

**ANALISIS SENTIMEN MENGENAI KEBIJAKAN MAKAN BERGIZI GRATIS
MENGUNAKAN METODE *SUPPORT VECTOR MACHINE***

SKRIPSI

Oleh :

AHMAD YASIR MU'AFI

NIM. 200605110057



**PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2025**

**ANALISIS SENTIMEN MENGENAI KEBIJAKAN MAKAN BERGIZI
GRATIS MENGGUNAKAN METODE *SUPPORT VECTOR MACHINE***

SKRIPSI

Diajukan kepada:

Universitas Islam Negeri Maulana Malik Ibrahim Malang
Untuk memenuhi Salah Satu Persyaratan dalam
Memperoleh Gelar Sarjana Komputer (S.Kom)

Oleh :
AHMAD YASIR MU'AFI
NIM. 200605110057

**PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2025**

HALAMAN PERSETUJUAN

**ANALISIS SENTIMEN MENGENAI KEBIJAKAN MAKAN BERGIZI
GRATIS MENGGUNAKAN METODE SUPPORT VECTOR MACHINE**

SKRIPSI

Oleh :

AHMAD YASIR MU'AFI

NIM. 200605110057

Telah Diperiksa dan Disetujui untuk Diuji:

Tanggal: 27 Mei 2025

Pembimbing I,



Fajar Rohman Hariri, M.Kom
NIP. 19890515 201801 1 001

Pembimbing II,



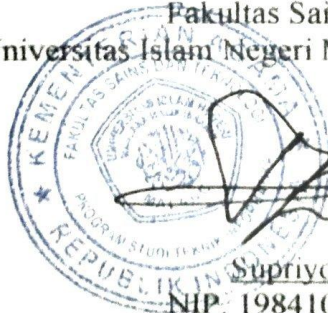
Dr. Cahyo Crysdian, M.Cs
NIP. 19740424 200901 1 008

Mengetahui,

Ketua Program Studi Teknik Informatika

Fakultas Sains dan Teknologi

Universitas Islam Negeri Maulana Malik Ibrahim Malang



Supriyono, M. Kom
NIP. 19841010 201903 1 012

HALAMAN PENGESAHAN

ANALISIS SENTIMEN MENGENAI KEBIJAKAN MAKAN SIANG GRATIS MENGGUNAKAN METODE SUPPORT VECTOR MACHINE

SKRIPSI

Oleh :

AHMAD YASIR MU'AFI

NIM. 200605110057

Telah Dipertahankan di Depan Dewan Penguji Skripsi
dan Dinyatakan Diterima Sebagai Salah Satu Persyaratan
Untuk Memperoleh Gelar Sarjana Komputer (S.Kom)
Tanggal: 25 Juni 2025

Susunan Dewan Penguji

Ketua Penguji : Okta Qomaruddin Aziz, M.Kom
NIP. 199110192019031013

()

Anggota Penguji I : Nur Fitriyah Ayu Tunjung Sari, M.Cs
NIP. 19911226 20201 2 2001

()

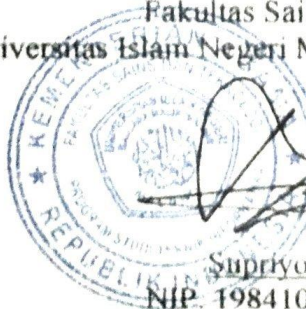

Anggota Penguji II : Fajar Rohman Hariri, M.Kom
NIP. 19890515 201801 1 001

()

Anggota Penguji III : Dr. Cahyo Crysdian, M.Cs
NIP. 19740424 200901 1 008

()

Mengetahui dan Mengesahkan,
Ketua Program Studi Teknik Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang



Supriyono, M. Kom
NIP. 19841010 201903 1 012

PERNYATAAN KEASLIAN TULISAN

Saya yang bertanda tangan di bawah ini:

Nama : Ahmad Yasir Mu'Afi
NIM : 200605110057
Fakultas / Prodi : Sains dan Teknologi / Teknik Informatika
Judul Skripsi : Analisis Sentimen Mengenai Kebijakan Makan Bergizi
Gratis Menggunakan Metode Vector Machine.

Menyatakan dengan sebenarnya bahwa Skripsi yang saya tulis ini benar-benar merupakan hasil karya saya sendiri, bukan merupakan pengambil alihan data, tulisan, atau pikiran orang lain yang saya akui sebagai hasil tulisan atau pikiran saya sendiri, kecuali dengan mencantumkan sumber cuplikan pada daftar pustaka.

Apabila dikemudian hari terbukti atau dapat dibuktikan skripsi ini merupakan hasil jiplakan, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Malang, 25 Juni 2025
Yang membuat pernyataan,



Ahmad Yasir Mu'Afi
NIM.200605110057

MOTTO

"Selama kita masih punya tekad, tidak ada yang benar-benar mustahil."

HALAMAN PERSEMBAHAN

Alhamdulillah, Puji syukur kehadiran Allah Subhanahu Wa Ta'ala, sholawat serta salam bagi Rasulullah Muhammad Shalallahu Alaihi Wassallam. Penulis mempersembahkan karya ini kepada :

Keluarga penulis yang senantiasa memberi dukungan baik dalam bentuk semangat, materi, do'a, serta rasa kasih sayang dan pengertian kepada penulis yang tiada hentinya, sehingga penulis tetap kuat dalam menjalani kehidupan hingga titik ini

Fajar Rohman Hariri, M.Kom selaku dosen pembimbing I yang telah membimbing dan memberikan jalan keluar dari keluh kesah dari penulis selama menjalani bimbingan skripsi. Dr.Cahyo Crysdian, M.Cs selaku dosen pembimbing II yang telah mempermudah serta memberikan dukungan dalam menjalani tahap skripsi.

Okta Qomaruddin Aziz, M.Kom selaku ketua penguji dan Nur Fitriyah Ayu Tunjung Sari, M.Cs selaku dosen penguji I yang telah memberi masukan dan memberikan bimbingan sehingga penulis memperoleh hasil skripsi yang lebih baik.

Seluruh Dosen Program Studi Teknik Informatika yang telah memberikan ilmunya selama masa perkuliahan sehingga peneliti mendapatkan pengetahuan yang lebih luas dan Insya Allah menjadikan manfaat bagi orang sekitar.

Pihak-pihak yang secara langsung maupun tidak langsung membantu dalam proses penyelesaian karya ini.

KATA PENGANTAR

Alhamdulillah robbil 'alamin, Segala puji bagi Allah Subhanahu wa Ta'ala yang telah memberikan kesehatan , dan rahmat sehingga penulis dapat menyelesaikan skripsi ini dengan baik. Shalawat dan salam tercurahkan kepada Nabi Muhammad Shallallahu 'alaihi wa sallam yang semoga kita mendapatkan syafaat beliau di hari akhirat.

Penyelesaian skripsi ini berhasil dilakukan atas dukungan beberapa pihak yang telah membantu secara langsung dan tidak langsung. Oleh karena itu, penulis ingin mengucapkan terima kasih kepada:

1. Prof. Dr. Hj. Ilfi Nur Diana, M.Si., CAHRM., CRMP, selaku Rektor Universitas Islam Negeri Maulana Malik Ibrahim Malang beserta jajarannya.
2. Dr. Agus Mulyono, M.Kes selaku Dekan Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang beserta jajarannya.
3. Supriyono, M. Kom, selaku ketua program studi Teknik Informatika Universitas Islam Negeri Maulana Malik Ibrahim Malang
4. Fajar Rohman Hariri, M.Kom selaku dosen pembimbing I yang telah membimbing penulis selama penyusunan tugas akhir ini dari awal sampai akhir.
5. Dr. Cahyo Crysdian, M.Cs selaku dosen pembimbing II yang telah membimbing penulis selama penyusunan tugas akhir.

6. Okta Qomaruddin Aziz, M.Kom selaku ketua penguji dan Nur Fitriyah Ayu Tunjung Sari, M.Cs sebagai penguji 1 yang telah memberikan saran dan arahan dalam penyelesaian tugas akhir ini.
7. Semua pihak yang terlibat, baik secara langsung maupun tidak langsung yang tidak dapat disebutkan satu per satu.

Penulis mengakui bahwa dalam penyusunan skripsi ini masih terdapat kekurangan dan berharap agar skripsi ini dapat memberikan manfaat bagi pembaca dan penulis secara pribadi.

