95%, sedangkan *Naïve Bayes* mencapai *akurasi* 91%, *precision* 87%, *recall* 97%, dan *F1-score* 92%.

Penelitian ini menyimpulkan bahwa *Logistic Regression* dengan optimasi *SMOTE* menjadi model terbaik untuk klasifikasi sentimen topik *metaverse.* Visualisasi dengan *wordcloud* juga digunakan untuk melihat frekuensi kata, yang menunjukkan bahwa topik seputar teknologi, dunia virtual, dan keamanan data mendominasi diskursus masyarakat. Penelitian ini berkontribusi dalam pengembangan analisis sentimen berbasis pembelajaran mesin untuk opini masyarakat terkait teknologi digital mutakhir(Ramadhani & Suryono, 2024).

2.18.3. *Hybrid Naïve Bayes TF-IDF Algorithm and Lexicon approach for sentimen analysis of reviews* Oleh Ramadhani (2024)

Ramadhani et al. (2024) dalam penelitiannya yang berjudul *"Hybrid Naive Bayes TF-IDF Algorithm and Lexicon Approach for Sentiment Analysis of Reviews"* mengusulkan pendekatan hibrida dalam analisis sentimen dengan mengintegrasikan algoritma *Naïve Bayes*, metode pembobotan *TF-IDF*, serta

pendekatan berbasis *leksikon* Bahasa Indonesia, yaitu *InSet* dan *SentiStrengthID.* Tujuan utama dari penelitian ini adalah meningkatkan akurasi klasifikasi sentimen terhadap ulasan aplikasi, yang dalam hal ini menggunakan dataset berjumlah 5.000 ulasan yang diperoleh dari platform *Kaggle*.

Proses analisis diawali dengan tahapan *pre-processing* data yang mencakup pembersihan teks, *case folding*, *tokenizing*, penghapusan *stopword*, *stemming*, serta *normalization*. Selanjutnya, pelabelan sentimen dilakukan secara otomatis berdasarkan skor *leksikal* dari dua kamus sentimen yang digunakan. Data dengan polaritas netral dieliminasi dari dataset guna meminimalkan *ambiguitas* dan meningkatkan kualitas klasifikasi.

Fitur teks kemudian ditransformasikan menggunakan metode *TF-IDF*, dan proses pelatihan dilakukan dengan model *Multinomial Naïve Bayes*. Evaluasi model menggunakan *confusion matrix* dengan metrik akurasi, *precision*, *recall*, dan *F1-score* sebagai indikator performa. Hasil eksperimen menunjukkan bahwa kombinasi *Naïve Bayes* dengan *leksikon SentiStrengthID* memberikan performa terbaik, dengan akurasi mencapai 84,6%, sementara kombinasi dengan *leksikon InSet* hanya memperoleh akurasi 68,6%.

Temuan ini menegaskan bahwa pemilihan jenis *leksikon* berperan penting dalam keberhasilan klasifikasi sentimen. Pendekatan *hibrida* yang menggabungkan teknik *leksikal* dan algoritma pembelajaran mesin dapat meningkatkan efektivitas dalam mengolah data ulasan berbahasa Indonesia, khususnya pada konteks aplikasi publik(Ahmad HaritsRamadhani, 2024).

2.18.4. *E-Commerce Brand Ranking Algorithm Based on User Evaluation and Sentiment Analysis* Oleh Chen & Zhang (2023)

Chen (2022), dalam artikelnya yang berjudul *"E-Commerce Brand Ranking Algorithm Based on User Evaluation and Sentiment Analysis"* yang diterbitkan dalam jurnal *Frontiers in Psychology*, mengusulkan sebuah algoritma pemeringkatan merek *e-commerce* yang didasarkan pada evaluasi pengguna dan hasil analisis sentimen. Tujuan utama dari penelitian ini adalah untuk mengidentifikasi figur atau pengguna yang berperan sebagai pemimpin opini dalam komunitas *e-commerce*, dengan mempertimbangkan interaksi pengguna serta polaritas sentimen dari komentar-komentar yang mereka hasilkan.

Data diperoleh melalui teknik *web scraping* dengan memanfaatkan pustaka Python seperti *BeautifulSoup* dan *urllib*, dan difokuskan pada tiga produk populer, yaitu Huawei Mate10, Lenovo Thinkpad, dan Sony A6300. Proses analisis data melibatkan segmentasi kalimat, tokenisasi, *part-of-speech tagging*, analisis *sintaksis*, serta analisis *semantik* menggunakan platform LTP yang dikembangkan oleh *Harbin Institute of Technology*.

Analisis sentimen dilakukan menggunakan pendekatan *deep learning* berbasis *Convolutional Neural Network* (CNN) yang dilatih dengan dataset berlabel emosi. Komentar pendek dianalisis dalam skema klasifikasi biner, yaitu sentimen positif dan negatif. Hasil dari klasifikasi sentimen tersebut kemudian digabungkan dengan algoritma *SRank*, yaitu pengembangan dari *LeaderRank* dan *PageRank*, untuk menghitung tingkat pengaruh dan posisi strategis pengguna dalam jaringan sosial *e-commerce.*

Hasil penelitian menunjukkan bahwa nilai pengaruh pengguna tidak hanya ditentukan oleh tingkat aktivitas atau jumlah pengikut, tetapi juga dipengaruhi oleh kualitas sentimen yang terkandung dalam komentar. Sebagai contoh, pengguna seperti "JD Jingdong" memiliki skor sentimen tinggi (0,90), sedangkan "Dangdang", meskipun aktif, memiliki skor lebih rendah (0,43). Hal ini menunjukkan bahwa penggabungan antara struktur jejaring sosial dan analisis sentimen dapat memberikan hasil yang lebih akurat dalam mengidentifikasi tokoh berpengaruh dalam ekosistem digital.

Penelitian ini menyimpulkan bahwa pendekatan berbasis gabungan antara aktivitas pengguna dan analisis sentimen memiliki potensi signifikan dalam pemeringkatan merek, identifikasi pemimpin opini, serta pengambilan keputusan strategis dalam pemasaran *e-commerce*(Chen, 2022).

2.18.5. *Sentiment analysis for e-commerce product reviews by deep learning model of Bert-BiGRU-Softmax* Oleh Li & Wang (2021)

Liu et al. (2021), dalam artikelnya yang berjudul *"Sentiment Analysis for E-Commerce Product Reviews by Deep Learning Model of Bert-BiGRU-Softmax"*, mengusulkan model *deep learning* gabungan untuk melakukan analisis sentimen terhadap ulasan produk *e-commerce* dalam skala besar. Model yang diusulkan terdiri atas tiga lapisan utama, yaitu *BERT* sebagai lapisan input untuk ekstraksi fitur kontekstual, *BiGRU* sebagai lapisan tersembunyi untuk menangani ketergantungan urutan data secara dua arah, dan *Softmax* dengan attention *mechanism* sebagai lapisan output untuk klasifikasi sentimen.

Penelitian ini dilakukan menggunakan dua dataset utama: *IMDB* dan *ChnSentiCorp*, serta satu dataset besar yang dikumpulkan dari platform *e-commerce* seperti *Sunning* dan *Taobao*, berisi lebih dari 500.000 ulasan produk. Model diuji terhadap aspek seperti kualitas, harga, layanan, dan logistik. Eksperimen menunjukkan bahwa model *Bert-BiGRU-Softmax* mencapai akurasi tertinggi sebesar 95,5%, mengungguli model lain seperti *RNN, LSTM, GRU, BiGRU,* dan bahkan *Bert-BiLSTM.*

Pendekatan ini memanfaatkan pelatihan dua tahap pada *BERT (Masked Language Model dan Next Sentence Prediction)* untuk menangkap nuansa sentimen dalam konteks kalimat penuh, sedangkan *BiGRU* digunakan untuk menyempurnakan penangkapan pola dari kedua arah konteks. *Softmax* dengan attention kemudian menghitung bobot sentimen yang paling relevan dari setiap dimensi ulasan (misalnya kamera, baterai, harga, dan sistem). Hasil akhir menunjukkan bahwa HW sebagai salah satu merek ponsel memiliki tingkat kepuasan tertinggi secara keseluruhan berdasarkan aspek-aspek yang dianalisis.

Kontribusi utama penelitian ini adalah peningkatan performa model dalam memahami sentimen multidimensi secara lebih mendalam melalui integrasi model *deep learning* canggih, serta penerapan dalam skenario dunia nyata dengan data ulasan dalam jumlah besar dari lingkungan e-commerce.(Liu et al., 2020).

2.18.6. Analisis Sentimen Ulasan Pengguna Aplikasi *ZenPro* dengan Implementasi Algoritma *Support Vector Machine (SVM)* oleh Buana & Arianto (2024)

Tujuan dari "*Analisis Sentimen Ulasan Pengguna Aplikasi ZenPro dengan Implementasi Algoritma Support Vector Machine (SVM)*" yang dilakukan oleh Buana dan Arianto (2024) adalah untuk membagi sentimen pengguna aplikasi *ZenPro* ke dalam dua kategori: sentimen positif dan negatif. Studi ini mengumpulkan 1000 ulasan pengguna dari *Google Play Store* dengan menggunakan teknik *web scraping* dengan bantuan *library Google Play Scraper*.

Case folding, cleaning, tokenizing, stopword removal, dan *stemming* adalah semua tahapan *pre-processing* yang digunakan. Data dilabelkan menggunakan pendekatan berbasis *leksikon*, dan kemudian dibagi menjadi data latih dan data uji dalam proporsi 80:20. Selanjutnya, metode *TF-IDF* digunakan untuk membobot fitur dan algoritma *SVM* dengan *kernel linier* digunakan untuk proses klasifikasi. Menurut evaluasi, model menunjukkan kinerja yang luar biasa, dengan akurasi sebesar 90%, ketepatan sebesar 93,42%, *recall* sebesar 94,04%, dan skor F1-sebesar 93,73%.

Hasil penelitian ini menunjukkan bahwa teknik pembelajaran mesin efektif untuk menganalisis perasaan dalam data teks berbahasa Indonesia. Hasil klasifikasi juga dapat menunjukkan persepsi pengguna terhadap fitur, kualitas, dan performa aplikasi. Penelitian ini menjadi referensi penting dalam pengembangan analisis sentimen, terutama dalam aspek pemrosesan teks,

pelabelan data, serta penerapan algoritma pembelajaran mesin(Indra Buana & Brahma Arianto, 2024).

2.18.7. analisis sentimen ulasan pelanggan terhadap produk roti macan artisan di *e-commerce* tokopedia menggunakan metode *clustering* oleh Azizah & Ernawati (2024)

Dalam penelitian mereka yang berjudul "*Analisis Sentimen Ulasan Pelanggan terhadap Produk Roti Macan Artisan di E-Commerce Tokopedia Menggunakan Metode Clustering*", Azizah dan Ernawati (2024) berusaha untuk mengetahui bagaimana konsumen melihat produk roti artisan melalui analisis ulasan yang dikumpulkan dari platform Tokopedia. Studi ini menggunakan algoritma *clustering K-Means* untuk mengelompokkan 26.500 data ulasan pelanggan dari Agustus hingga Desember 2023. Data ini dikelompokkan berdasarkan beberapa variabel, termasuk rating, komentar, harga produk, lokasi, dan kuantitas pembelian.

Untuk memastikan bahwa data bersih dan siap untuk analisis, proses *preprocessing* teks dimulai, yang mencakup penghapusan duplikat, normalisasi, penghapusan *stopword*, tokenisasi, dan *stemming*. Untuk mendapatkan jumlah klaster yang ideal, metode *Elbow* digunakan. Bahasa pemrograman Python digunakan untuk menjalankan proses *clustering* dengan bantuan pustaka seperti *Scikit-learn dan Pandas*.

Hasil analisis menunjukkan bahwa data ulasan terbagi ke dalam dua kelompok besar, di mana sebagian besar menunjukkan sentimen positif terhadap produk, khususnya terkait rasa dan kualitas roti dengan varian

unggulan seperti *Cranberry Cheese* dan *Double Chocolate*. Di sisi lain, ulasan dengan sentimen negatif umumnya mengkritisi aspek sistem pemesanan dan pengemasan produk yang dianggap masih perlu perbaikan.

Temuan dalam penelitian ini menegaskan bahwa kepuasan pelanggan memiliki peran penting dalam membentuk reputasi toko. Ulasan positif mampu meningkatkan kepercayaan pelanggan dan berdampak langsung pada peningkatan penjualan, sementara ulasan negatif memberikan masukan konstruktif bagi pengembangan layanan. Pendekatan *clustering* terbukti efektif dalam mengidentifikasi pola dan preferensi pelanggan, sehingga dapat dimanfaatkan oleh manajemen toko untuk merumuskan strategi bisnis yang lebih adaptif dan berbasis data(Azizah & Ernawati, 2024).

2.18.8. Penerapan Algoritma *K-Nearest Neighbor (K-NN)* Untuk Analisis Sentimen Publik Terhadap Pembelajaran *Daring* Oleh Supriyanto, Alita, dan Isnain (2023)

Penelitian berjudul "*Penerapan Algoritma K-Nearest Neighbor (K-NN) Untuk Analisis Sentimen Publik Terhadap Pembelajaran Daring*" dilakukan oleh Supriyanto, Alita, dan Isnain (2023) dan menggunakan 1.825 *tweet* yang dikumpulkan dari Februari hingga September 2020. Tujuan dari penelitian ini adalah untuk mengklasifikasikan pendapat masyarakat tentang kebijakan pembelajaran *online* selama pandemi COVID-19.