Malang, 19 Juni 2025

Penulis

DAFTAR ISI

HALAMAN JUDUL	ii
HALAMAN PERSETUJUAN	ii
HALAMAN PENGESAHAN	iii
PERNYATAAN KEASLIAN TULISAN	ii
HALAMAN PENGAJUAN	v
HALAMAN PERSEMBAHAN	vii
MOTTO	ivii
HALAMAN PERSEMBAHAN	viii
KATA PENGANTAR	ix
DAFTAR ISI	xi
DAFTAR GAMBAR	xiii
DAFTAR TABEL	xiv
ABSTRAK	xv
ABSTRACT	xvi
البحث مستخلص	xvi
BAB I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Peryataan Masalah	4
1.3 Tujuan Penelitian	4
1.4 Batasan Masalah	5
1.5 Manfaat Penelitian	5
BAB II STUDI PUSTAKA	6
2.1 Analisis Sentimen	6
2.2 <i>Support Vector Machine</i>	7
BAB III DESAIN DAN IMPLEMENTASI	10
3.1 Desain Penelitian	10
3.2 Pengumpulan Data	11
3.3 Pelabelan Data	12
3.4 Desain Sistem	14
3.5 <i>Preprocessing</i> Data	14
3.6 <i>Random Undersampling</i>	19
3.7 Frekuensi Kemunculan Kata	20
3.8 TF-IDF	20
3.9 <i>Support Vector Machine</i>	29
3.10 <i>Confusion Matrix</i>	34
BAB IV HASIL DAN PEMBAHASAN	36
4.1 Data Penelitian	36
4.2 <i>Preprocessing</i> Data	37
4.3 <i>Random Undersampling</i>	41
4.4 TF-IDF	43
4.5 Perhitungan Algoritma <i>Support Vector Machine</i>	41
4.6 Skenario Pengujian	46
4.7 Hasil Uji Coba	46

4.8 Pembahasan	53
BAB V KESIMPULAN DAN SARAN	61
5.1 Kesimpulan	61
5.2 Saran	62
DAFTAR PUSTAKA	
LAMPIRAN-LAMPIRAN	

DAFTAR GAMBAR

Gambar 3.1 Desain Penelitian	10
Gambar 3.2 Desain Sistem	14
Gambar 3.3 Flowchart Proses <i>Cleaning</i>	15
Gambar 3.4 Flowchart Proses <i>Case Folding</i>	16
Gambar 3.5 Flowchart Proses <i>Remove Stopwords</i>	17
Gambar 3.6 Flowchart Proses <i>Tokenizing</i>	18
Gambar 3.7 Flowchart Proses <i>Stemming</i>	19
Gambar 3.8 Flowchart Algoritma <i>Support Vector Machine</i>	30
Gambar 3.9 Flowchart pencarian <i>hyperplane</i>	31
Gambar 4.1 Kode proses <i>Cleaning</i>	37
Gambar 4.2 Kode proses <i>Case Folding</i>	38
Gambar 4.3 Kode proses <i>Remove Stopwords</i>	39
Gambar 4.4 Kode proses <i>Tokenizing</i>	40
Gambar 4.5 Kode proses <i>Stemming</i>	41
Gambar 4.6 Kode proses <i>Random Undersampling</i>	42
Gambar 4.7 Kode Proses TF-IDF	43
Gambar 4.8 Kode proses SVM	44
Gambar 4.9 Perbandingan Akurasi dengan Jumlah Dataset dan Kernel SVM	55
Gambar 4.10 Perbandingan Akurasi dengan Rasio Data	57

DAFTAR TABEL

Tabel 2.1 Perbandingan Penggunaan metode <i>Support Vector Machine</i>	8
Tabel 3.1 Pengumpulan Data	11
Tabel 3.2 Data Berlabel	13
Tabel 3.3 Frekuensi Kemunculan Kata Dataset 2024	20
Tabel 3.4 Frekuensi Kemunculan Kata Dataset 2025	23
Tabel 3.5 Frekuensi Kemunculan Kata Dataset Gabungan	25
Tabel 3.6 Hasil Perhitungan TF	28
Tabel 3.7 <i>Confusion Matrix</i>	34
Tabel 4.1 Perbandingan Jumlah Data	36
Tabel 4.2 Skenario Pengujian	46
Tabel 4.3 Hasil proses <i>Cleaning</i>	38
Tabel 4.4 Hasil Proses <i>Case Folding</i>	38
Tabel 4.5 Hasil Proses <i>Remove Sropwords</i>	39
Tabel 4.6 Hasil Proses <i>Tokenizing</i>	40
Tabel 4.7 Hasil Proses <i>Stemming</i>	41
Tabel 4.8 Hasil <i>Random Undersampling</i>	42
Tabel 4.9 Hasil Perhitungan TF-IDF	43
Tabel 4.10 Hasil Klasifikasi SVM	45
Tabel 4.11 Hasil <i>Confusion Matrix</i> Dataset 2024	48
Tabel 4.12 Hasil Akurasi, Presisi, <i>Recall</i> , dan <i>F-measure</i> Dataset 2024	48
Tabel 4.13 Hasil <i>Confusion Matrix</i> Dataset 2025	49
Tabel 4.14 Hasil Akurasi, Presisi, <i>Recall</i> , dan <i>F-measure</i> Dataset 2025	50
Tabel 4.15 Hasil <i>Confusion Matrix</i> Dataset Gabungan	51
Tabel 4.16 Hasil Akurasi, Presisi, <i>Recall</i> , dan <i>F-measure</i> Dataset Gabungan	52
Tabel 4.17 Hasil Akurasi, Presisi, <i>Recall</i> , dan <i>F-measure</i> Semua Dataset	53

ABSTRAK

Mu'Afi, Ahmad Yasir, 2025. **ANALISIS SENTIMEN MENGENAI KEBIJAKAN MAKAN SIANG GRATIS MENGGUNAKAN METODE *SUPPORT VECTOR MACHINE***. Skripsi. Jurusan Teknik Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing: (I) Fajar Rohman Hariri, M.Kom (II) Dr.Cahyo Crys dian, M.Cs.

Kata kunci: Analisis Sentimen, *Support Vector Machine*, TF-IDF, Makan Bergizi Gratis.

Salah satu kebijakan baru yang saat ini diterapkan pada sistem pendidikan Indonesia yang bertujuan untuk memperbaiki gizi dan meningkatkan kualitas sistem pendidikan Indonesia adalah Makan Bergizi Gratis. Konsep utama kebijakan makan bergizi gratis adalah pemberian makan siang dan susu gratis kepada sekolah, pesantren, serta pemberian gizi kepada balita dan ibu hamil. Pada penerapan kebijakan tersebut menimbulkan sentimen publik baik yang netral, bersikap mendukung bahkan tidak mendukung. Sehingga perlu dilakukan analisis sentimen untuk mengetahui perspektif publik terhadap kebijakan Makan Bergizi Gratis. Pada penelitian ini menggunakan algoritma SVM dengan pembobotan TF-IDF. Data sentimen yang digunakan pada penelitian ini diperoleh dari postingan di *Instagram* yang membahas tentang makan bergizi gratis pada postingan tahun 2024 dan 2025. Skenario pengujian akan dilakukan dengan 3 *dataset* berbeda yaitu *dataset* 2024, 2025, dan gabungan keduanya, serta menggunakan 2 rasio data latih dan data uji yang berbeda dan 2 kernel yang berbeda yaitu *Linear* dan *Polynomial*. Hasil klasifikasi pada *dataset* 2024 (536 data) adalah yang terburuk dengan akurasi rata-rata 70 %, dan yang terbaik *dataset* 2025 dengan rata-rata akurasi 78%. Sedangkan untuk perbandingan kernel hasil klasifikasi pada kernel *Polynomial* mendapatkan hasil yang terbaik dari hampir semua variasi *dataset* dan variasi rasio data latih dan uji. Sedangkan untuk perbandingan rasio data uji dan data latih hasil klasifikasi pada rasio 80 : 20 unggul pada semua skenario pengujian pada rasio 70: 30.

ABSTRACT

Mu'Afi, Ahmad Yasir, 2025. **SENTIMENT ANALYSIS ON FREE LUNCH POLICY USING *SUPPORT VECTOR MACHINE METHOD***. Thesis. Department of Informatics Engineering, Faculty of Science and Technology, State Islamic University of Maulana Malik Ibrahim Malang. Promotor: (I) Fajar Rohman Hariri, M.Kom (II) Dr.Cahyo Crysdian, M.Cs.

One of the new policies currently being implemented in the Indonesian education system that aims to improve nutrition and improve the quality of the Indonesian education system is Free Nutritious Meals. The main concept of the free nutritious meal policy is the provision of free lunch and milk to schools, Islamic boarding schools, and the provision of nutrition to toddlers and pregnant women. The implementation of this policy has raised public sentiment, both neutral, supportive and even non-supportive. Therefore, it is necessary to conduct a sentiment analysis to determine the public perspective on the Free Nutritious Meal policy. This study uses the Support Vector Machine algorithm with TF-IDF weighting. The sentiment data used in this study was obtained from Instagram posts discussing free nutritious meals in posts in 2024 and 2025. The test scenario will be carried out with 3 different datasets, namely the 2024, 2025, and a combination of both datasets, and using 2 different ratios of training and test data and 2 different kernels, namely Linear and Polynomial. The classification results on the 2024 dataset (536 data) are the worst with an average accuracy of 70%, after that the combined dataset with an average of 76%, and the best is the 2025 dataset with an average accuracy of 78%. Meanwhile, for the comparison of kernels, the classification results on the Polynomial kernel are superior to almost all datasets and variations in the ratio of training and test data. Meanwhile, for the comparison of the ratio of test data and training data, the classification results at a ratio of 80: 20 are superior in all test scenarios at a ratio of 70: 30.

Key words: Sentiment Analysis, Support Vector Machine, TF-IDF, Free Nutritious Meals.

البحث مستخلص

معافي، أحمد ياسر، 2025. تحليل المشاعر حول سياسة الغذاء المجاني باستخدام طريقة آلة ناقلات الدعم. الأطروحة. قسم هندسة المعلوماتية، كلية العلوم والتكنولوجيا، جامعة مولانا مالك إبراهيم مالانج الإسلامية الحكومية. المشرف: (ط) فجر رحمان حريري، ماجستير كوم (2) د. كاهيو كريسيديان، ماجستير.

مجانية غذائية وجبة، TF-IDF، المساندة، المتجهات دعم آلة المشاعر، تحليل المفتاحية الكلمات

إحدى السياسات الجديدة المطبقة حاليًا في نظام التعليم الإندونيسي والتي تهدف إلى تحسين التغذية وتعزيز جودة نظام التعليم الإندونيسي هي الوجبة الغذائية المجانية. ويتمثل المفهوم الرئيسي لسياسة الوجبة الغذائية المجانية في توفير وجبة الغذاء والحليب مجاناً للمدارس والمدارس الداخلية الإسلامية، فضلاً عن توفير التغذية للأطفال الصغار والنساء الحوامل. وقد تسبب تنفيذ هذه السياسة في أن تكون مشاعر الجمهور محايدة أو مؤيدة أو حتى غير مؤيدة. لذلك من الضروري إجراء تحليل المشاعر لمعرفة وجهة نظر الجمهور حول سياسة الوجبات الغذائية المجانية. ويستخدم هذا يتم الحصول على بيانات المشاعر المستخدمة في هذا TF-IDF البحث خوارزمية آلة دعم المتجهات مع ترجيح البحث من المنشورات على إنستغرام التي تناقش الوجبات الغذائية المجانية في منشورات عامي 2024 و2025. سيتم إجراء سيناريو الاختبار باستخدام 3 مجموعات بيانات مختلفة، وهي مجموعات بيانات 2024 و2025 ومجموعات البيانات المدمجة، وكذلك باستخدام نسبتين مختلفتين من بيانات التدريب وبيانات الاختبار ونواتين مختلفتين، وهما النواة الخطية ومتعددة الحدود. نتائج التصنيف على مجموعة بيانات 2024 (536 بيانات) هي الأسوأ بمتوسط دقة 70%، ثم مجموعة البيانات المدمجة بمتوسط دقة 76%، وأفضل مجموعة بيانات 2025 بمتوسط دقة 78%. أما بالنسبة لمقارنة نتائج تصنيف النواة على النواة متعددة الحدود فهي متفوقة على جميع مجموعات البيانات تقريباً والاختلافات في نسبة بيانات التدريب والاختبار. أما بالنسبة لمقارنة نسبة بيانات الاختبار د

BAB I

PENDAHULUAN

1.1 Latar Belakang

Pemilihan Umum presiden dan wakil presiden tahun 2024 di Indonesia melahirkan berbagai wacana kepada publik, salah satu wacana yang banyak dibicarakan adalah kebijakan makan siang gratis. Kebijakan makan bergizi gratis adalah pemberian makan siang dan susu gratis kepada sekolah, pesantren, serta pemberian gizi kepada balita dan ibu hamil (Zaman et al., 2024) . Berdasarkan penelitian terdahulu menyatakan bahwa nutrisi dan gizi yang baik dapat memberikan energi dan meningkatkan daya ingat sehingga memungkinkan siswa untuk berkonsentrasi dan aktif dalam pembelajaran (Eliza et al., 2024). Berdasarkan pengalaman negara Amerika Serikat kebijakan makan siang gratis di sekolah, telah memberikan pengaruh yang positif (Fanny et al., 2024) . Dengan adanya program tersebut diharapkan dapat mengatasi masalah kekurangan gizi anak di Indonesia. Pada wacana penerapan kebijakan makan siang gratis menimbulkan kontroversi, disebabkan muncul wacana rencana anggaran yang digunakan akan menggunakan dana BOS (Fasha & Tesniyadi, 2024).

Beberapa kebijakan yang telah dibuat oleh pemerintah dilakukan hanya untuk menjadikan rakyat makmur dan sejahtera. Oleh karena itu kebijakan yang telah dibuat harus ditetapkan dengan adil. Hal ini sudah dijelaskan dalam Surah An-Nisa ayat 58:

إِنَّ اللَّهَ يَأْمُرُكُمْ أَنْ تُؤَدُّوا الْأَمَانَاتِ إِلَىٰ أَهْلِهَا وَإِذَا حَكَمْتُمْ بَيْنَ النَّاسِ أَنْ تَحْكُمُوا بِالْعَدْلِ إِنَّ اللَّهَ نِعِمَّا يَعِظُكُمْ بِهِ إِنَّ اللَّهَ كَانَ سَمِيعاً بَصِيراً (58)

“Sesungguhnya Allah menyuruh kalian menyampaikan amanat kepada yang berhak menerimanya, dan (menyuruh kalian) apabila menetapkan hukum di antara manusia supaya kalian menetapkan dengan adil. Sesungguhnya Allah memberi pengajaran yang sebaik-baiknya kepada kalian. Sesungguhnya Allah adalah Maha Mendengar lagi Maha Melihat”(Q.S. AN-Nisa: 58)

Menurut tafsir ibnu katsir disebutkan pada ayat ini memerintahkan untuk menyampaikan amanat kepada yang berhak menerimanya. Amanat ini mencakup hak seorang hamba kepada tuhanya dan hak hamba kepada hamba. Selain itu, ayat tersebut juga diturunkan kepada umara atau pembuat hukum diantara manusia untuk adil. Sehingga dapat disimpulkan bahwa pemerintah sebagai pembuat kebijakan harus membuat dengan seadil-adilnya.

Pada rencana penerapan makan siang gratis menimbulkan komentar publik yang netral, bersikap pro bahkan kontra dalam kebijakan tersebut (Fasha & Tesniyadi, 2024) . Sehingga diperlukan analisis sentimen untuk mengetahui perspektif publik terhadap kebijakan makan siang gratis.

Analisis sentimen adalah sebuah ilmu pengetahuan yang merupakan cabang dari data mining yang digunakan untuk mengolah dan menganalisis data tekstual dari sebuah organisasi, individu, dan topik tertentu (Lubis et al., 2024). Data yang didapat dari analisis sentimen merupakan data yang tidak terstruktur, kemudian datanya dapat diklasifikasikan ke dalam sentmen positif, negatif atau netral (Ramadhani et al., 2024). Pengelompokan klasifikasi tersebut dapat menggunakan klasifikasi teks.

Dalam melakukan analisis sentimen data paling mudah di dapat dari sebuah *platform* media sosial. Media sosial merupakan sebuah media atau *platform* yang memungkinkan komunikasi atau interaksi sosial secara online (Lailita & Khoirunnisa, 2024). Dikutip dari *website* Sekretariat Kabinet Republik Indonesia tahun 2016, bahwa sekitar 132 juta pengguna internet aktif di Indonesia (Zena Lusi et al., 2024). Analisis sentimen pada media sosial adalah mengetahui penentuan atau respon opini, pandangan terhadap sebuah isu, produk yang dikumpulkan atau analisis datanya didapat dari media sosial (Kaharudin et al., 2023). Sosial media yang akan digunakan untuk pengambilan data adalah sosial media Instagram.

Instagram adalah sosial media yang populer yang bagi semua kalangan dari yang tua sampai yang muda (Khatib Sulaiman et al., n.d.) . Alasan pengambilan data dari Instagram dikarenakan menyediakan dataset yang kaya dan beragam yang disampaikan melalui unggahan, komentar, Instagram TV, dan hashtag, yang membuat sentimen menjadi lebih komprehensif (Ezra Rofran & Joanda Kaunang, 2024).

Algoritma yang sering digunakan dalam analisis sentimen adalah algoritma klasifikasi *Naive Bayes* dan SVM (Setiawan & Suryono, 2024). Dalam penelitian ini algoritma klasifikasi yang akan digunakan adalah algoritma SVM. SVM adalah algoritma klasifikasi yang cocok dan kuat untuk menganalisis data, serta menghasilkan hasil yang optimal (Setiawan & Suryono, 2024). SVM dikenal dengan algoritma dengan akurasi yang tinggi, yang kinerjanya lebih baik dari algoritma klasifikasi lainya (Nufairi et al., 2024). Keunggulan SVM adalah fungsi

kernel yang dapat memisahkan data *nonlinear* yang besar secara efektif (Nufairi et al., 2024). Dalam penelitian yang berjudul Analisis Sentimen Aplikasi Ruang Guru di Twitter menggunakan algoritma klasifikasi, terbukti SVM memiliki akurasi yang tinggi dibandingkan yang lain, SVM memiliki akurasi 78,55%, Naive Bayes memiliki akurasi 67,32%, dan K-Nearest Neighbour 77,21 (Giovani et al., 2020). Selain itu, alasan dipilihnya algoritma SVM adalah cocok dengan pembobotan TF-IDF dalam penerapannya (Ipmawati et al., 2024).

Berdasarkan penjelasan pada bagian latar belakang, penelitian ini dilakukan untuk melakukan analisis sentimen terhadap kebijakan makan bergizi gratis menggunakan metode SVM. Tujuan dari penelitian ini adalah Mengukur besar nilai akurasi, presisi, *recall*, dan *f-measure*, serta menganalisis faktor-faktor yang berpengaruh terhadap performa SVM.

1.2 Pernyataan Masalah

Berdasarkan uraian pada bagian latar belakang, berikut adalah rumusan masalahnya:

1. Berapakah nilai dari akurasi, *recall*, presisi, dan *f-measure* yang didapatkan dalam analisis sentimen kebijakan MBG dengan menggunakan SVM?
2. Faktor apa yang mempengaruhi performa SVM pada analisis sentimen kebijakan MBG?