Sebelum pengolahan, proses pengolahan data terdiri dari *case folding, cleansing, removal of stopwords, stemming, dan tokenization*. Setelah teks diproses, metode *TF-IDF* digunakan untuk membedakan fiturnya. Selanjutnya,

algoritma *K-Nearest Neighbor (K-NN)* digunakan untuk klasifikasi dengan parameter nilai K=10. Dengan menggunakan *confusion matrix*, evaluasi performa model dilakukan. Hasilnya menunjukkan akurasi sebesar 84,93%, ketepatan sebesar 87%, *recall* sebesar 86%, skor F1 sebesar 87%, dan tingkat kesalahan sebesar 0,12%.

Hasil klasifikasi menunjukkan bahwa mayoritas opini masyarakat terhadap kebijakan pembelajaran daring bersifat positif, dengan proporsi sebesar 56,24% sentimen positif dan 43,76% sentimen negatif. Penelitian ini memperlihatkan efektivitas algoritma *machine learning* dalam menganalisis sentimen publik dari data media sosial, serta menegaskan pentingnya tahap *text preprocessing* dan pembobotan fitur untuk meningkatkan performa model klasifikasi(Supriyanto et al., 2023).

2.18.9. Perancangan Aplikasi Analisis Sentimen pada Ulasan Maskapai Penerbangan di *Tripadvisor* Menggunakan Metode *Naïve Bayes* Oleh Christian dan Fenriana (2023)

Tujuan dari studi yang ditulis oleh Christian dan Fenriana (2023), "*Perancangan Aplikasi Analisis Sentimen pada Ulasan Maskapai Penerbangan di Tripadvisor Menggunakan Metode Naïve Bayes*" adalah untuk membuat sistem aplikasi yang dapat mengkategorikan ulasan pengguna menjadi ulasan positif atau negatif.Penelitian ini menekankan pentingnya analisis sentimen dalam menyaring informasi dari banyaknya ulasan pengguna mengenai layanan maskapai penerbangan yang ada di platform *Tripadvisor*.

Penelitian ini mengumpulkan 3.356 ulasan dari tiga maskapai penerbangan: *Citilink, Lion Air*, dan *Garuda Indonesia*. *Cleaning, case folding, tokenizing, filtering, dan stemming* adalah beberapa langkah dalam proses *preprocessing* teks. Setelah data dibersihkan, kamus sentimen digunakan untuk memberikan label dan algoritma *Naive Bayes* digunakan untuk memprosesnya.

Untuk menilai model, tiga skenario digunakan untuk membagi data latihan dan data uji, yaitu 70:30, 80:20, dan 90:10. Skenario 90:10 menunjukkan hasil terbaik dengan akurasi sebesar 77%, *presisi* sebesar 77%, *recall* sebesar 100%, dan skor F1 sebesar 87,1%.

Penelitian ini menunjukkan bahwa algoritma *Naïve Bayes* efektif digunakan untuk klasifikasi sentimen pada ulasan pengguna berbahasa Indonesia. Selain itu, aplikasi yang dikembangkan juga dilengkapi dengan visualisasi data, berupa grafik batang dan grafik lingkaran, yang memudahkan pengguna dalam memahami persepsi umum terhadap masing-masing maskapai. Studi ini memberikan kontribusi signifikan terhadap pengembangan sistem berbasis analisis sentimen yang aplikatif dan bermanfaat dalam proses pengambilan keputusan berbasis ulasan pengguna(Christian & Fenriana, 2023).

2.18.10. Penerapan *Text Mining* Dalam Menganalisis Pendapat Masyarakat Terhadap Pemilu 2024 Pada Media Sosial X Menggunakan Metode *Naive Bayes* Oleh Mauliza & Sipayung (2024)

Jurnal berjudul "*Penerapan Text Mining dalam Menganalisis Pendapat Masyarakat terhadap Pemilu 2024 pada Media Sosial X Menggunakan Metode Naive Bayes*" ditulis oleh Mauliza dan Sipayung (2024). Tujuan dari penelitian

ini adalah untuk mengklasifikasikan pendapat masyarakat tentang pemilihan 2024 berdasarkan data ulasan yang diperoleh dari media sosial X (sebelumnya dikenal sebagai *Twitter*). Penelitian ini menggunakan teknik *text mining* dan algoritma klasifikasi *Naïve Bayes* untuk mengidentifikasi dan memahami persepsi publik terhadap masalah politik nasional.

Data yang digunakan terdiri dari 300 tweet yang dikumpulkan menggunakan library *tweet-harvest* dengan filter hashtag seperti #pemilu2024 dan #Pilpres2024. Proses *preprocessing* data dilakukan melalui tahapan *case folding*, *cleansing*, *tokenizing*, *filtering*, dan *stemming*, yang seluruhnya diimplementasikan dalam aplikasi *RapidMiner*. Data yang telah dibersihkan kemudian dilabeli secara manual menjadi tiga kategori sentimen, yaitu positif, negatif, dan netral, dengan komposisi 100 data latih dan 200 data uji.

Pembobotan fitur dilakukan menggunakan metode *TF-IDF*, dan proses klasifikasi dilakukan dengan algoritma *Naïve Bayes Classifier*. Hasil pengujian menunjukkan bahwa model mampu mengidentifikasi 103 *tweet* dengan sentimen positif, 47 *tweet* dengan sentimen negatif, dan 50 *tweet* dengan sentimen netral. Temuan ini mengindikasikan bahwa mayoritas masyarakat memberikan tanggapan positif terhadap penyelenggaraan Pemilu 2024.

Penelitian ini menunjukkan bahwa algoritma *Naïve Bayes* efektif dalam menangani klasifikasi teks berbahasa Indonesia, khususnya dalam konteks sosial dan politik. Selain itu, penggunaan *RapidMiner* sebagai alat bantu analisis memberikan kemudahan dalam memvisualisasikan dan memproses data. Studi ini juga menegaskan pentingnya analisis sentimen dalam memahami dinamika opini publik berbasis media sosial secara *real-time*, yang dapat

dimanfaatkan dalam pengambilan keputusan strategis oleh berbagai pihak terkait(Mauliza & Sipayung, 2024).

2.18.11. *A review on sentiment analysis from social media platforms* oleh Rodríguez-Ibánez et al. (2023)

Rodríguez-Ibáñez et al. (2023) dalam jurnal berjudul *"A Review on Sentiment Analysis from Social Media Platforms"* menyajikan tinjauan menyeluruh terhadap berbagai metode, tantangan, dan penerapan analisis sentimen dalam konteks media sosial. Penelitian ini tidak hanya merangkum metode tradisional seperti pendekatan *leksikon* dan *machine learning*, tetapi juga membahas perkembangan terkini seperti penggunaan model berbasis *Transformer* (misalnya *BERT, T5, GPT-3*) serta analisis temporal dan kausalitas dalam domain analisis sentimen.

Dalam studi ini, para penulis menyoroti pentingnya dinamika temporal, yaitu perubahan sentimen dari waktu ke waktu, serta pendekatan kausalitas, terutama *Granger causality*, untuk memahami hubungan antara perubahan sentimen dengan variabel eksternal, seperti pergerakan pasar saham atau opini politik. Selain itu, mereka mengevaluasi tantangan reproduktibilitas dalam penelitian analisis sentimen dan kesenjangan antara dunia akademik dan industri, yang tercermin dari perbedaan penggunaan metode dan kompleksitas model.

Penelitian ini juga mencatat bahwa meskipun model berbasis *Transformer* menunjukkan kinerja unggul, penerapannya dalam praktik masih terbatas oleh kebutuhan sumber daya komputasi yang tinggi dan keterbatasan

aksesibilitas. Berdasarkan analisis lebih dari 2.300 publikasi dan 8.000 paten selama lima tahun terakhir, ditemukan bahwa aplikasi analisis sentimen paling banyak diterapkan pada bidang pemasaran, politik, dan ekonomi. Namun demikian, sektor-sektor seperti kesehatan dan manajemen bencana masih relatif kurang dieksplorasi.

Kontribusi utama dari penelitian ini adalah menekankan pentingnya pendekatan temporal dan kausal dalam meningkatkan akurasi dan relevansi hasil analisis sentimen. Selain itu, makalah ini menyarankan perlunya penelitian lanjutan terkait penyusunan model analisis sentimen yang lebih dapat direplikasi dan diaplikasikan secara luas, baik dalam ranah akademik maupun industri(Rodríguez-Ibánez et al., 2023).

2.18.12. Eksplorasi Algoritma *Support Vector Machine* untuk Analisis Sentimen Destinasi Wisata di Indonesia Oleh Junaedi et al. (2024)

Tujuan dari penelitian berjudul "*Eksplorasi Algoritma Support Vector Machine untuk Analisis Sentimen Destinasi Wisata di Indonesia*" oleh Junaedi, Gunawan, Kuswanto, dan Jonathan (2024) adalah untuk mengetahui seberapa baik algoritma *Support Vector Machine (SVM)* berfungsi untuk menganalisis ulasan wisatawan yang diambil dari platform media sosial *Twitter* (*X*). Fokus utama penelitian ini adalah penggunaan *text mining* untuk mengidentifikasi persepsi wisatawan dalam tiga kategori sentimen: positif, negatif, dan netral. Tujuannya adalah untuk membantu proses pengambilan keputusan berbasis data dalam manajemen industri pariwisata.

Data ulasan yang dikumpulkan melalui *API Twitter* kemudian diproses melalui teknik *preprocessing* teks seperti *casefolding, tokenizing, dan stemming*.Metode *TF-IDF* digunakan untuk memembobot kata, dan kemudian klasifikasi dilakukan menggunakan algoritma *SVM* yang menggunakan tiga jenis *kernel: linear, radial basis function (RBF)*, dan *sigmoid*. Selain itu, penelitian ini memeriksa rasio data pelatihan dan pengujian (70:30 dan 80:20) untuk mengetahui dampaknya terhadap kinerja model.

Hasil penelitian menunjukkan bahwa *SVM* dengan *kernel linear dan rasio* data 70:30 memiliki performa terbaik dengan akurasi 92,89%, ketepatan 92,89%, *recall* 74%, dan *F1-score* 81%. Menurut evaluasi berdasarkan kelas sentimen, model memiliki kinerja tinggi pada sentimen positif (*F1-score* 96%), kinerja cukup baik pada sentimen negatif (*F1-score* 80%), tetapi kinerja masih moderat pada sentimen netral (*F1-score* 67%), yang mungkin disebabkan oleh ketidakseimbangan distribusi data.

Studi ini menunjukkan bahwa algoritma *SVM* sangat cocok untuk klasifikasi sentimen dalam industri pariwisata. Analisis sentimen dapat membantu manajer destinasi wisata memahami opini wisatawan, membuat strategi promosi yang lebih baik, dan meningkatkan layanan dengan data *real-time*(Junaedi et al., 2024).

2.18.13. *Aspect-Based Sentiment Analysis of Online Marketplace Reviews Using Convolutional Neural Network* Oleh Ari Bangsa M, Priyanta S, Suyanto Y (2020)

Penelitian ini mengungkapkan bahwa penggunaan *Convolutional Neural Network (CNN)* untuk analisis sentimen berbasis aspek dalam ulasan *marketplace online* menunjukkan hasil yang signifikan. *CNN*, yang dipadukan dengan ekstraksi fitur menggunakan *Word2Vec*, berhasil mengidentifikasi aspek-aspek spesifik dari produk yang dievaluasi oleh konsumen dan mengklasifikasikan sentimen yang terkait dengan aspek tersebut (positif, negatif, atau netral). Model ini dapat menangani kompleksitas data teks dengan lebih efektif dibandingkan dengan metode tradisional, memberikan wawasan yang lebih mendalam tentang pendapat konsumen terhadap produk. Penelitian ini juga menekankan pentingnya pengolahan dan pemilihan data untuk mencapai kinerja yang optimal dalam analisis sentimen berbasis aspek. Temuan ini membuka potensi penggunaan *CNN* dalam sistem analisis sentimen yang lebih kompleks dan mendalam untuk platform *e-commerce*, yang dapat membantu pengambilan keputusan bisnis yang lebih tepat(Ari Bangsa et al., 2020).

2.18.14. Sentimen Analisis Aplikasi *E-Commerce* Berdasarkan Ulasan Pengguna Menggunakan Algoritma *Stochastic Gradient Descent* Oleh Kurniawan & Pradana (2023)

Dalam penelitiannya yang berjudul "*Sentimen Analisis Aplikasi E-Commerce Berdasarkan Ulasan Pengguna Menggunakan Algoritma Stochastic Gradient Descent*", Kurnia (2023) mengklasifikasikan ulasan pengguna untuk

aplikasi *Tokopedia* dan *Shopee* yang diperoleh dari *Google Play Store*. Tujuan utama dari penelitian ini adalah untuk mengelompokkan ulasan pengguna ke dalam dua kategori sentimen, yaitu positif dan negatif, menggunakan algoritma *Stochastic Gradient Descent.*

Data ulasan dikumpulkan melalui teknik *web scraping* untuk periode Maret hingga Mei 2021, dengan masing-masing platform menyumbang sebanyak 10.000 data ulasan. Proses analisis dimulai dengan tahapan *text preprocessing*, pelabelan data secara manual, dan penerapan algoritma *SGD* untuk proses klasifikasi.Hasil evaluasi model menunjukkan performa yang bervariasi antara kedua aplikasi Untuk *Tokopedia*, model menghasilkan akurasi sebesar 84%, *precision* 87%, dan *recall* 90%.Sedangkan pada Shopee, akurasi model hanya mencapai 66%, dengan *precision* 65% dan *recall* 66%.

Penelitian ini juga menampilkan visualisasi kata dalam bentuk *word cloud* untuk mengilustrasikan kata-kata yang paling sering muncul dalam ulasan positif dan negatif. Temuan penelitian menunjukkan bahwa opini pengguna dapat direpresentasikan secara efektif melalui pendekatan *text mining*, dan bahwa algoritma *SGD* cukup efisien dalam menangani data dalam jumlah besar. Meski begitu, perbedaan performa antar platform mengindikasikan adanya pengaruh konteks data terhadap efektivitas model klasifikasi.