1.3 Tujuan Penelitian

1. Mengukur nilai dari akurasi, *recall*, presisi, dan *f-measure* yang didapat dalam analisis sentimen kebijakan MBG dengan menggunakan SVM.

2. Menganalisis faktor-faktor yang berpengaruh terhadap performa SVM pada analisis sentimen kebijakan MBG.

1.4 Batasan Masalah

Data sentimen yang digunakan berbahasa Indonesia dan diambil pada postingan akun berita yang membahas “makan siang gratis” di Instagram berjumlah 933 data pada postingan tahun 2024 dan 641 data pada postingan tahun 2025.

1.5 Manfaat Penelitian

Manfaat yang diharapkan dalam penelitian ini adalah sebagai berikut:

1. Bagi Presiden dan Wakil Presiden terpilih diharapkan mampu menjadi bahan pertimbangan untuk menyiapkan strategi kebijakan makan siang gratis yang paling baik untuk memperbaiki gizi anak-anak sekolah.

Memberikan pengetahuan dan pemahaman mengenai algoritma *Support Vector Machine* dan pembobotan dengan TF-IDF pada penelitian analisis sentimen.

BAB II

STUDI PUSTAKA

2.1 Analisis Sentimen

Analisis sentimen adalah sebuah studi komputasi mengenai ulasan-ulasan, serta sentimen yang diekspresikan dalam sebuah teks (Nurian et al., 2024) . Analisis sentimen memiliki tugas mengelompokkan polaritas dari sebuah ulasan sehingga dapat diketahui apakah ulasanya cenderung ke positif atau ke negatif (Nurian et al., 2024) . Analisis sentimen berperan penting mengolah data dalam jumlah besar menjadi sebuah informasi penting bagi suatu individu atau organisasi tertentu.

Analisis sentimen memiliki tujuan untuk mengetahui kecenderungan sebuah sentimen apakah ke arah netral, positif, atau negatif. Analisis sentimen dapat mengacu pada berbagai bidang dari penulisan mengenai suatu layanan atau topik, serta menganalisa sentimen, ulasan, emosi pada seseorang, ataupun kegiatan tertentu (Khusnul et al., 2024) . Dalam hal ini, analisis sentimen telah sering digunakan atau menjadi populer dalam menganalisis ulasan terhadap suatu layanan (Nurian et al., 2024).

Analisis sentimen adalah suatu studi penyelidikan sentimen, ulasan, komentar, emosi terhadap suatu objek seperti individu, masalah, topik, produk, atau atribut terkait. Analisis sentimen melibatkan berbagai aspek yang memiliki fungsi berbeda, termasuk analisis sentimen, data mining, ekstraksi sentimen, analisis subjek, dan peninjauan sentimen (Karimah & Dwilestari, 2024).

Meskipun memiliki fungsi berbeda, semuanya dapat masuk dalam analisis sentimen.

2.2 Support Vector Machine

Hasil dari penelitian yang pernah dilakukan oleh Pamungkas & Cahyono (2024) dengan melakukan analisis sentimen terhadap aplikasi *ChatGP*. Penelitian ini menggunakan algoritma SVM dan KNN. Data diambil menggunakan library *google play store* di Google colab, dataset yang diambil berjumlah 4712 komentar. Labeling data dibagi menjadi tiga yaitu netral, positif, dan negatif. Selanjutnya, data hasil preprocessing dibagi menjadi 3769 data latih dan 943 data uji. Hasil analisis menggunakan algoritma SVM menunjukkan konsistensi hasil dengan rata-rata akurasi 80%, presisi 79%, *recall* 77%, dan *F-1 Score* 77% yang dilakukan dalam lima kali pengujian.

Lubis & Setyawan (2024) melakukan penelitian analisis sentimen mengenai aplikasi pospay di playstore. Algoritma yang digunakan dalam melakukan klasifikasi adalah SVM dan *Naive Bayes*. Data diambil dari web scraping dari halaman ulasan aplikasi Pospay pada Google Play Store. Setelah mengalami proses preprocessing diperoleh jumlah 2258 komentar positif, 2517 komentar negatif, dan 225 komentar netral. Kemudian data dibagi menjadi 80% data training dan 20% data testing. Hasil dari penelitian menggunakan algoritma SVM memperoleh hasil akurasi 87%, presisi 96%, *recall* 87%, dan *f-1 score* 91%.

Penelitian dilakukan oleh Riyadiiban & Riyadi (2024) melakukan analisis sentimen terhadap stadion Jakarta Internasional Stadium (Jis) dengan membandingkan algoritma SVM dan *Naive Bayes*. Pengambilan data dengan

menggunakan API Twitter, yang diambil sebanyak 940 komentar. Labeling data dibagi menjadi dua yaitu sentimen positif berjumlah 46 dan sentimen negatif berjumlah 894. Hasil dari klasifikasi menggunakan SVM mendapatkan nilai akurasi 99,68%, *recall* positif 97,73%, *recall* negatif 99,78%, presisi positif 100%, presisi negatif 97,71%.

Viriya et al (2024) melakukan analisis sentimen terhadap aplikasi mobile Gapura UB pada Google Play Store dengan algoritma SVM. Data diambil dari halaman ulasan aplikasi mobile Gapura UB di Play Store dengan web scraping. Data yang diambil berjumlah 345 ulasan negatif dan 149 ulasan positif. Dengan pembagian data 90% data latih dan 10% data uji, mendapatkan akurasi sebesar 98%, presisi 99%, *recall* 97%, dan *f-1 score* 98%.

Sabna et al (2024) melakukan analisis sentimen mengenai kurikulum merdeka menggunakan algoritma SVM dan *Naive Bayes*. Data diambil dari *platform* Twitter yang berjumlah 399 data. Data dibagi menjadi dua yaitu 319 data positif dan 80 data negatif. Hasil pengujian dengan algoritma SVM mendapatkan nilai akurasi 79,95%, presisi 79,95%, dan *recall* 100%.

Berikut adalah perbandingan dari beberapa algoritma SVM yang disajikan dalam sebuah Tabel 2.1.

Tabel 2.1 Perbandingan Penggunaan metode *Support Vector Machine*

NO	Objek	Metode	Hasil
1	aplikasi ChatGP	Support Vector Machine	Hasil analisis menggunakan algoritma SVM menunjukkan konsistensi hasil dengan rata-rata akurasi 80%, presisi 79%, <i>recall</i> 77%, dan F-1 Score 77% yang dilakukan dalam lima kali pengujian.

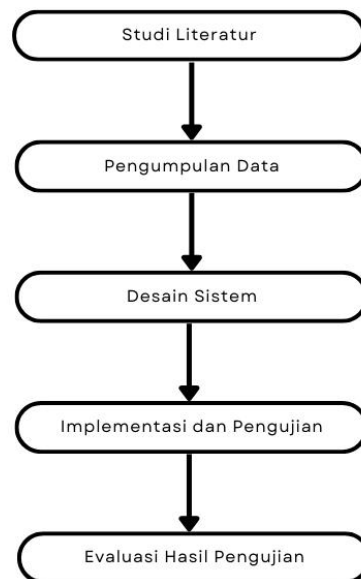
2	aplikasi pospay	Support Vector Machine	Hasil dari penelitian menggunakan algoritma SVM memperoleh hasil akurasi 87%, presisi 96%, recall 87%, dan f-1 score 91%
3	stadion Jakarta Internasional Stadium (Jis)	Support Vector Machine	Hasil dari klasifikasi menggunakan SVM mendapatkan nilai akurasi 99,68%, recall positif 97,73%, recall negatif 99,78%, presisi positif 100%, presisi negatif 97,71%.
4	aplikasi mobile Gapura UB	Support Vector Machine	Data yang diambil berjumlah 345 ulasan negatif dan 149 ulasan positif. Dengan pembagian data 90% data latih dan 10% data uji, mendapatkan akurasi sebesar 98%, presisi 99%, recall 97%, dan f-1 score 98%.
5	kurikulum merdeka	Support Vector Machine	Hasil pengujian dengan algoritma SVM mendapatkan nilai akurasi 79,95%, presisi 79,95%, dan recall 100%

BAB III

DESAIN DAN IMPLEMENTASI

3.1 Desain Penelitian

Pada bagian ini mempresentasikan langkah-langkah yang akan dilakukan dalam melakukan penelitian. Desain penelitian ditulis secara sistematis dan teratur agar mendapatkan target penelitian yang baik. Berikut adalah desain penelitian yang dapat dilihat pada Gambar 3.1.



Gambar 3.1 Desain Penelitian

Langkah diawali dengan melakukan studi literatur yang kemudian dilakukan pengumpulan data dari Instagram melalui proses *scrapping* data, yang selanjutnya dilakukan *labeling* data menjadi label netral, positif, atau negatif. Pada tahapan

desain sistem akan dibagi menjadi tiga yaitu, *preprocessing* data, pembobotan kata, dan pembagian data. Pada *preprocessing* data terdapat 5 tahapan *preprocessing* yaitu, *cleaning*, *case folding*, *tokenizing*, *remove stopwords*, dan *stemming*. Pada pembobotan kata akan dilakukan dengan TF-IDF, yang kemudian akan dilakukan pengklasifikasian data menggunakan algoritma *Support Vector Machine*. Tahapan terakhir adalah evaluasi hasil yang sudah dilakukan dalam penelitian.

3.2 Pengumpulan Data

Pengumpulan data diperoleh dari kolom komentar sosial media Instagram dalam postingan dari beberapa akun yang membahas tentang makan siang gratis. Pengumpulan data dilakukan dengan Data Miner yang adalah sebuah ekstensi pada sebuah *browser*. Data sentimen akan dibagi menjadi dua yaitu data sentimen pada postingan 2024, dan data sentimen pada postingan tahun 2025. Pengumpulan data sentimen berhasil mendapatkan 933 data untuk tahun 2024 dan 641 data untuk tahun 2025. Berikut adalah tabel hasil sampel dari pengumpulan data yang dapat dilihat pada Tabel 3.1.

Tabel 3.1 Pengumpulan Data

Komentar
“@necthophyle112 Sejujurnya banyak anak sekolah yg bahkan gak jajan karena gak bawa uang jajan, makanan gratis ini semoga bisa bermanfaat.”
“@wawan_rudalf_26_12 Semoga program berjalan lancar dan menyeluruh sampai daerah perbatasan Indonesia”
“@agung_rudes Pendidikan gratis lebih penting drpd makan gratis...”
“@tsaqip_rachman Sy lebih setuju pendidikan gratis, kalau makan insya Allah masih sanggup walaupun hanya makan singkong”

“@evakristyana Ak setuju bgt ada makan siang gratis ,soalnya pusing jg kl hrus mikir menu bawain bekal apa, palagi kl pekerja kaya ak yg hrus mencari nafkah buat bantu suami juga.”

3.3 Pelabelan Data

Pelabelan data dilakukan dengan membagi data menjadi dua label yaitu positif atau negatif. Pelabelan dilakukan dengan pelabelan otomatis dengan RoBERTa (Robustly Optimized BERT Pretraining Approach). Pelabelan otomatis dengan RoBERTa merupakan proses pemberian label sentimen secara otomatis terhadap data teks menggunakan model pra-latih berbasis *transformer*. Proses dimulai dengan memuat model dan *tokenizer*, lalu data teks (misalnya komentar pengguna) dianalisis melalui *pipeline* analisis sentimen dari *HuggingFace Transformers*.

Pada pelabelan otomatis dengan RoBERTa setiap teks akan diberikan label sentimen seperti *positive*, *neutral*, atau *negative*, yang kemudian dikonversi ke format biner, misalnya P untuk positif dan N untuk negatif. Pendekatan ini memungkinkan analisis cepat dan konsisten terhadap jumlah data besar tanpa memerlukan pelabelan manual, sehingga sangat efisien untuk aplikasi seperti pemantauan opini publik, evaluasi produk, atau penelitian media sosial.

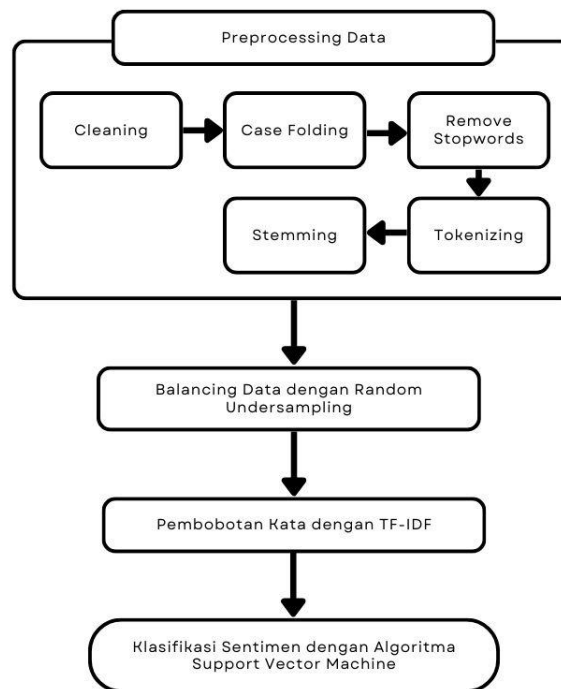
Hasil dari pelabelan otomatis dengan RoBERTa mendapatkan 268 data berlabel positif dan 665 data berlabel negatif pada dataset 2024, serta 325 data berlabel positif dan 316 berlabel negatif pada dataset 2025. Berikut adalah data yang berhasil dikumpulkan yang dapat dilihat pada Tabel 3.2.

Tabel 3.2 Data Komentar dan Label

Komentar	Label
“@necthophyle112 Sejujurnya banyak anak sekolah yg bahkan gak jajan karena gak bawa uang jajan, makanan gratis ini semoga bisa bermanfaat.”	P
“@wawan_rudalf_26_12 Semoga program berjalan lancar dan menyeluruh sampai daerah perbatasan Indonesia”	P
“@agung_rudes Pendidikan gratis lebih penting drpd makan gratis...”	N
“@tsaqip_rachman Sy lebih setuju pendidikan gratis, kalau makan insya Allah masih sanggup walaupun hanya makan singkong”	N
“@evakristyana Ak setuju bgt ada makan siang gratis ,soalnya pusing jg kl hrus mikir menu bawain bekal apa, palagi kl pekerja kaya ak yg hrus mencari nafkah buat bantu suami juga.”	P

3.4 Desain Sistem

Berikut adalah beberapa langkah desain sistem pada Gambar 3.2.



Gambar 3.2 Desain Sistem

Proses diawali dengan *preprocessing* data yang mana dibagi menjadi 5 tahapan yaitu, *cleaning*, *case folding*, *tokenizing*, *remove stopwords*, dan *stemming*. Data yang sudah melalui proses *preprocessing* akan menghasilkan data bersih. Kemudian akan dilakukan *Random Undersampling* data untuk menyeimbangkan data. Kemudian data bersih tersebut dilakukan pembobotan kata dengan TF-IDF. Hasil dari pembobotan kata akan menghasilkan data bobot kata yang nantinya akan digunakan untuk proses selanjutnya.

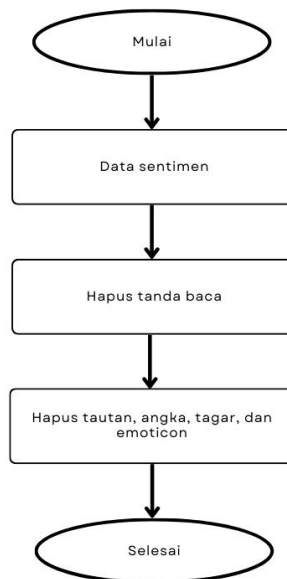
3.5 Preprocessing Data

Data hasil proses data mining perlu melalui tahap *preprocessing* terlebih dahulu. Tahap ini bertujuan untuk mengolah data mentah agar menjadi data yang

lebih terstruktur dan bersih, sehingga mempermudah proses klasifikasi. Pada penelitian ini digunakan lima langkah utama dalam *preprocessing*, yaitu *cleaning*, *case folding*, *tokenizing*, *stopword removal*, serta *stemming*.

3.5.1 *Cleaning*

Tahapan *cleaning* adalah tahapan untuk menghilangkan *noise* yang tidak ada kaitanya seperti tagar, angka, tanda baca, *emoticon*, dan tautan. *Cleaning* berfungsi untuk menghilangkan komponen yang tidak penting sehingga membuat data menjadi bersih dan memudahkan untuk proses klasifikasi. Berikut adalah contoh *flowchart* dari proses *cleaning* yang dapat dilihat pada Gambar 3.3.

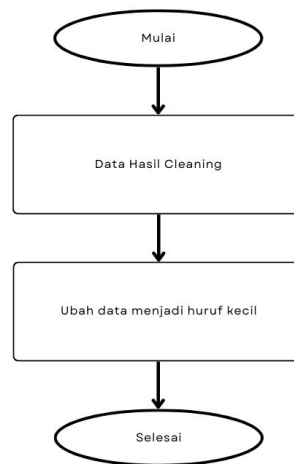


Gambar 3.3 *Flowchart* Proses *Cleaning*

Langkah diawali dengan menginput data sentimen hasil pengumpulan data yang telah dilakukan. Kemudian akan dilakukan penghapusan tanda baca, tautan, angka, dan emoticon.

3.5.2 Case Folding

Case folding merupakan tahap mengubah seluruh huruf kapital menjadi huruf kecil. Tujuan dari proses ini adalah menyeragamkan bentuk huruf agar konsisten, yakni hanya dalam bentuk huruf kecil. Dengan adanya *case folding*, sistem dapat lebih mudah mengenali kata dan menghasilkan perhitungan yang lebih optimal. Ilustrasi alur tahapan *case folding* dapat dilihat pada Gambar 3.4..