2.18.15. *On the optimal scheme of the possible three predictors in discrete optimization problems using decision-making algorithms* Oleh Yulia Terentyeva32 (2024)

Terentyeva (2024), dalam artikelnya yang berjudul *"On the Optimal Scheme of the Possible Three Predictors in Discrete Optimization Problems Using Decision-Making Algorithms"*, membahas strategi optimal dalam penggunaan fungsi prediktor pada penyelesaian masalah optimisasi diskrit melalui algoritma pengambilan keputusan. Fokus utama penelitian ini terletak pada efektivitas proses pengambilan keputusan ketika satu hingga tiga prediktor digunakan untuk memberikan masukan secara iteratif dalam tahapan optimisasi.

Dalam studi ini, penulis mengembangkan dan membandingkan berbagai skema pengambilan keputusan seperti *unanimous voting* dan *majority voting*, baik pada kelompok prediktor yang memiliki probabilitas keberhasilan yang seragam maupun berbeda. Dengan memanfaatkan pendekatan probabilistik dan teori keputusan, skema-skema tersebut disusun berdasarkan kombinasi probabilitas dari tiap prediktor dan dianalisis melalui simulasi matematis serta integrasi multivariat.

Hasil analisis menunjukkan bahwa secara umum, penggunaan satu prediktor tunggal dengan probabilitas keberhasilan tertinggi menghasilkan performa terbaik dibandingkan dengan pendekatan kolektif seperti *voting* mayoritas atau konsensus. Penelitian ini juga didukung oleh pembahasan matematis yang mendalam, termasuk penerapan teorema Weierstrass, fungsi

ekstrem, matriks Hessian, serta metode integrasi ganda untuk mengevaluasi wilayah preferensi antar strategi pengambilan keputusan.

Kontribusi utama dari penelitian ini adalah penegasan bahwa penambahan jumlah prediktor tidak selalu menghasilkan peningkatan kualitas keputusan, dan bahkan dapat menambah beban komputasi secara signifikan. Temuan ini relevan untuk diterapkan dalam pengembangan sistem pendukung keputusan, baik yang berbasis perangkat lunak maupun sistem penilaian yang melibatkan opini manusia, khususnya dalam konteks penyaringan biner berbasis logika prediktif (Yulia Terentyeva32, 2024).

BAB III

METODOLOGI PENELITIAN

3.1. Analisa Kebutuhan

Untuk mendapatkan pemahaman yang lebih baik tentang tanggapan masyarakat, penelitian ini akan menganalisis opini yang belum pernah diteliti sebelumnya. Fokus utama penelitian ini adalah melihat bagaimana orang-orang, terutama mereka yang menggunakan aplikasi *X*, berperilaku terhadap program makan siang gratis. Untuk memperoleh pemahaman yang lebih baik tentang tanggapan masyarakat, penelitian ini akan menganalisis opini-opini yang belum pernah diteliti sebelumnya dengan menggunakan metode klasifikasi data *Naive Bayes*. Diharapkan analisis ini akan memberikan gambaran yang lebih jelas tentang persepsi publik terhadap program tersebut.

3.1.1. Analisa Kebutuhan Pemakai

Dalam implementasi program makan siang dan distribusi susu gratis yang diinisiasi oleh pemerintah, analisis kebutuhan pengguna diarahkan pada pemahaman terhadap persepsi publik, khususnya dari kalangan pengguna aplikasi Media *X*. Para pengguna platform tersebut secara aktif memberikan opini dan tanggapan melalui kolom ulasan dan komentar, yang berpotensi menjadi sumber data yang bernilai untuk menilai efektivitas program. Namun demikian, tingginya volume data opini tersebut menjadikan proses analisis secara manual kurang efisien dan berisiko menimbulkan subjektivitas dalam interpretasi. Oleh karena itu, untuk mengotomatisasi proses klasifikasi opini ke dalam kategori sentimen positif dan negatif diperlukan penggunaan sistem

analisis sentimen berbasis pembelajaran mesin seperti algoritma *Naive Bayes*. Diharapkan bahwa metode ini dapat memberikan gambaran yang lebih cepat dan jujur tentang bagaimana masyarakat bertindak terhadap kebijakan yang diterapkan.Maka berikut kebutuhan didalam aplikasi:

Tabel 3. 1 Analisis Kebutuhan Pemakai

No	Analisis Kebutuhan Pemakai Saya Berharap Sistem Dapat :	Keterangan
1	Mampu Menampilkan hasil prediksi dengan tepat dan dapat dipercaya.	v
2	Tampilan Aplikasi yang sederhana sehingga mudah di pahami	v
3	Melihat dan memahami sentimen publik terhadap program pemerintah.	v

Berdasarkan Analisis Kebutuhan pemakai,Maka dikembangkan aplikasi untuk memenuhi kebutuhan pengguna.Berikut ini adalah tabel untuk menggambarkan kebutuhan aplikasi:

Tabel 3. 2 Analisis Kebutuhan aplikasi

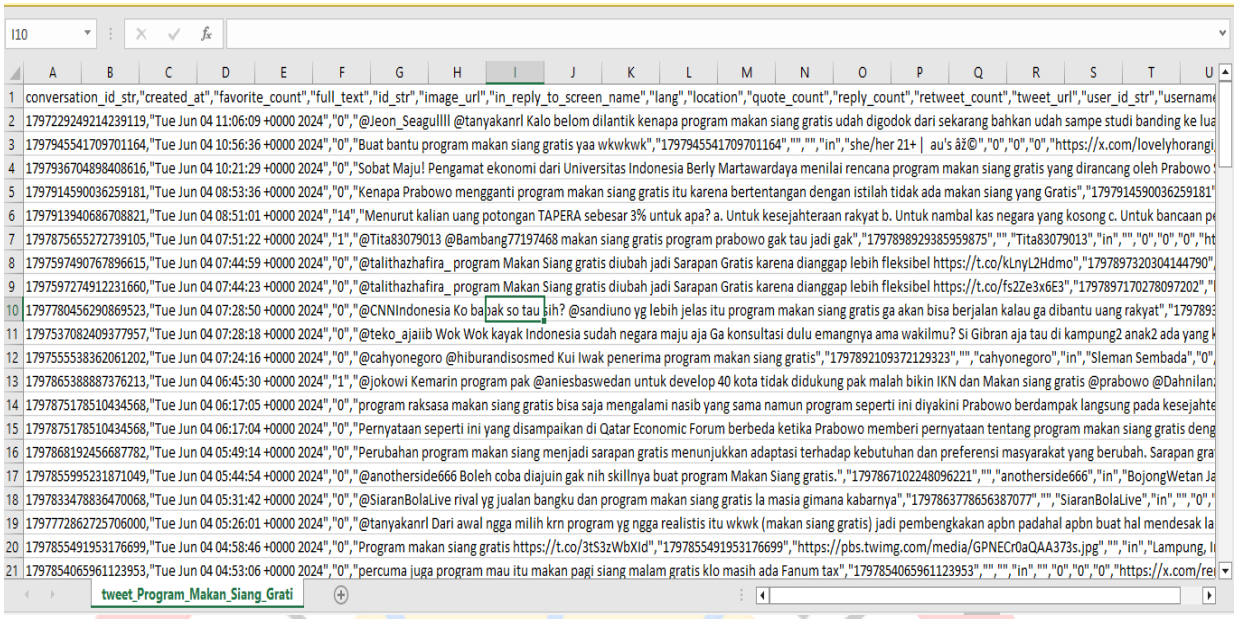
No	Analisis Kebutuhan Aplikasi Saya Berharap Sistem Dapat :	Keterangan
1	Memberikan Prediksi Sentimen Positif dan negatif.	v
2	Tampilan Aplikasi mudah digunakan dan di pahami.	v
3	Menampilkan hasil analisis dalam bentuk tabel dan grafik.	v
4	Login dan logout pengguna (admin atau pengguna umum).	v
5	Pengguna dapat mengunggah dataset ulasan dalam format CSV.	v
6	Menampilkan akurasi model berdasarkan data uji.	v

3.2. Metode Yang Digunakan

3.2.1. Pengumpulan Data

Pada tahap pengumpulan data, informasi yang digunakan berupa teks cuitan yang diperoleh dari media sosial *Twitter*. Pengambilan data dilakukan secara otomatis menggunakan pemrograman di lingkungan *Google Colab*. Metode yang digunakan adalah *web scraping* terhadap data publik *Twitter* dengan bantuan pustaka *tweet-harvest*, yang dioperasikan melalui autentikasi token untuk mendapatkan akses data.

Langkah awal dalam proses ini melibatkan pemasangan dependensi seperti *Node.js* dan pustaka pendukung lainnya yang diperlukan untuk menjalankan skrip pengambilan data. Setelah lingkungan siap, eksekusi perintah *tweet-harvest* dilakukan dengan menyertakan parameter pencarian seperti kata kunci ("Program Makan Siang dan Susu Gratis"), rentang waktu tertentu, serta batas maksimum jumlah data (*limit*) yang akan dikumpulkan. Hasil dari proses ini berupa file berformat .csv yang memuat isi teks cuitan beserta atribut tambahan seperti tanggal, nama pengguna, dan metadata lainnya.



	conversation_id_str	created_at	favorite_count	full_text	id_str	image_url	in_reply_to_screen_name	lang	location	quote_count	reply_count	retweet_count	tweet_url	user_id_str	username
2	1797229249214239119	Tue Jun 04 11:06:09 +0000 2024	0	@Jeon_SeagulIII @tanyakanrl Kalo belom dilantik kenapa program makan siang gratis udah digodok dari sekarang bahkan udah sampe studi banding ke lua											
3	1797945541709701164	Tue Jun 04 10:56:36 +0000 2024	0	Buat bantu program makan siang gratis yaa wkwkwk	1797945541709701164			in	she/her 21+ | au's âž©	0	0	0	https://x.com/lovelyhorangi		
4	1797936704898408616	Tue Jun 04 10:21:29 +0000 2024	0	Sobat Maju! Pengamat ekonomi dari Universitas Indonesia Berly Martawardaya menilai rencana program makan siang gratis yang dirancang oleh Prabowo											
5	1797914590036259181	Tue Jun 04 08:53:36 +0000 2024	0	Kenapa Prabowo mengganti program makan siang gratis itu karena bertentangan dengan istilah tidak ada makan siang yang Gratis	1797914590036259181										
6	1797913940686708821	Tue Jun 04 08:51:01 +0000 2024	14	Menurut kalian uang potongan TAPERA sebesar 3% untuk apa? a. Untuk kesejahteraan rakyat b. Untuk nambal kas negara yang kosong c. Untuk bancaan pe											
7	1797875655272739105	Tue Jun 04 07:51:22 +0000 2024	1	@Tita83079013 @Bambang77197468 makan siang gratis program prabowo gak tau jadi gak	1797898929385959875			in		0	0	0	0	ht	
8	1797597490767896615	Tue Jun 04 07:44:59 +0000 2024	0	@talithazhafira_ program Makan Siang gratis diubah jadi Sarapan Gratis karena dianggap lebih fleksibel https://t.co/kLnyL2Hdmo	1797897320304144790										
9	1797597274912231660	Tue Jun 04 07:44:23 +0000 2024	0	@talithazhafira_ program Makan Siang gratis diubah jadi Sarapan Gratis karena dianggap lebih fleksibel https://t.co/fs2Ze3x6E3	1797897170278097202										
10	1797780456290869523	Tue Jun 04 07:28:50 +0000 2024	0	@CNNIndonesia Ko bapak so tau sih? @sandiuno yg lebih jelas itu program makan siang gratis ga akan bisa berjalan kalau ga dibantu uang rakyat	1797893										
11	1797537082409377957	Tue Jun 04 07:28:18 +0000 2024	0	@teko_ajaiib Wok Wok kayak Indonesia sudah negara maju aja Ga konsultasi dulu emangnya ama wakilmu? Si Gibran aja tau di kampung2 anak2 ada yang k											
12	1797555538362061202	Tue Jun 04 07:24:16 +0000 2024	0	@cahyonegoro @hiburandisosmed Kui Iwak penerima program makan siang gratis	1797892109372129323				cahyonegoro					in	Sleman Sembada
13	1797865388887376213	Tue Jun 04 06:45:30 +0000 2024	1	@jokowi Kemarin program pak @aniesbaswedan untuk develop 40 kota tidak didukung pak malah bikin IKN dan Makan siang gratis @prabowo @Dahnilan											
14	1797875178510434568	Tue Jun 04 06:17:05 +0000 2024	0	program raksasa makan siang gratis bisa saja mengalami nasib yang sama namun program seperti ini diyakini Prabowo berdampak langsung pada kesejahte											
15	1797875178510434568	Tue Jun 04 06:17:04 +0000 2024	0	Pernyataan seperti ini yang disampaikan di Qatar Economic Forum berbeda ketika Prabowo memberi pernyataan tentang program makan siang gratis deng											
16	1797868192456687782	Tue Jun 04 05:49:14 +0000 2024	0	Perubahan program makan siang menjadi sarapan gratis menunjukkan adaptasi terhadap kebutuhan dan preferensi masyarakat yang berubah. Sarapan gra											
17	1797855995231871049	Tue Jun 04 05:44:54 +0000 2024	0	@anotherside666 Boleh coba diajuin gak nih skillnya buat program Makan Siang gratis.	1797867102248096221				anotherside666					in	BojongWetan Ja
18	1797833478836470068	Tue Jun 04 05:31:42 +0000 2024	0	@SiaranBolaLive rival yg jualan bangku dan program makan siang gratis la masia gimana kabarnya	1797863778656387077				SiaranBolaLive					in		0
19	1797772862725706000	Tue Jun 04 05:26:01 +0000 2024	0	@tanyakanrl Dari awal ngga milih krn program yg ngga realistis itu wkwk (makan siang gratis) jadi pembengkakan apbn padahal apbn buat hal mendesak la											
20	1797855491953176699	Tue Jun 04 04:58:46 +0000 2024	0	Program makan siang gratis https://t.co/3tS3zWbXId	1797855491953176699	https://pbs.twimg.com/media/GPNECr0aQAA373s.jpg								in	Lampung,
21	1797854065961123953	Tue Jun 04 04:53:06 +0000 2024	0	percuma juga program mau itu makan pagi siang malam gratis klo masih ada Fanum tax	1797854065961123953									in		0

tweet_Program_Makan_Siang_Grati

Gambar 3. 1 Hasil *Scraping* Data *Tweet*

3.2.2. Teks *Preprocessing* Data

Pemosresan teks merupakan bagian penting dari analisis sentimen, *Preprocessing* adalah langkah penting yang harus diselesaikan sebelum menerapkan algoritma klasifikasi pada sebuah dokumen.(Rahel Lina Simanjuntak et al., 2023) Dalam tahapan *proprocessing* ini dimulai dari tahapan yaitu,*case folding , noise removal, tokenization, filtering (stopword removal)* serta *stemming*.