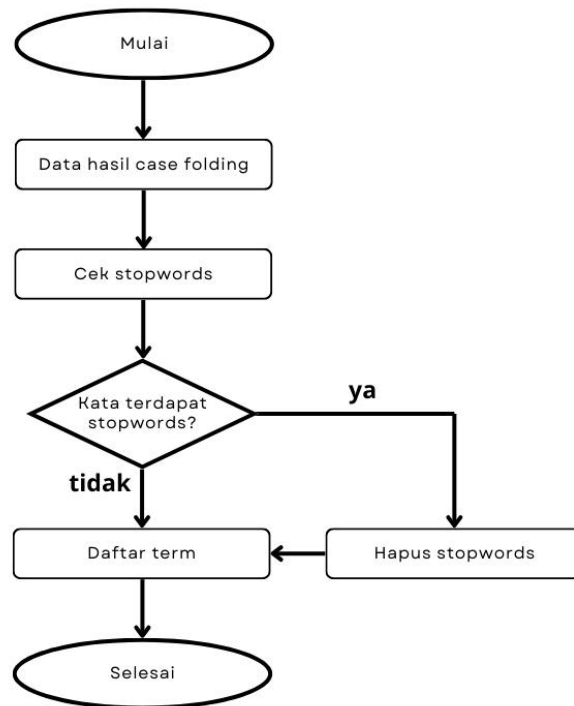
Gambar 3.4 Flowchart Proses Case Folding

Proses pada gambar diatas merupakan tahap *case folding*, di mana input berasal dari teks yang telah melewati pembersihan sebelumnya. Teks yang telah dibersihkan kemudian dikonversi ke huruf kecil seluruhnya untuk menyamakan format penulisan.

3.5.3 Remove Stopwords

Remove stopwords adalah tahapan menghilangkan kata-kata umum, kata-kata yang tidak penting, dan kata yang tidak mempengaruhi hasil dari proses

analisis sentimen. Tahapan ini bertujuan untuk mengurangi jumlah kata yang ada dalam data. Berikut adalah *flowchart* dari tahapan *remove stopwords* pada Gambar 3.5.



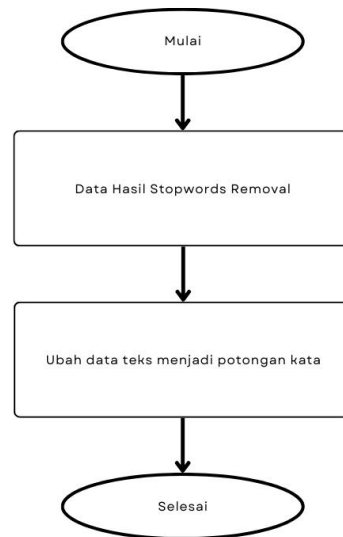
Gambar 3.5 *Flowchart Proses Remove Stopwords*

Gambar di atas menunjukkan langkah *remove stopwords*, di mana input yang digunakan berasal hasil proses *stemming*. Apabila teks yang diuji terdapat kata-kata *stopwords*, maka harus dihapus.

3.5.4 *Tokenizing*

Tokenizing merupakan tahapan untuk memecah atau memisahkan kalimat menjadi kata-kata yang menyusunnya. Kalimat yang dipisah menjadi kata berdasarkan spasinya, kata-kata yang telah dipecah biasa disebut dengan token. Pemisahan kalimat menjadi kata dalam tahapan *tokenizing* bertujuan untuk

digunakan dalam proses pembobotan kata dalam tahapan selanjutnya. Berikut *flowchart* dari proses *tokenizing* pada Gambar 3.6.

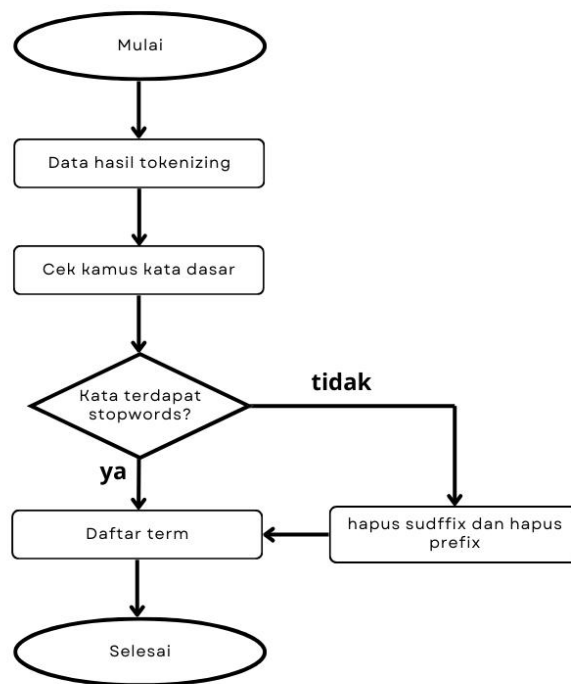


Gambar 3.6 *Flowchat* Proses *Tokenizing*

Gambar di atas menggambarkan tahap *tokenizing*, yang menggunakan data keluaran dari proses *remove stopwords* sebagai input. Selanjutnya, teks dipecah menjadi bagian-bagian kecil berupa token kata.

3.5.5 *Stemming*

Stemming adalah tahapan untuk menghilangkan imbuhan dalam sebuah kata sehingga menjadi kata dasar atau kata asli. Tujuan dilakukan proses *stemming* adalah untuk mengurangi variasi kata sehingga membuat proses selanjutnya menjadi lebih efisien. *Flowchart* dari proses *stemming* dapat dilihat pada Gambar 3.7.



Gambar 3.7 Flowchart Proses Stemming

Gambar di atas menunjukkan langkah proses stemming dengan menggunakan data hasil tokenisasi sebagai input. Apabila kata-kata dalam teks memiliki imbuhan *prefix* atau *suffix*, maka imbuhan tersebut akan dihilangkan.

3.6 Random Undersampling

Random Undersampling adalah suatu teknik untuk mengurangi kelas mayoritas menjadi berjumlah sama dengan kelas minoritas. Pengurangannya dilakukan secara acak sampai masing-masing kelas berjumlah sama. Tujuan dari penyamarataan kelas dengan menggunakan *Random Undersampling* adalah untuk mencegah bias model terhadap kelas mayoritas dan meningkatkan performa model secara keseluruhan, terutama pada kelas minoritas.

3.7 Frekuensi Kemunculan Kata

Frekuensi kemunculan kata merupakan salah satu metode dasar dalam analisis teks yang digunakan untuk mengetahui seberapa sering sebuah kata muncul dalam sekumpulan dokumen. Dengan menghitung jumlah kemunculan setiap kata, kita dapat mengidentifikasi kata-kata yang paling dominan, yang sering kali mencerminkan topik, sentimen, atau fokus utama dari suatu teks. Frekuensi ini juga berperan penting dalam berbagai teknik pemrosesan bahasa, seperti pembobotan TF-IDF, analisis sentimen, dan klasifikasi teks. Selain itu, informasi frekuensi kemunculan kata juga berguna dalam pembersihan data, seperti menghapus kata-kata yang terlalu jarang muncul (misalnya hanya satu kali), karena kata-kata tersebut cenderung tidak memberikan kontribusi yang signifikan dalam analisis dan justru dapat menambah *noise*. Berikut adalah frekuensi kemunculan data pada setiap dataset penelitian.

3.7.1 Frekuensi Kemunculan Kata Dataset 2024

Uji coba dataset tahun 2024 menggunakan dataset yang sudah difilter yang berjumlah 268 komentar positif dan 268 komentar negatif. Berikut adalah 20 kata teratas hasil frekuensi kemunculan data pada dataset 2024 yang dapat dilihat pada Tabel 3.3.

Tabel 3.3 Frekuensi Kemunculan Kata Dataset 2024

Frekuensi Kata Dataset 2024			
Kata	Frekuensi Positif	Frekuensi Negatif	Total Frekuensi
Makan	232	294	526
Gratis	177	295	472
Siang	152	248	400

Program	58	57	115
Anak	67	43	110
Sekolah	55	44	99
Didik	12	61	73
Pilih	17	45	62
Pakai	27	24	51
Negara	13	34	47
Moga	45	2	47
Kasih	21	25	46
Rakyat	15	30	45
Indonesia	28	12	40
Orang	14	23	37
Dana	11	25	36
Butuh	14	21	35
Gas	18	14	32
Uang	17	12	29
Gizi	21	6	27

Dari tabel diatas dapat dilihat hasil frekuensi kemunculan kata pada data berlabel positif dan negatif menunjukkan bahwa topik mengenai makan siang gratis menjadi pembahasan yang paling dominan di kedua label. Kata seperti makan, gratis, dan siang memiliki frekuensi tertinggi secara keseluruhan, mengindikasikan bahwa kebijakan makan siang gratis menjadi sorotan utama dalam diskusi, baik dalam konteks dukungan maupun kritik. Sementara itu, kata-kata seperti moga, kasih, dan gizi lebih sering muncul dalam label positif, menunjukkan nada harapan dan apresiasi terhadap program tersebut. Sebaliknya,

kata-kata seperti didik, negara, dana, dan pilih cenderung lebih dominan dalam label negatif, mencerminkan adanya kekhawatiran, kritik terhadap kebijakan, atau isu teknis pelaksanaan program. Secara keseluruhan, frekuensi kemunculan kata ini mencerminkan dinamika opini publik, di mana sebagian masyarakat mendukung program tersebut karena nilai sosial dan kesehatannya, sementara sebagian lainnya meragukan efektivitas atau keberlanjutan program dari sisi kebijakan dan anggaran.