A. Noise Removal

Noise Removal adalah proses yang digunakan untuk menghilangkan karakter atau simbol non-teks, seperti angka, tanda baca, URL, *emotikon*, serta spasi yang berlebihan. Dalam langkah ini, data dibersihkan dari elemen yang tidak penting atau tidak berguna untuk analisis.URL atau simbol-simbol khusus sering kali tidak memberikan kontribusi berarti dalam proses analisis teks dan justru dapat mengganggu model pembelajaran mesin. Dengan

membersihkan data dari "*noise*" tersebut, data training menjadi lebih terstruktur dan relevan, sehingga mampu mendukung analisis teks yang lebih akurat sesuai dengan kebutuhan model pembelajaran mesin.Berikut adalah contoh data teks sebelum dan sesudah di lakukan *cleaning* teks.

Tabel 3. 3 Tabel contoh teks sebelum di lakukan *noise removal*

@RikeSants Program yangg di jual 01 dan 03 di fahami oleh kalangan akademis sedang 02 jualan nya menengah kebawah dengan makan siang gratis gemoy nya.siapa yang menang ide dan program bagus pasti dipakai
@mimih6mei siap kawal program makan siang gratis sih. karena ini program bagus banget apalagi buat gizi anak2.
@pro_gibran_ Program makan siang gratis Gak cuma bermanfaat untuk anak anak ya tapi utk berbagai lapisan masyarakat.
@bakanosan1 @PNS_Garis_Lucu @KemenkeuRI Cocok ini untuk program makan siang gratis wkwk....
@Andiarief__ Program makan siang gratis gak dikawal ni besar anggarannya jadi gak jadi gitu ?.

Tabel 3. 4 Tabel contoh teks sesudah dilakukan *noise removal*

Program yang di jual dan di fahami oleh kalangan akademis sedang jualan nya menengah kebawah dengan makan siang gratis gemoy nya siapa yang menang ide dan program bagus pasti dipakai
siap kawal program makan siang gratis sih karena ini program bagus banget apalagi buat gizi anak

Program makan siang gratis Gak cuma bermanfaat untuk anak anak ya tapi untuk berbagai lapisan masyarakat
Cocok ini untuk program makan siang gratis wkwk
Program makan siang gratis gak dikawal ni besar anggarannya jadi gak jadi gitu

B. *Case Folding*

Proses ini dilakukan untuk memastikan teks konsisten dengan mengubah semua huruf menjadi huruf kecil. Tujuan dari proses ini adalah untuk menyamakan format teks sehingga sistem lebih mudah memprosesnya, terutama untuk analisis data. Proses ini juga mencakup penghapusan tanda baca, simbol, atau karakter tertentu yang dianggap tidak relevan atau tidak diperlukan untuk analisis.Dengan langkah ini, data teks dapat disederhanakan dan difokuskan hanya pada informasi yang penting, sehingga meningkatkan akurasi dalam pengolahan data selanjutnya(Rahel Lina Simanjuntak et al., 2023).Berikut merupakah output dari text yang telah dilakukan proses *case folding*.

Tabel 3. 5 contoh teks sebelum dilakukan *case folding*

Program yang di jual dan di fahami oleh kalangan akademis sedang jualan nya menengah kebawah dengan makan siang gratis gemoy nya siapa yang menang ide dan program bagus pasti dipakai
siap kawal program makan siang gratis sih karena ini program bagus banget apalagi buat gizi anak
Program makan siang gratis Gak cuma bermanfaat untuk anak anak ya tapi untuk berbagai lapisan masyarakat

Cocok ini untuk program makan siang gratis wkwk
Program makan siang gratis gak dikawal ni besar anggarannya jadi gak jadi gitu

Tabel 3. 6 contoh teks sesudah dilakukan case folding

program yang di jual dan di fahami oleh kalangan akademis sedang jualan nya menengah kebawah dengan makan siang gratis gemoy nya siapa yang menang ide dan program bagus pasti dipakai
siap kawal program makan siang gratis sih karena ini program bagus banget apalagi buat gizi anak
program makan siang gratis gak cuma bermanfaat untuk anak anak ya tapi untuk berbagai lapisan masyarakat
cocok ini untuk program makan siang gratis wkwk
program makan siang gratis gak dikawal ni besar anggarannya jadi gak jadi gitu

C. *Filtering(stopword)*

Penghapusan *stopword*, yaitu kata-kata umum yang sering muncul dalam teks tetapi tidak memiliki makna khusus atau nilai informatif dalam analisis teks, adalah langkah selanjutnya dalam proses *preprocessing*.*Stopword* biasanya meliputi kata seperti "di", "ke", "dari", "yang", dan "dan", yang meskipun sering digunakan, tidak berkontribusi signifikan terhadap proses pengolahan data. Penghapusan *stopword* dilakukan untuk meningkatkan efisiensi analisis, sehingga model dapat lebih fokus pada kata-kata yang

benar-benar relevan dalam memahami konteks dan makna teks(Muzaki et al., 2024).Dalam penelitian ini, daftar *stopword* yang digunakan diambil dari dataset yang tersedia di *Kag gle* https://www.kaggle.com/datasets/oswinrh/indonesianstoplist. Y ang berisikan list list kata *stopword*.Dengan langkah ini, teks dapat dianalisis dengan lebih efisien dan relevan terhadap tujuan klasifikasi sentimen.Berikut ini adalah contoh list list kata *stopword*:

Tabel 3. 7 Daftar List Kata *Stopword*

Daftar List Kata *Stopword*	
ada	akhiri
adalah	akhirnya
adanya	aku
adapun	akulah
agak	amat
agaknya	amatlah
agar	anda
akan	andalah
akankah	antar
akhir	antara

Setelah list kata *stopword* ditentukan, tahap selanjutnya adalah melakukan eliminasi kata-kata tersebut dari setiap kalimat

dalam data ulasan. Proses ini bertujuan untuk menyaring kata-kata yang tidak memiliki kontribusi signifikan terhadap makna sentimen yang dianalisis, sehingga tahap ekstraksi fitur dapat dilakukan secara lebih efisien dan terfokus. Dengan menghilangkan kata-kata yang bersifat umum dan kurang informatif, model analisis dapat lebih mudah mengenali pola sentimen yang relevan dalam teks.

Tabel 3. 8 contoh teks sebelum di*stopword*

Program yang di jual dan di fahami oleh kalangan akademis sedang jualan nya menengah kebawah dengan makan siang gratis gemoy nya siapa yang menang ide dan program bagus pasti dipakai
siap kawal program makan siang gratis sih karena ini program bagus banget apalagi buat gizi anak
program makan siang gratis gak cuma bermanfaat untuk anak anak ya tapi untuk berbagai lapisan masyarakat
cocok ini untuk program makan siang gratis wkwk
program makan siang gratis gak dikawal ni besar anggarannya jadi gak jadi gitu

Tabel 3. 9 contoh teks sesudah di*stopword*

program jual fahami kalangan akademis jualan menengah makan siang gratis gemoy menang ide program bagus dipakai
kawal program makan siang gratis program bagus gizi anak
program makan siang gratis bermanfaat anak anak lapisan masyarakat
cocok program makan siang gratis
program makan siang gratis dikawal anggarannya

D. Tokenization

Tokenisasi adalah proses memecah atau membagi sebuah kalimat menjadi bagian-bagian kecil yang disebut *term*. Pemisahan ini dilakukan berdasarkan tanda spasi sebagai delimiter, sehingga setiap kata dalam kalimat dapat diidentifikasi sebagai unit analisis yang terpisah. Langkah ini penting untuk mempersiapkan teks agar dapat diproses lebih lanjut dalam analisis teks atau pemodelan pembelajaran mesin(Rahel Lina Simanjuntak et al., 2023).Berikut merupakan contoh teks yang sudah *ditokenisas*i.

Tabel 3. 10 contoh data teks yang sudah di *tokenisasi*

TOKENIZATION
['program', 'jual', 'fahami', 'kalangan', 'akademis', 'jualan', 'menengah', 'makan', 'siang', 'gratis', 'gemoy', 'menang', 'ide', 'program', 'bagus', 'dipakai']
['kawal', 'program', 'makan', 'siang', 'gratis', 'program', 'bagus', 'gizi', 'anak']

['program', 'makan', 'siang', 'gratis', 'bermanfaat', 'anak', 'anak', 'lapisan', 'masyarakat']
['cocok', 'program', 'makan', 'siang', 'gratis']
['program', 'makan', 'siang', 'gratis', 'dikawal', 'anggarannya']

E. *Stemming*

Stemming adalah proses yang dimaksudkan untuk menghilangkan bentuk kata yang tidak sesuai dengan standar Kamus Besar Bahasa Indonesia (KBBI). Proses ini mencakup penghapusan imbuhan, kata depan, kata ganti, dan akhiran, sehingga setiap kata dapat kembali ke bentuk dasarnya. *Library Sastrawi*, yang dirancang khusus untuk stemming teks berbahasa Indonesia, digunakan untuk melakukan proses stemming dalam penelitian ini. Teks menjadi lebih standarisasi dengan penggunaan *stemming*, yang memudahkan analisis dan meningkatkan akurasi pemrosesan data teks (Rahel Lina Simanjuntak et al., 2023).

Tabel 3. 11 contoh teks sesudah dilakukan Stemming

STEMMING
['program', 'jual', 'faham', 'kalang', 'akademi', 'jual', 'tengah', 'makan', 'siang', 'gratis', 'gemoy', 'menang', 'ide', 'program', 'bagus', 'pakai']
['kawal', 'program', 'makan', 'siang', 'gratis', 'program', 'bagus', 'gizi', 'anak']
['program', 'makan', 'siang', 'gratis', 'manfaat', 'anak', 'anak', 'lapis', 'masyarakat']

['cocok', 'program', 'makan', 'siang', 'gratis']
['program', 'makan', 'siang', 'gratis', 'kawal', 'anggar']

3.2.3. Pelabelan *Leksikon*

Setelah data melewati beberapa tahapan *preprocessing* teks, pelabelan sentimen dilakukan menggunakan teknik berbasis *leksikon*. Metode ini dilakukan dengan membandingkan setiap kata dalam ulasan dengan daftar kata yang dikategorikan ke dalam dua kategori sentimen: positif dan negatif.Bing Liu dan Minqing Hu membuat *Opinion Lexicon*, kamus kata sentimen yang digunakan dalam penelitian ini. Kamus ini telah diubah dan disesuaikan untuk konteks Bahasa Indonesia oleh Wahid,D.H. dan Azhari,S.N., dan terdiri dari 2.408 kata positif dan 1.182 kata negatif.Pelabelan dilakukan menggunakan *Python*, bahasa pemrograman yang dijalankan di lingkungan *Google Colab*. Sistem dimaksudkan untuk menghitung berapa banyak kata yang memiliki nilai positif dan negatif dalam setiap teks ulasan. Komentar diberi label positif jika lebih banyak kata positif, dan lebih sedikit kata negatif.Dalam kasus di mana jumlah kata positif dan negatif seimbang atau tidak ditemukan kecocokan dengan kamus, ulasan tidak dilabeli lebih lanjut, karena dalam penelitian ini hanya digunakan dua kategori sentimen.Sebagai ilustrasi, berikut ditampilkan sebagian contoh dari daftar kata sentimen yang digunakan dalam proses pelabelan.

Tabel 3. 12 Contoh Daftar Kata Sentimen Positif dan Negatif

Kata Bersentimen Positif		Kata Bersentimen Negatif	
acungan jempol	aman	anak nakal	antipati
adaptif	amanah	anak yatim	antisosial
adil	amat	anarki	antitesis
afinitas	ambisius	anarkis	apak
afirmasi	andal	anarkisme	apati
agilely	bagus	ancaman	apatis
agung	bahagia	aneh	apek
ahli	baik	aneh lagi	apokaliptik
ahlinya	bakat	anehnya	apologis
ajaib	bangga	angkuh	argumentatif

Melalui pendekatan ini, proses pelabelan dapat dilakukan secara otomatis dan efisien terhadap data ulasan dalam jumlah besar. Hasil pelabelan ini selanjutnya digunakan sebagai data *training* dan data *testing* dalam pembangunan model klasifikasi sentimen menggunakan algoritma seperti *Naive Bayes*.

A	B
ulasan	label
Kayaknya program makan siang gratis itu juga gitu rakyat dibuat nyaman untuk melanggengkan kekuasaan	positif
@RikeSants Program yg di jual 01 dan 03 di fahami oleh kalangan akademis sedang 02 jualan nya menengah kebawah dgn makan siang gratis gemoy nya..siapa yg menang ide dan program bagus pasti dipakai.	positif
Jokowi dan para menterinya mulai mengutak-atik anggaran untuk program makan siang gratis. Tidak patut dan berisiko besar. #MajalahTempo https://t.co/oLOtjGV6VA	negatif
Apalagi ada program makan siang gratis. Menggila gk tuh?	negatif
@araaa2236 @p3d3584n937 @Bull_winner Kan emang jokowi 3 periode versi second account hahaha Jika 02 jadi presiden dan wakil presiden. Visi misinya yang beda cuma makan siang gratis sama naikin tax ratio 23% selebih nya melanjutkan program pak jokowi	positif
@dreamaker202 @henrysubiakto @kasito17 MAKAN SIANG GRATIS PROGRAM NGEMIS MINTA MAKAN GAK MENDIDIK SAMA SEKALI SAMA SEPERTI PROGRAM BANSOS PEMERINTAH... https://t.co/AvcFdIuuEM	positif
@DrEvaChaniago Diajak milih calon pemimpin perubahan dengan program pendidikan gratis biar cerdas malah ikut pilih melanjutkan yg punya program makan siang gratis akhirnya ya begini nasib bangsa kedepannya.	positif
hanya orang gak waras yg gak dukung program makan siang gratis untuk anak di sekolah biarkan anak tersenyum bahagia karena mereka bisa makan siang gratis di sekolah nanti nya https://t.co/EX0lR8ntdT	positif
Gambaran makan siang gratis.. nnti kalo ganti presiden ganti program bakal kena marah masyarakat yg ketergantungan padahal anggaran abis xixixi	negatif
Buka Suara Soal Isu Pengurangan KJMU DPRD DKI Singgung Program Makan Siang Gratis Prabowo-Gibran https://t.co/PfarnwtzNv	negatif
tentang free lunch setelah semua mahal Makan siang gratis ini bakal orang bergantung ke dinasti jokowi. Perubahan bakal ditolak rakyat meski dibilang program makan siang tuh ga masuk akal. Ini negara udah ga akan ketolong. Rusak atas bawah	negatif
@expelliarmess Bayangin se f*cked up apa kita kayak semua itu bisa dilakuin tapi malah pilih program makan siang gratis??? Dan banyak yang vote??????????	positif
@affahafra79 @Gamal_Albinsaid Tentunya bukan makan siang gratis karna program ini belum ada. Perjuangan dr gamal dan kedua orangtuanya membuat dr gamal seperti skrng ini. Tentunya juga isnting dr gamal dng memilih partai politik yang baik dan benar.	negatif
Ngomong-ngomong disuapin uang doang jadi inget program makan siang gratis.	positif
Propaganda org besar Amerika. Btw ada dibahas juga nih di channel Sepulang Sekolah perihal siapa yg PALING diuntungkan kalau program makan siang gratis beneran dilaksanakan . Ga persis sih tapi ada lah nyerempetÂ² @sepulangsekolah	negatif
Terkait adanya rencana peningkatan belanja negara ini ada kaitannya sama program dia di masa depan yaitu makan siang gratis. (sebenernya ikn pun bakal ambil sebagian dana APBN kita sih wkwkkw)	positif
@WongAlasRoban Gak perlu ada warteg mungkin karena ada program makan siang gratis tis tis	positif
@tempodotco Sumpah ga setuju bgt makan siang gratis wlpn aku suka makan. Betul2 perilaku konsumtif. Pdhl bisa dialihkan ke program kesehatan atau pendidikan gratis lebih banyak.	positif
Inilah kenapa program makan siang gratis ga akan bikin kita maju	positif
Kami dukung program makan siang dan susu gratis. Mari prihatin dg anak dari saudara kita yang kurang beruntung.	positif
@RAKemal dari sini kita sadar bahwa timnas amin terlalu khusnudzon dengan mayoritas rakyat indo padahal kenyataannya memang mayoritas mental rakyat kita itu ya mental ngemis. pantes program makan siang gratis laku	positif
Menterinya masih ngga yakin sama Program makan siang gratis https://t.co/QuSMVo2DJp	positif
@Juna120821 @mas_veel Program prabowo apa?? Keberlanjutan program jokowi kan terus ini program jokowi yang dilanjutkan sama prabowo makan siang gratis ajah tau2 udah di bahas sama jokowi padahal kan itu program prabowo upss salam keberlanjutan	positif
Ekonom Senior Soroti Program Makan Siang Gratis Prabowo-Gibran: Bukan Ide Jelek Tapi... https://t.co/UIX5d6VVo5	negatif
@Heraloebss @aniesbaswedan @ganjarpranowo Pemerintah Pak Jokowi seimbang antara program kerakyatan dan pembangunan. hanya Prabowo yang menyanggupi melanjutkan program kerakyatan Jokowi dan Hilirisasi dengan ditambah makan siang gratis. pasion lain setenga	positif
@blantik_pedhet2 Pilih yang diusung Golkar. Biar klop sepaket se Indonesia kuningisasi jika menang. Supaya Program kasih makan siang gratis se Indonedia lancar	positif
@Chavesz08 @aniesbaswedan Yang katanya menang ini buzzer masih aja senggol senggol. Mending bantuin cari dana buat program makan siang gratis	positif
@tempodotco Wkwkwk oke gass dibatasin duit anggarannya dialihkan ke program makan siang gratis	positif
@rivdhilah sibuk nyiapin program makan siang gratis	positif

Gambar 3. 2 Hasil Pelabelan Menggunakan *Leksikon*

3.3. Algoritma Yang Digunakan

Dalam pembahasan mengenai algoritma ini, peneliti menyajikan perhitungan manual menggunakan data yang berisi sentimen dengan nilai positif dan negatif, menurut pandangan peneliti. Berikut ini adalah alur pengimplementasian algoritma terhadap studi kasus yang dijadikan fokus dalam penelitian ini. Proses ini menggambarkan langkah-langkah yang dilakukan untuk menerapkan algoritma dalam menganalisis dan mengklasifikasikan sentimen positif dari data ulasan yang sudah di tokenisasi sebagai berikut.

Tabel 3. 13 contoh Teks Ulasan

Ulasan	Label
Sumpah makan siang GRATIS itu PROGRAM AMPAS. Sekolah gue dijadiin percobaan selama dua kali dan halah bikin SUSAH aja.	Negatif
Program makan siang gratis malah memakan korban	Negatif
Yes semua sadar manfaat program makan siang gratis agar generasi muda sehat gak ada kosong perut di sekolah. Generasi cerah. Keren Asik Mantab	Positif
Program Makan Siang Gratis Mulai Terlihat Hasilnya Siswa & Siswi Terlihat Tumbuh Sehat	Positif

Teks ulasan pada tabel sebelumnya masih dalam bentuk asli atau mentah. Untuk mempersiapkannya ke tahap analisis selanjutnya, dilakukan proses pembersihan data (*preprocessing*) agar teks menjadi lebih terstruktur dan representatif. Hasil dari proses ini ditampilkan pada tabel berikut.

Tabel 3. 14 Contoh Teks Sesudah di *Preprocessing*

Sesudah Text *Preprocessing*
sumpah,makan,siang,gratis,progam,ampas,sekolah,coba,bikin,susah
program,makan,siang,gratis,makan,korban
sadar,manfaat,program,makan,siang,gratis,generasi,muda,sehat,kosong,perut,sekolah, generasi,cerah,keren,asik,mantab
program makan siang gratis hasil siswa siswi tumbuh sehat

3.3.1. Proses Pembobotan Kata *TF-IDF*

Setelah proses pemisahan kalimat menjadi token atau kata-kata individual (*tokenization*) selesai dilakukan, langkah selanjutnya adalah melakukan pembobotan kata untuk keperluan klasifikasi. Metode pembobotan yang digunakan dalam penelitian ini adalah *TF-IDF (Term Frequency-Inverse Document Frequency)*.Tahapan awal dalam perhitungan *TF-IDF* adalah menghitung *Term Frequency (TF)*, yaitu seberapa sering suatu kata muncul dalam suatu dokumen dibandingkan dengan total jumlah kata dalam dokumen tersebut. Rumus perhitungan *TF* adalah sebagai berikut:

$$TF(t,d) = \frac{ft,d}{\sum kfk,d}$$

Keterangan:

ft,d = frekuensi kata *t* dalam dokumen *d*

∑kfk,d = total jumlah kata dalam dokumen

Tabel 3. 15 Pembobotan *Term Frequency*

Term	D1 (Negatif)	D2 (Negatif)	D3 (Positif)	D4 (Positif)
sumpah	1			
makan	1	2	1	1
siang	1	1	1	1
gratis	1	1	1	1
program	1	1	1	1
ampas	1			
sekolah	1		1	
coba	1			
bikin	1			
susah	1			
korban		1		
sadar			1	
manfaat			1	
generasi			2	
muda			1	
sehat			1	1
kosong			1	
perut			1	
cerah			1	
keren			1	
asik			1	
mantab			1	

hasil				1
siswa				1
siswi				1
tumbuh				1
Total Term	**10**	**6**	**17**	**9**
Total Bobot	**42**			

Berdasarkan tabel di atas, dapat dilihat nilai *Term Frequency (TF)* yang menunjukkan seberapa sering suatu kata muncul dalam masing-masing dokumen. Nilai *TF* dihitung dengan membagi jumlah kemunculan suatu kata dalam dokumen dengan total jumlah kata dalam dokumen tersebut. Nilai *TF* ini kemudian digunakan sebagai dasar dalam menghitung pembobotan selanjutnya, yaitu *Inverse Document Frequency (IDF)*, guna memperhitungkan tingkat kekhususan sebuah kata terhadap keseluruhan dokumen. Rumus perhitungan *IDF* adalah sebagai berikut:

$$IDF(t) = log \frac{N}{dft}$$

Keterangan :

IDF(t) = inverse document frequency untuk term t

N = total jumlah dokumen

Dft = jumlah dokumen yang mengandung term t

log = biasanya log basis 10, tapi bisa juga natural log (ln), tergantung implementasi

Tabel 3. 16 Pembobotan *Inverse Document Frequency*

Term	D1 (Negatif)	D2 (Negatif)	D3 (Positif)	D4 (Positif)	IDF
sumpah	1				0,60206
makan	1	2	1	1	0
siang	1	1	1	1	0
gratis	1	1	1	1	0
program	1	1	1	1	0
ampas	1				0,60206
sekolah	1		1		0,30103
coba	1				0,60206
bikin	1				0,60206
susah	1				0,60206
korban		1			0,60206
sadar			1		0,60206
manfaat			1		0,60206
generasi			2		0,60206
muda			1		0,60206
sehat			1	1	0,30103
kosong			1		0,60206
perut			1		0,60206
cerah			1		0,60206
keren			1		0,60206
asik			1		0,60206

mantab			1		0,60206
hasil				1	0,60206
siswa				1	0,60206
siswi				1	0,60206
tumbuh				1	0,60206
Total Term	**10**	**6**	**17**	**9**	
Total Bobot	**42**				

Nilai *Inverse Document Frequency (IDF)* pada tabel di atas menggambarkan tingkat keunikan suatu kata dalam seluruh kumpulan dokumen. Semakin jarang sebuah kata muncul di berbagai dokumen, semakin tinggi nilai *IDF*-nya. Sebaliknya, jika sebuah kata muncul di hampir semua dokumen, maka nilai *IDF*-nya akan mendekati nol. Nilai *IDF* ini selanjutnya akan dikalikan dengan nilai *TF* untuk menghasilkan bobot akhir berupa *TF-IDF*. Rumus perhitungan *TF-IDF* adalah sebagai berikut:

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

Keterangan :

TF(t,d) = frekuensi kemunculan term t dalam dokumen d

IDF(t) = inverse document frequency dari term t

Tabel 3. 17 ***Term Frequency Inverse Document Frequency***

Term	D1 (Negatif)	D2 (Negatif)	D3 (Positif)	D4 (Positif)
sumpah	0,60206	0,00000	0,00000	0,00000
makan	0,00000	0,00000	0,00000	0,00000
siang	0,00000	0,00000	0,00000	0,00000
gratis	0,00000	0,00000	0,00000	0,00000
program	0,00000	0,00000	0,00000	0,00000
ampas	0,60206	0,00000	0,00000	0,00000
sekolah	0,30103	0,00000	0,30103	0,00000
coba	0,60206	0,00000	0,00000	0,00000
bikin	0,60206	0,00000	0,00000	0,00000
susah	0,60206	0,00000	0,00000	0,00000
korban	0,00000	0,60206	0,00000	0,00000
sadar	0,00000	0,00000	0,60206	0,00000
manfaat	0,00000	0,00000	0,60206	0,00000
generasi	0,00000	0,00000	1,20412	0,00000
muda	0,00000	0,00000	0,60206	0,00000
sehat	0,00000	0,00000	0,30103	0,30103
kosong	0,00000	0,00000	0,60206	0,00000
perut	0,00000	0,00000	0,60206	0,00000
cerah	0,00000	0,00000	0,60206	0,00000
keren	0,00000	0,00000	0,60206	0,00000
asik	0,00000	0,00000	0,60206	0,00000
mantab	0,00000	0,00000	0,60206	0,00000
hasil	0,00000	0,00000	0,00000	0,60206

siswa	0,00000	0,00000	0,00000	0,60206
siswi	0,00000	0,00000	0,00000	0,60206
tumbuh	0,00000	0,00000	0,00000	0,60206
Total Term	**3,31133**	**0,60206**	**7,22472**	**2,70927**
Total Bobot	**13,84738**			

Tabel *TF-IDF* di atas menunjukkan hasil dari proses pembobotan kata berdasarkan kombinasi antara frekuensi kemunculan kata dalam dokumen (*TF*) dan tingkat keunikan kata dalam seluruh dokumen (*IDF*). Nilai *TF-IDF* yang lebih tinggi menunjukkan bahwa kata tersebut penting dan relevan terhadap dokumen tertentu. Hasil pembobotan ini kemudian digunakan sebagai input fitur untuk proses klasifikasi menggunakan algoritma *Naive Bayes*.