Hasil frekuensi kemunculan kata menunjukkan bahwa banyak kata dominan seperti makan, gratis, siang, dan program muncul hampir seimbang pada label positif dan negatif. Keberadaan kata-kata ini dalam kedua kategori menunjukkan bahwa kata-kata tersebut sangat bergantung pada konteks kalimat secara keseluruhan. Dalam analisis sentimen menggunakan SVM, kondisi ini dapat menjadi tantangan karena SVM mengandalkan pola pada fitur kata untuk membedakan antar kelas. Apabila kata-kata utama yang digunakan sebagai fitur sering muncul di kedua kelas, maka margin pemisah SVM akan semakin sempit dan dapat menurunkan akurasi model. Ini menjadi indikasi bahwa data tersebut mungkin tidak sepenuhnya dapat dipisahkan secara linear, sehingga model SVM linear bisa saja tidak cukup kuat untuk menghasilkan klasifikasi yang optimal. Oleh karena itu, perlu dilakukan penggunaan representasi teks berbasis konteks seperti TF-IDF, dan pemilihan fitur. Dengan langkah-langkah tersebut, gangguan dari kata-kata yang muncul di kedua label dapat diminimalisasi dan performa model dapat ditingkatkan.

3.7.2 Frekuensi Kemunculan Kata Dataset 2025

Uji coba dataset tahun 2025 menggunakan dataset yang sudah difilter yang berjumlah 316 komentar positif dan 316 komentar negatif. Berikut adalah 20 kata teratas hasil frekuensi kemunculan data pada dataset 2025 yang dapat dilihat pada Tabel 3.4.

Tabel 3.4 Frekuensi Kemunculan Kata Dataset 2025

Frekuensi Kata Dataset 2025			
Kata	Frekuensi Positif	Frekuensi Negatif	Total Frekuensi
Makan	149	76	225
Anak	84	32	116
Sekolah	69	31	100
Gratis	58	41	99
Program	36	59	95
Syukur	84	3	87
Gizi	29	35	64
Korupsi	7	46	53
Kasih	30	23	53
Siang	29	14	43
Uang	27	15	42
Moga	37	2	39
Mending	5	26	31
Nasi	20	10	30
Rakyat	2	26	28
Sma	25	3	28
Senang	26	0	26
Mbg	4	21	25
Duit	4	20	24

Menu	14	9	23
------	----	---	----

Hasil analisis frekuensi kata menunjukkan bahwa data teks yang dianalisis banyak membahas topik-topik penting seperti pendidikan, bantuan sosial, emosi masyarakat, serta isu-isu politik dan ekonomi. Kata-kata seperti anak, sekolah, gizi, dan sma mengindikasikan bahwa topik pendidikan menjadi perhatian utama, dengan banyak sentimen positif yang menyertainya. Di sisi lain, ekspresi seperti syukur, moga, dan senang menunjukkan respons positif dan harapan masyarakat terhadap isu-isu tertentu. Namun, kata-kata seperti korupsi, rakyat, duit, dan uang lebih sering muncul pada label negatif, yang mencerminkan ketidakpuasan terhadap aspek-aspek pemerintahan atau pengelolaan dana publik. Temuan ini menunjukkan bahwa data mengandung opini yang kuat terhadap kebijakan dan kondisi sosial, sehingga penting bagi proses analisis sentimen untuk memahami konteks kata-kata tersebut guna menghasilkan klasifikasi yang lebih akurat dan bermakna.

Hasil analisis frekuensi kata yang ditampilkan tetap memungkinkan untuk dilakukan analisis sentimen menggunakan algoritma SVM, namun beberapa hal penting perlu diperhatikan agar hasil klasifikasi tidak terganggu. Distribusi frekuensi kata yang tidak seimbang antar label, seperti kata program dan korupsi yang lebih sering muncul dalam label negatif, serta syukur dan senang yang dominan pada label positif, justru dapat membantu SVM dalam membedakan karakteristik masing-masing kelas. Meski demikian, adanya kata-kata yang muncul pada kedua label dengan frekuensi tinggi, seperti gratis, anak, dan kasih,

dapat menyulitkan SVM dalam membentuk margin pemisah yang tegas. Oleh karena itu, penggunaan metode pembobotan seperti TF-IDF sangat disarankan agar kata-kata umum tersebut mendapatkan bobot yang sesuai. Selain itu, keberadaan kata-kata bermuatan emosional atau topikal seperti korupsi, rakyat, syukur, dan duit sangat membantu dalam membentuk fitur yang mendukung akurasi model. Meski begitu, perlu diwaspadai potensi overfitting terhadap kata-kata unik yang hanya muncul satu kali, sehingga pembersihan terhadap kata-kata dengan frekuensi rendah tetap perlu dilakukan. Dengan preprocessing yang tepat seperti penyeimbangan data, pembobotan, dan filter kata jarang, analisis sentimen menggunakan SVM tidak hanya tetap dapat dilakukan, tetapi juga memiliki peluang untuk menghasilkan model klasifikasi yang efektif dan akurat.

3.7.3 Frekuensi Kemunculan Kata Dataset Gabungan

Uji coba dataset gabungan menggunakan dataset yang sudah difilter yang berjumlah 593 komentar positif dan 593 komentar negatif. Berikut adalah 20 kata teratas hasil frekuensi kemunculan data pada dataset 2025 pada Tabel 3.5.

Tabel 3.5 Frekuensi Kemunculan Kata Dataset Gabungan

Frekuensi Kata Dataset Gabungan			
Kata	Frekuensi Positif	Frekuensi Negatif	Total Frekuensi
Makan	384	515	899
Gratis	236	485	721
Siang	181	385	566
Program	94	138	232
Anak	151	79	230
Sekolah	124	85	209

Didik	12	114	126
Kasih	52	63	115
Pilih	22	73	95
Syukur	87	4	91
Moga	82	5	87
Gizi	51	33	84
Rakyat	17	62	79
Uang	44	34	78
Indonesia	41	32	73
Pakai	35	32	67
Negara	18	49	67
Orang	29	37	66
Dana	12	52	64
Butuh	29	27	56

Kata-kata dengan frekuensi tinggi seperti makan, gratis, siang, anak, sekolah, dan gizi menunjukkan bahwa topik terkait program makanan gratis, khususnya untuk anak-anak sekolah, merupakan tema sentral dalam dataset ini. Hal ini diperkuat oleh kemunculan kata program, didik, dan kasih, yang menunjukkan adanya narasi bantuan sosial atau kebijakan pemerintah yang berorientasi pada pendidikan dan kesejahteraan anak. Selanjutnya, munculnya kata rakyat, negara, pilih, dan dana mengindikasikan keterlibatan topik politik dan kebijakan publik, mungkin dalam konteks janji kampanye, kritik kebijakan, atau harapan masyarakat. Di sisi emosional, kata syukur, moga, dan kasih yang banyak muncul dalam label positif mencerminkan ekspresi dukungan, harapan, atau rasa

terima kasih dari masyarakat. Sebaliknya, kata didik, pilih, dan negara yang lebih dominan dalam label negatif bisa jadi mengandung kritik atau ketidakpuasan terhadap kebijakan terkait. konteks data sebelum melakukan klasifikasi sentimen, agar dapat menyesuaikan preprocessing dan pemilihan fitur secara lebih kontekstual.

Berdasarkan hasil gabungan frekuensi kata positif dan negatif tersebut, penggunaan SVM untuk analisis sentimen tidak akan terganggu secara signifikan, namun ada beberapa hal penting yang perlu diperhatikan agar hasil klasifikasi tetap akurat. Beberapa kata seperti makan, gratis, dan siang memiliki frekuensi yang sangat tinggi di kedua label (positif dan negatif), yang artinya kata-kata tersebut tidak secara eksklusif mencerminkan satu jenis sentimen saja. Ini bisa menyebabkan ambiguitas dalam pemodelan karena SVM akan kesulitan memisahkan sentimen hanya berdasarkan kata-kata tersebut. Namun, ini bukan hambatan serius, selama model juga mempertimbangkan konteks kata lain dalam kalimat yang membedakan sentimen.

Salah satu strategi mengatasi hal ini adalah dengan menggunakan metode pembobotan seperti TF-IDF, yang dapat menurunkan pengaruh kata-kata umum dan menaikkan bobot kata-kata yang lebih khas untuk masing-masing sentimen. Selain itu, pemrosesan lanjutan seperti penghapusan kata *non-discriminatif*, pemilihan fitur yang informatif, atau menggunakan n-gram (bukan hanya unigram) juga bisa membantu memperjelas konteks sentimen.

Dengan demikian, meskipun terdapat kata-kata yang muncul secara dominan di kedua label, hal ini tidak secara langsung mengganggu performa analisis sentimen dengan SVM asalkan preprocessing, pemilihan fitur, dan representasi data dilakukan secara tepat. Model SVM justru sangat efektif dalam mencari garis pemisah terbaik meski ada tumpang tindih pada sebagian fitur.

3.8 TF-IDF

TF-IDF merupakan suatu teknik yang bertujuan untuk mengevaluasi suatu dokumen untuk mengetahui tingkat kepentingan kata di dalamnya. *Term Frequency* (TF) digunakan untuk mengukur jumlah kata yang sering muncul dalam sebuah dokumen. Sementara itu, *Inverse Document Frequency* (IDF) digunakan untuk mengukur jumlah kata yang jarang muncul dalam suatu dokumen (Husain et al., 2024). Dalam analisis sentimen TF-IDF digunakan untuk mengubah kalimat sentimen menjadi sebuah angka. Rumus dalam menghitung TF dapat dilihat dalam Persamaan 3.1.

$$TF_{(t,d)} = \frac{f_{t,d}}{\sum_k f_{k,d}} \quad (3.1)$$

$f_{t,d}$ = jumlah kemunculan kata t dalam dokumen d

$\sum_k f_{k,d}$ = jumlah total kata dalam dokumen d

Berikut adalah contoh hasil perhitungan TF yang dapat dilihat dalam bentuk Tabel 3.6.