3.3.2. Proses Klasifikasi Data *Training*

Setelah tahap perhitungan bobot untuk setiap *term* selesai dilakukan, langkah selanjutnya adalah menghitung nilai probabilitas untuk masing-masing *term* dalam rangka membangun model klasifikasi berbasis probabilistik, seperti *Naïve Bayes.*

A. Menghitung Nilai Probabilitas *Prior*

Pada tahap ini, dilakukan perhitungan terhadap atribut kelas, yang dalam konteks penelitian ini terdiri dari dua kategori, yaitu positif dan negatif. Nilai probabilitas *prior* (*prior* probability) menggambarkan proporsi masing-masing kelas dalam data pelatihan (*training data*).

$$p(c) = \frac{nc}{n}$$

Dalam rumus tersebut:

P(c) = probabilitas *prior* kelas *c*, yaitu peluang awal sebelum melihat data.

Nc = jumlah dokumen dalam data latih (*training data*) yang termasuk ke dalam kategori atau kelas tertentu, yaitu kelas c (misalnya: positif atau negatif).

N = total jumlah seluruh dokumen dalam data latih.

$$p(Positif) = \frac{2}{4} = 0{,}5$$

$$p(Negatif) = \frac{2}{4} = 0{,}5$$

B. Menghitung *Conditional Probability*

Pada tahap ini, dilakukan perhitungan probabilitas kondisional (*conditional probability*), yaitu menghitung kemungkinan kemunculan setiap kata (*term*) dalam suatu kelas tertentu misalnya positif atau negatif.Proses ini bertujuan untuk mengetahui seberapa besar peluang suatu kata muncul dalam dokumen yang termasuk ke dalam kelas tertentu, yang merupakan bagian penting dari proses klasifikasi dengan algoritma *Naïve Bayes Multinomial*.

$$p(w|c) = \frac{count\ (w,c) + 1)}{(count\ c +)|V|}$$

Dalam rumus tersebut:

Count(w,c) = jumlah kemunculan kata w dalam dokumen yang termasuk kategori c

Count(c) = total seluruh kata yang muncul dalam kategori c

|V| = jumlah kata unik (*vocabulary size*) dalam seluruh dataset.

+1 = *Laplace Smoothing*, digunakan agar probabilitas tidak menjadi nol apabila kata *w* tidak pernah muncul pada kelas *c*.

|V| = jumlah total kata unik (*vocabulary size*) dalam seluruh *korpus* data Penambahan nilai +1 pada pembilang bertujuan untuk memberi bobot minimal pada kata yang tidak muncul dalam kategori tertentu, sementara penambahan |V| pada penyebut menyesuaikan total distribusi agar tetap valid sebagai probabilitas. Pendekatan ini penting untuk menjaga kestabilan model saat menghadapi data baru yang mengandung kata-kata tidak dikenal pada saat pelatihan. Berikut contoh perhitungan menggunakan model *Multinominal Naïve Bayes*:

Dokumen1="sumpah,makan,siang,gratis,progam,ampas,sekolah,coba,bikin,susah".

1. Probabilitas Kata "sumpah"

$$p(sumpah|negatif) = \frac{(0{,}60206 + 0) + 1}{(3{,}31133 + 0{,}60206) + 26}$$

$$= 0{,}05356$$

$$p(sumpah|positif) = \frac{(0 + 0) + 1}{(7{,}22472 + 2{,}70927) + 26}$$

$$= 0{,}02783$$

2. Probabilitas Kata "makan"

$$p(makan|negatif) = \frac{(0 + 0) + 1}{(3{,}31133 + 0{,}60206) + 26}$$

$$= 0{,}03343$$

$$p(makan|positif) = \frac{(0+0)+1}{(7{,}22472+2{,}70927)+26}$$

$$= 0{,}02783$$

3. Probabilitas Kata "siang"

$$p(siang|negatif) = \frac{(0+0)+1}{(3{,}31133+0{,}60206)+26}$$

$$= 0{,}03343$$

$$p(siang|positif) = \frac{(1+1)+1}{(7{,}22472+2{,}70927)+26}$$

$$= 0{,}02783$$

4. Probabilitas Kata "gratis"

$$p(gratis|negatif) = \frac{(0+0)+1}{(3{,}31133+0{,}60206)+26}$$

$$= 0{,}03343$$

$$p(gratis|positif) = \frac{(0+0)+1}{(7{,}22472+2{,}70927)+26}$$

$$= 0{,}02783$$

5. Probabilitas Kata "program"

$$p(program|negatif)$$

$$= \frac{(0+0)+1}{(3{,}31133+0{,}60206)+26}$$

$$= 0{,}03343$$

$$p(program|positif)$$

$$= \frac{(0+0)+1}{(7{,}22472+2{,}70927)+26}$$

$$= 0{,}02783$$

6. Probabilitas Kata “ampas”

$$p(ampas|negatif) = \frac{(0{,}60206 + 0) + 1}{(3{,}31133 + 0{,}60206) + 26}$$

$$= 0{,}05356$$

$$p(ampas|positif) = \frac{(0 + 0) + 1}{(7{,}22472 + 2{,}70927) + 26}$$

$$= 0{,}02783$$

7. Probabilitas Kata “sekolah”

$$p(sekolah|negatif) = \frac{(0{,}30103 + 0) + 1}{(3{,}31133 + 0{,}60206) + 26}$$

$$= 0{,}04349$$

$$p(sekolah|positif) = \frac{(0{,}30103 + 0) + 1}{(7{,}22472 + 2{,}70927) + 26}$$

$$= 0{,}03621$$

8. Probabilitas Kata “coba”

$$p(coba|negatif) = \frac{(0{,}60206 + 0) + 1}{(3{,}31133 + 0{,}60206) + 26}$$

$$= 0{,}05356$$

$$p(coba|positif) = \frac{(0 + 0) + 1}{(7{,}22472 + 2{,}70927) + 26}$$

$$= 0{,}02783$$

9. Probabilitas Kata “bikin”

$$p(bikin|negatif) = \frac{(0{,}60206 + 0) + 1}{(3{,}31133 + 0{,}60206) + 26}$$

$$= 0{,}05356$$

$$p(bikin|positif) = \frac{(0+0)+1}{(7{,}22472+2{,}70927)+26}$$

$$= 0{,}02783$$

10. Probabilitas Kata "susah"

$$p(\text{susah}|negatif) = \frac{(0+0{,}60206)+1}{(3{,}31133+0{,}60206)+26}$$

$$= 0{,}05356$$

$$p(\text{susah}|positif) = \frac{(0+0)+1}{(7{,}22472+2{,}70927)+26}$$

$$= 0{,}02783$$

Setelah nilai probabilitas untuk setiap *term* berhasil dihitung, langkah selanjutnya adalah melakukan perhitungan pada data uji. Tahap ini memanfaatkan nilai *conditional probability* yang telah diperoleh sebelumnya untuk mengklasifikasikan dokumen uji berdasarkan kategori sentimen. Berikut merupakan hasil perhitungan dari probabilitas bersyarat tersebut:

Tabel 3. 18 Hasil *Conditional Probability*

Term	P(W|negatif)	P(W|positif)
sumpah	0,05355662	0,02782881
makan	0,03342985	0,02782881
siang	0,03342985	0,02782881
gratis	0,03342985	0,02782881
program	0,03342985	0,02782881
ampas	0,05355662	0,02782881
sekolah	0,04349323	0,03620611
coba	0,05355662	0,02782881

bikin	0,05355662	0,02782881
susah	0,05355662	0,02782881
korban	0,05355662	0,02782881
sadar	0,03342985	0,04458342
manfaat	0,03342985	0,04458342
generasi	0,03342985	0,06133803
muda	0,03342985	0,04458342
sehat	0,03342985	0,04458342
kosong	0,03342985	0,04458342
perut	0,03342985	0,04458342
cerah	0,03342985	0,04458342
keren	0,03342985	0,04458342
asik	0,03342985	0,04458342
mantab	0,03342985	0,04458342
hasil	0,03342985	0,04458342
siswa	0,03342985	0,04458342
siswi	0,03342985	0,04458342
tumbuh	0,03342985	0,04458342

3.3.3. Proses Klasifikasi Data *Testing*

Setelah tahap perhitungan nilai *conditional probability* untuk setiap kata dalam masing-masing kelas sentimen, langkah berikutnya adalah menghitung nilai probabilitas *posterior*. Perhitungan nilai *posterior* ini digunakan sebagai dasar proses klasifikasi pada data *testing*, yaitu untuk menentukan label sentimen berdasarkan nilai probabilitas tertinggi.

A. Menghitung Nilai Probabilitas *Posterior*

Pada tahap ini, dilakukan proses perhitungan untuk menentukan nilai posterior pada masing-masing kategori sentimen. Nilai *posterior* ini digunakan untuk memutuskan kelas dari suatu dokumen, di mana kelas dengan nilai posterior tertinggi akan ditetapkan sebagai label dari data tersebut.Perhitungannya menggunakan rumus berikut:

$$P(d|c) = \prod_{\{w \in d\}} P(w|c)$$

Dalam rumus tersebut:

$P(c \mid d)$ merupakan probabilitas *posterior*, yaitu peluang bahwa dokumen d termasuk ke dalam kelas c.

$P(c)$ adalah probabilitas awal (*prior*) dari kategori c.

$\prod w \in dP(w \mid c)$menunjukkan perkalian dari semua probabilitas kondisional untuk setiap kata w yang terdapat dalam dokumen d.

Berikut ini disajikan contoh perhitungan klasifikasi terhadap empat dokumen uji berdasarkan rumus tersebut.

"sumpah,makan,siang,gratis,progam,ampas,sekolah,coba,bikin,susah"

Negatif:

P(negatif|d1) = P(negatif) * P(sumpah|negatif) * P(makan|negatif) * P(siang|negatif) * P(gratis|negatif) * P(program|negatif) * P(ampas|negatif) * P(sekolah|negatif) * P(coba|negatif) * P(bikin|negatif) * P(susah|negatif)

P(negatif|d1) = 0,5 * 0,05356* 0,03343* 0,03343* 0,03343* 0,03343* 0,05356* 0,04349* 0,05356* 0,05356* 0,05356

P(negatif|d1) = 1,19673E-14

Positif :

P(positif|d1) = P(positif) * P(sumpah| positif) * P(makan| positif) * P(siang| positif) * P(gratis| positif) * P(program| positif) * P(ampas| positif) * P(sekolah| positif) * P(coba| positif) * P(bikin| positif) * P(susah| positif)

P(positif|d1) = 0,5 * 0,02783* 0,02783* 0,02783* 0,02783* 0,02783*0,02783* 0,03621* 0,02783* 0,02783* 0,02783

P(positif|d1) = 1,81219E-16

Berikut ini adalah hasil perhitungan klasifikasi data testing:

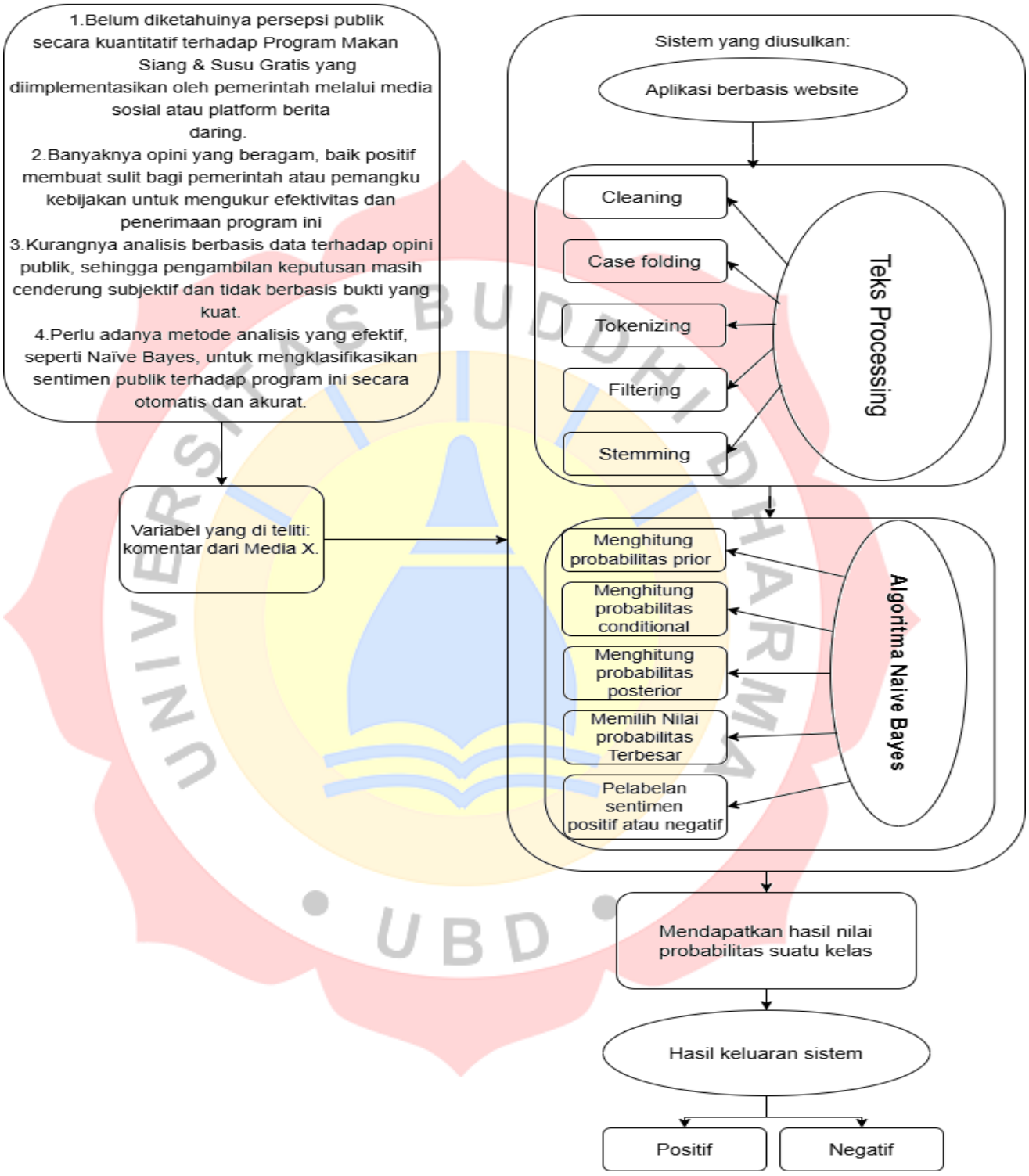
Tabel 3. 19 Hasil Klasifikasi Data Testing

doc1			
Teks	Peluang Kelas	Skor Peluang	Kelas Baru
sumpah,makan,siang,gratis,progam,ampas,sekolah,coba,bikin,susah	P(negatif)	**1,19673E-14**	**Negatif**
	P(positif)	1,81219E-16	
doc2			
Teks	Peluang Kelas	Skor Peluang	Kelas Baru
program,makan,siang,gratis,makan,korban	P(negatif)	**1,11803E-09**	**Negatif**
	P(positif)	2,32240E-10	
doc3			
Teks	Peluang Kelas	Skor Peluang	Kelas Baru
sadar,manfaat,program,makan,siang,gratis,generasi,muda,sehat,cerah,keren,asik,mantab,kosong,perut,sekolah,generasi,	P(negatif)	5,29105E-26	**Positif**
	P(positif)	**1,26743E-24**	
doc4			
Teks	Peluang Kelas	Skor Peluang	Kelas Baru
	P(negatif)	2,60723E-14	**Positif**

program makan siang gratis hasil siswa siswi tumbuh sehat	P(positif)	**5,28220E-14**	

Berdasarkan hasil perhitungan nilai probabilitas *posterior*, dokumen akan diklasifikasikan ke dalam kategori dengan nilai peluang tertinggi. Sebagai contoh, pada dokumen 1, jika nilai peluang untuk kategori negatif lebih besar dibandingkan dengan kategori positif, maka dokumen tersebut akan dimasukkan ke dalam kelas negatif, sesuai dengan nilai *posterior* tertinggi yang diperoleh. Setelah proses klasifikasi dilakukan terhadap data uji menggunakan algoritma *Naive Bayes*, langkah selanjutnya adalah melakukan evaluasi terhadap performa model klasifikasi. Evaluasi ini menggunakan metrik pengukuran seperti akurasi, *precision, recall, dan F1-score*, yang dihitung berdasarkan *confusion matrix* dari hasil klasifikasi.

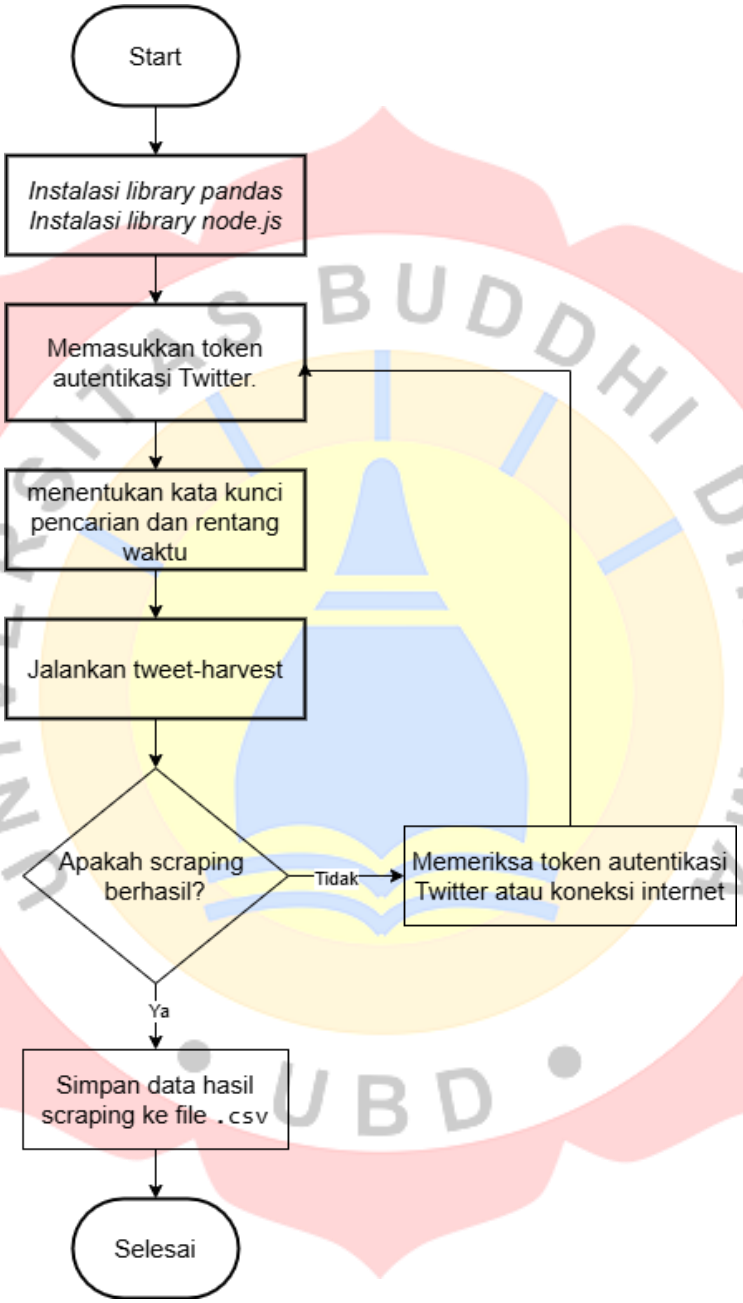
3.4. Kerangka Pemikiran



Gambar 3. 3 Kerangka Pemikiran

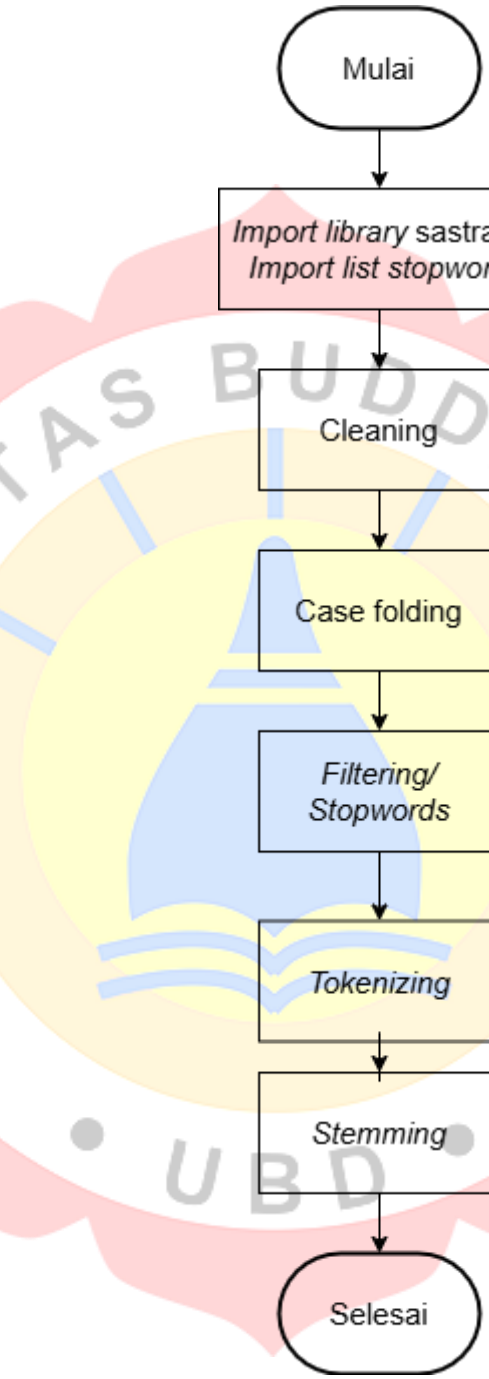
3.5. Perancangan *Flowchart*

3.5.1. *Flowchart* Pengumpulan Data



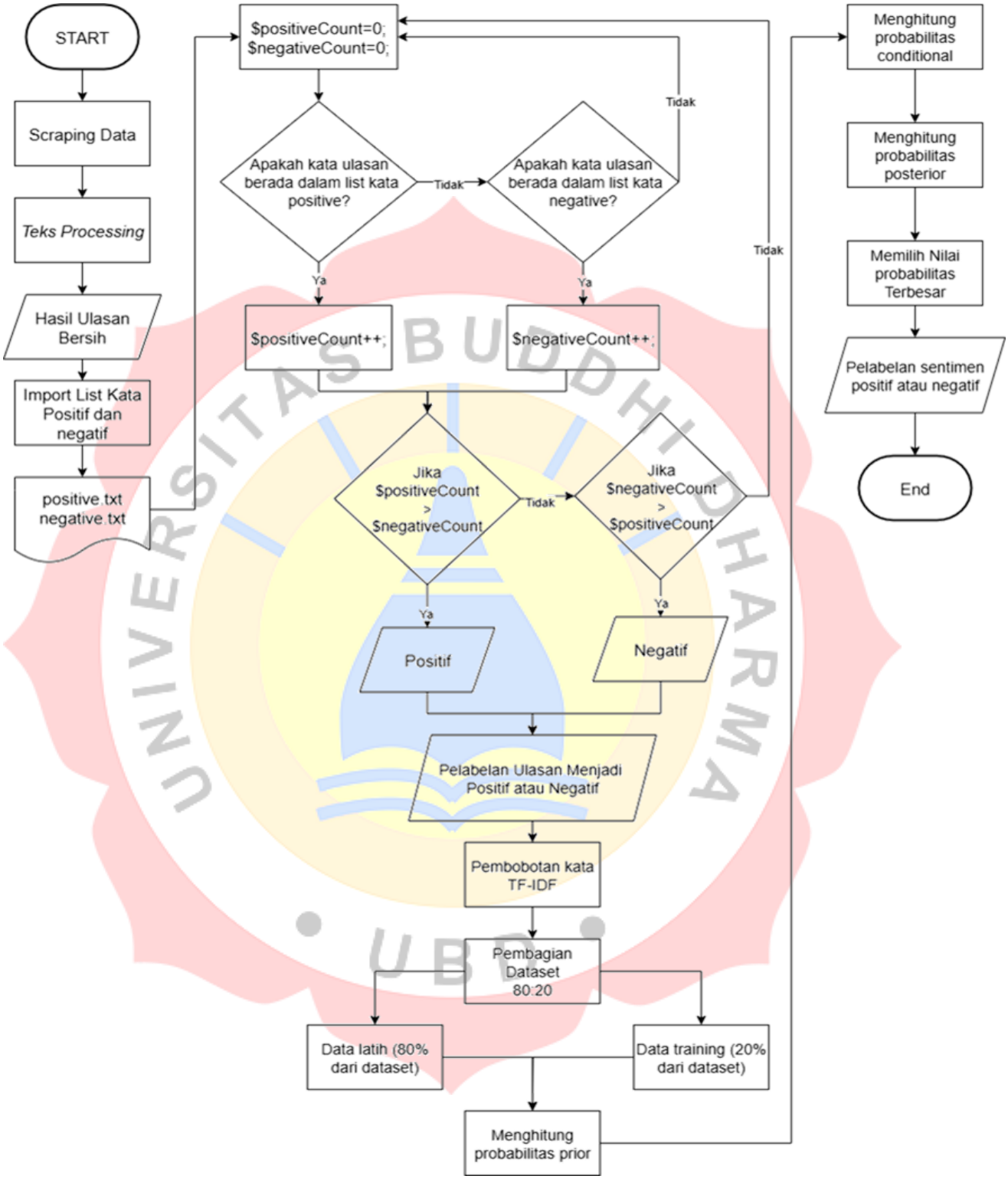
Gambar 3. 4 *Flowchart Scraping Data*

3.5.2. *Flowchart Text Preprocessing*



Gambar 3. 5 ***Flowchart Text Preprocessing***

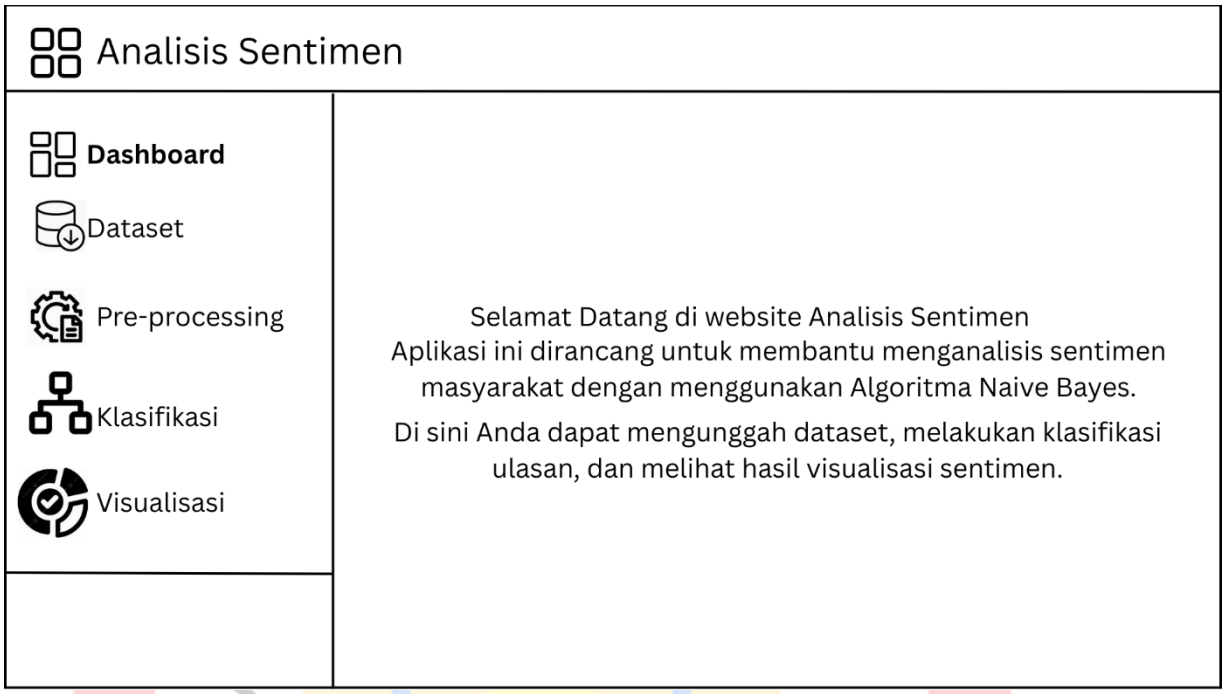
3.5.3. *Flowchart* Klasifikasi *Naïve Bayes*



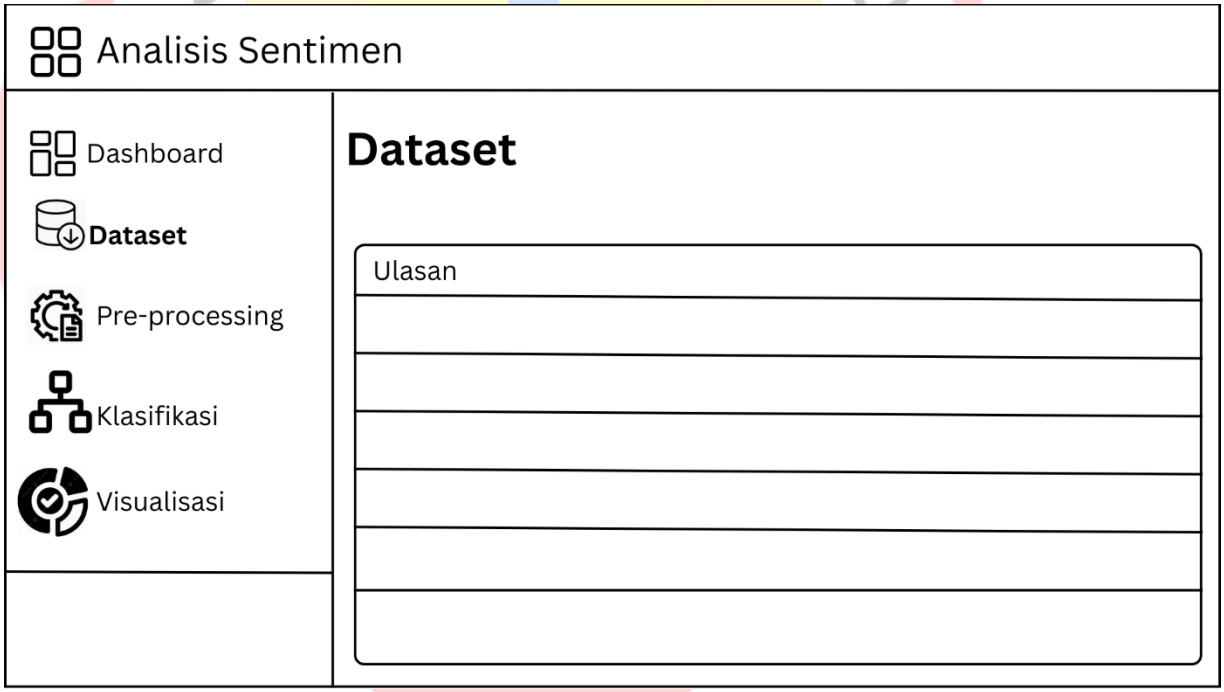
Gambar 3. 6 *Flowchart* Sentimen analisis

3.6. Perancangan Tampilan Layar, Menu

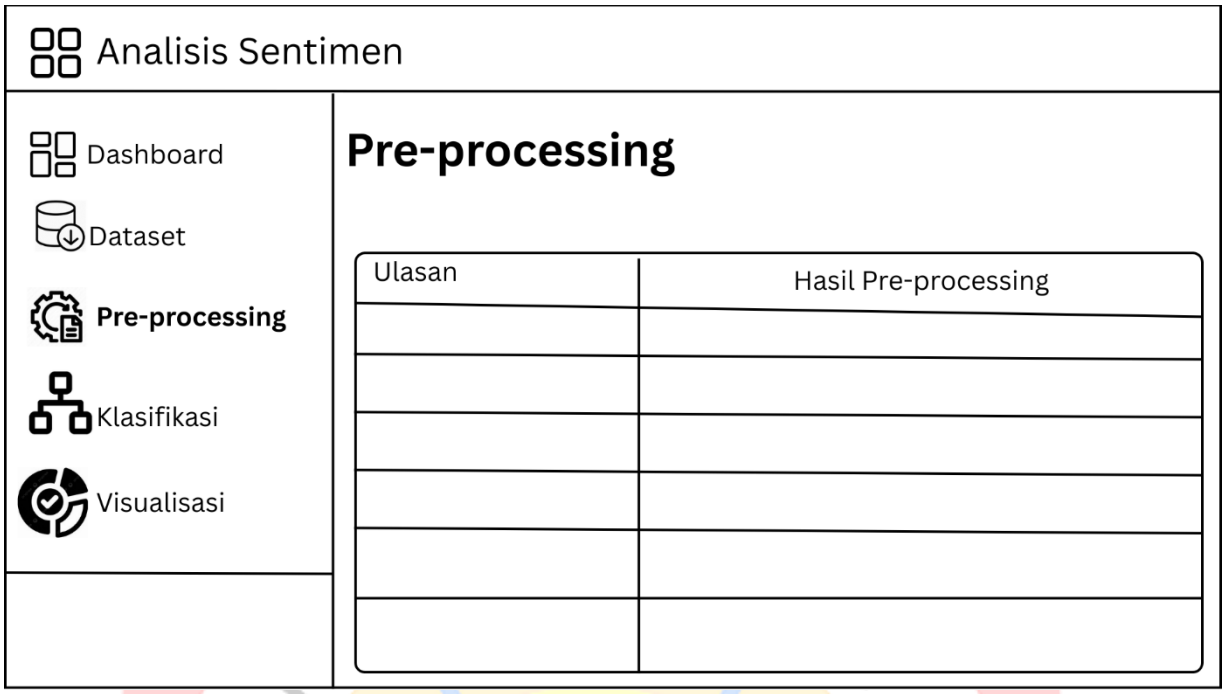
Aplikasi ini dirancang dengan lima halaman utama. Halaman pertama berfungsi sebagai *dashboard*, yang menyajikan ucapan selamat datang serta deskripsi singkat mengenai tujuan dan fungsi aplikasi. Halaman kedua merupakan menu unggah data, di mana pengguna dapat mengunggah berkas ulasan yang akan diproses lebih lanjut. Selanjutnya, halaman ketiga menyediakan fitur prapemrosesan teks (*text preprocessing*), yang mencakup tahapan pembersihan data (*cleaning*), normalisasi huruf (*case folding*), tokenisasi (*tokenizing*), penyaringan kata (*filtering*), dan *stemming*. Setelah tahap prapemrosesan selesai, pengguna dapat ke halaman keempat, yaitu proses klasifikasi sentimen, yang dilakukan dengan menerapkan algoritma *Naïve Bayes* berbasis pendekatan *lexicon-based* untuk mengklasifikasikan ulasan menjadi dua kategori sentimen: positif dan negatif. Hasil dari proses klasifikasi tersebut kemudian divisualisasikan dalam bentuk grafik yang ditampilkan pada halaman kelima. Adapun rancangan antarmuka dari aplikasi yang dikembangkan dapat dilihat pada gambar.



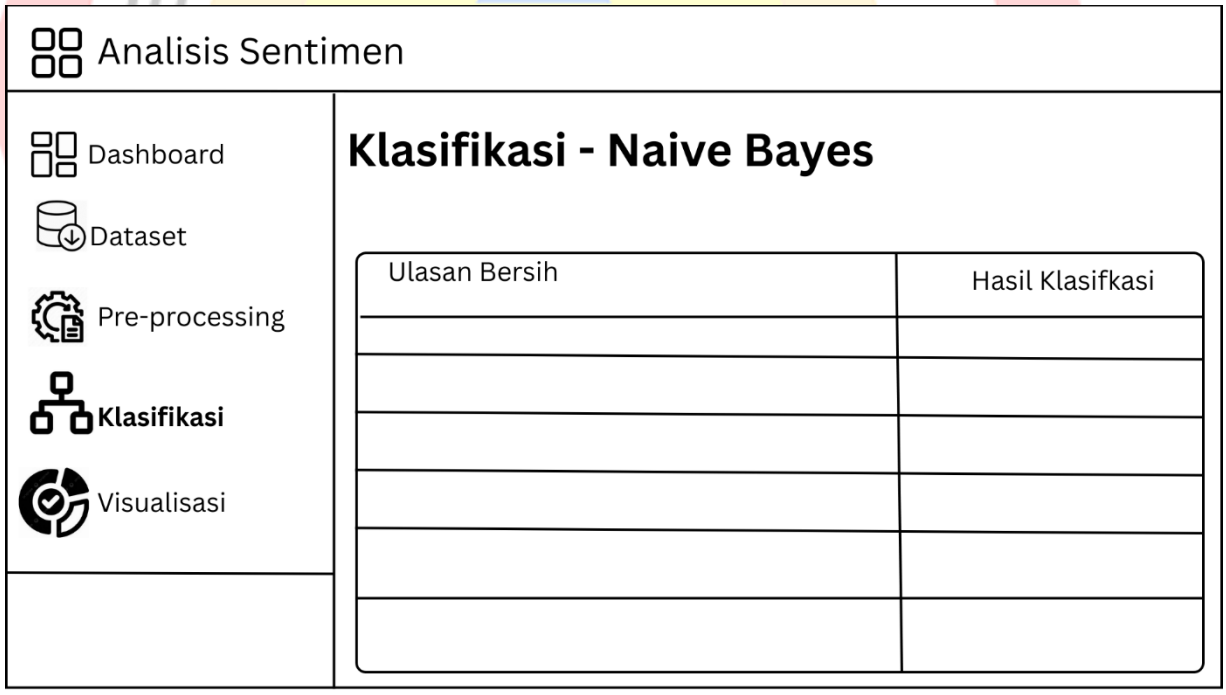
Gambar 3. 7 Rancangan Tampilan *Dashboard*



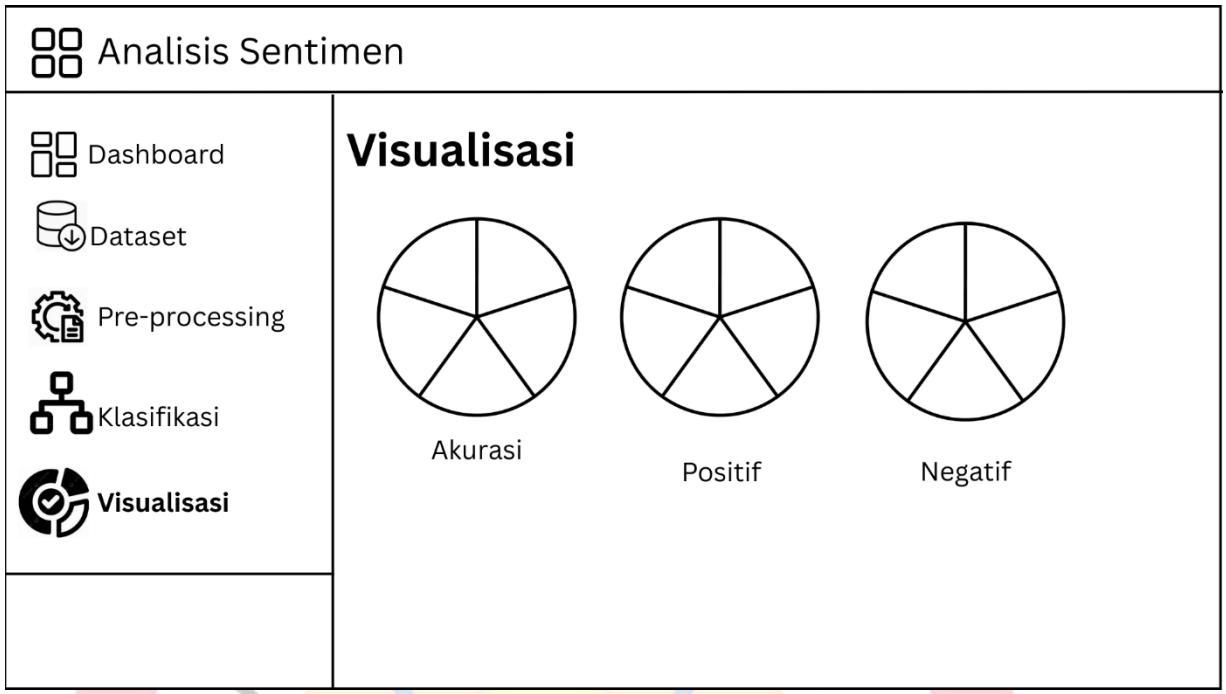
Gambar 3. 8 Rancangan Tampilan *Dataset*



Gambar 3. 9 Rancangan Tampilan *Pre-processing*



Gambar 3. 10 Rancangan Tampilan Klasifikasi



Gambar 3. 11 Rancangan Tampilan Visualisasi