Tabel 3.6 Hasil Perhitungan TF

Dokumen	Jumlah Kata	Kata	Frekuensi	TF
1	25		1	1/25

2	9	gratis	1	1/9
3	21		1	1/21
4	4		1	1/4
5	10		1	1/10

Langkah selanjutnya adalah menghitung nilai dari IDF. Berikut adalah rumus dari perhitungan IDF yang dapat dilihat dalam Persamaan 3.2.

$$IDF_{(t)} = \log_e \frac{N}{df_{(t)}} \quad (3.2)$$

N = jumlah total dokumen dalam korpus

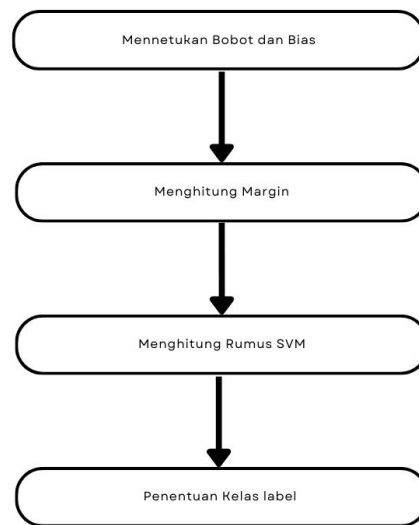
$df_{(t)}$ = jumlah dokumen yang mengandung kata t

Dalam perhitungan TF-IDF menggabungkan antara TF dan IDF untuk memperoleh nilai yang mempresentasikan relevansi kata dalam dokumen tersebut. Dalam metode ini, semakin sering sebuah kata dalam sebuah dokumen maka semakin kecil nilai bobotnya. Sebaliknya, semakin jarang sebuah kata dalam sebuah dokumen maka semakin besar bobot nilainya. Berikut adalah rumus perhitungan TF-IDF yang dapat dilihat dalam Persamaan 3.3.

$$TF - IDF_{(t,d)} = TF_{(t,d)} \times IDF_{(t)} \quad (3.3)$$

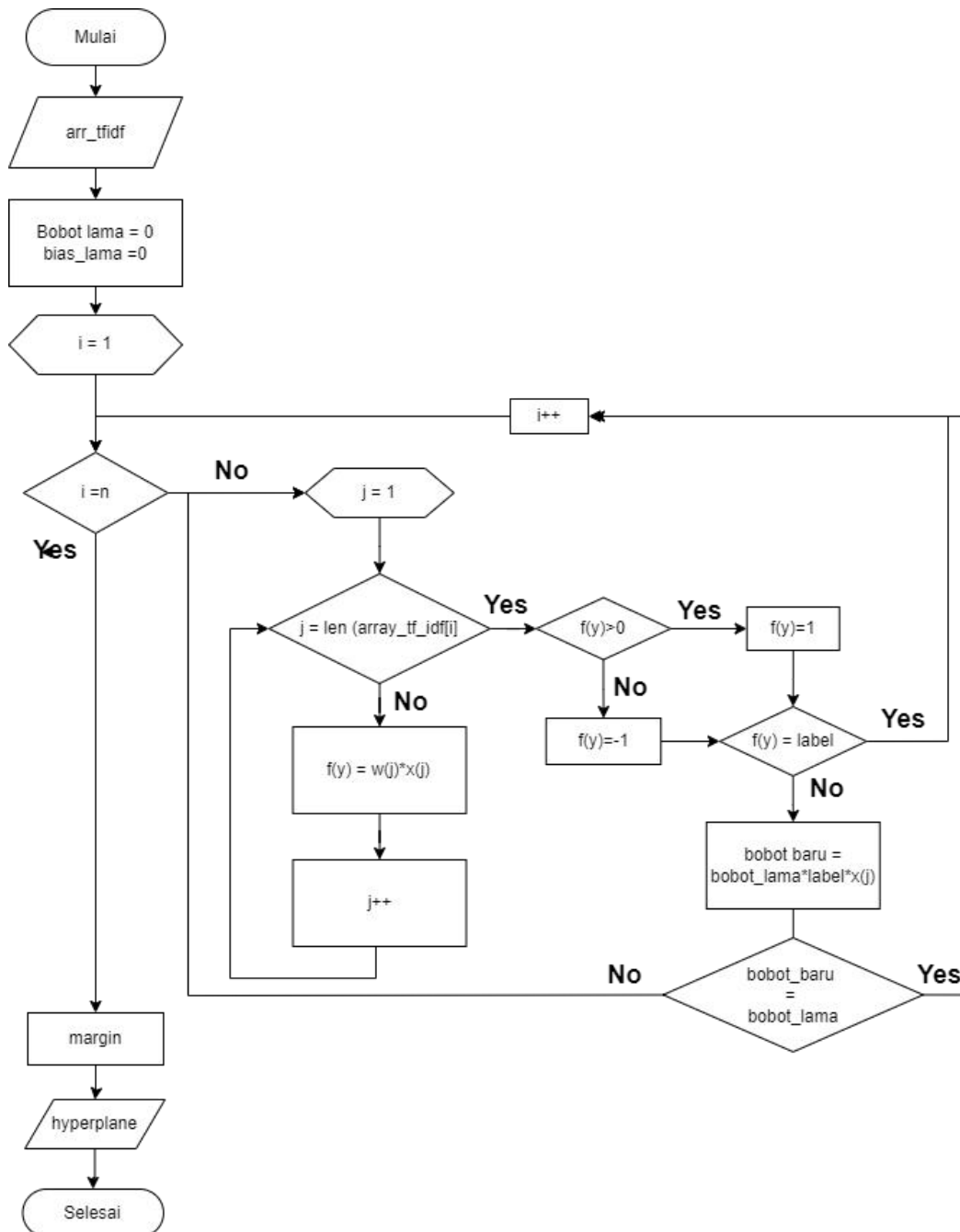
3.9 Support Vector Machine

Penelitian ini menggunakan algoritma SVM, yang menggunakan model *binary linear* karena hanya menggunakan dua label yaitu sentimen negatif dan positif. Berikut adalah *flowchart* dari desain algoritma SVM pada Gambar 3.8.



Gambar 3.8 *Flowchart* Algoritma *Support Vector Machine*

Hasil klasifikasi dengan menggunakan algoritma SVM didapatkan berdasarkan nilai input terhadap *hyperplane* yang telah dibuat, Berikut adalah *flowchart* dalam pembentukan *hyperplane* pada algoritma SVM pada Gambar 3.9 (Andhika et al., 2023).



Gambar 3.9 Flowchart pencarian hyperplane

SVM merupakan algoritma *supervised learning*, yang sering digunakan untuk melakukan klasifikasi pada analisis sentimen. Tujuan algoritma SVM adalah untuk mencari sebuah *hyperplane* atau garis yang memisahkan dua kelas. Dalam SVM terdapat beberapa kernel, kernel yang digunakan dalam penelitian ini

akan menggunakan SVM kernel *Linear* dan *Polynomial*. Secara umum rumus SVM kernel linear dapat dilihat dalam pada Persamaan 3.4.

$$w \cdot x + b = 0 \quad (3.4)$$

W merupakan vektor bobot dari fitur yang mendefinisikan orientasi *hyperplane*, X merupakan vektor fitur dari hasil TF-IDF, dan b merupakan bias atau *offset* dari *hyperplane*. Pada analisis sentimen pasti memiliki lebih dari satu bobot yang didapat dari dilakukannya proses TF-IDF, sehingga persamaan SVM diubah dapat dilihat dalam Persamaan 3.5.

$$y = \sum_{i=1}^n w_i \cdot x_i + b \quad (3.5)$$

y = Hasil prediksi

W_i = Bobot fitur ke-I

X_i = Nilai fitur ke-I

B = bias

Dari perhitungan klasifikasi ini, untuk memperoleh nilai bobot dan bias yang optimal maka dilakukan dengan cara *Gradient Descent*. *Gradient Descent* digunakan untuk meminimalisir *loss* dengan memperbaharui nilai bobot dan bias melibatkan *gradient* dan *learning rate*. Berikut adalah rumus untuk memperbaharui bobot pada Persamaan 3.6

$$W(\text{baru}) = W(\text{lama}) - \alpha * W(\text{lama}) + \alpha * \text{gradien } W \quad (3.6)$$

Sedangkan untuk pada rumus SVM *Polynomial* akan ditunjukkan pada Persamaan 3.7

$$K(x, x') = (\gamma \cdot x \cdot x' + r)^d \quad (3.7)$$

x, x' = Dua vektor data hasil TF-IDF

$x \cdot x'$ = Perkalian dua vektor

γ = Skala (biasanya default 1 / jumlah fitur)

r = Konstanta (bisa nol atau satu)

d = Derajat polinomial

Sedangkan untuk analisis sentimen yang menggunakan banyak data menggunakan rumus seperti dalam Persamaan 3.8.

$$\sum_{i=1}^n w_i y_i \cdot K(x_i, x) + b \quad (3.8)$$

x = Data baru yang akan diprediksi

xI = Data pelatihan yang jadi support vektor

w_i = Bobot (dihitung saat traning)

γI = Label dari data pelatihan (+1 positif, -1 negatif)

$K(x_i, x)$ = Kernel Polynomial

b = bias

3.10 Confusion Matrix

Dari skenario tersebut, akan dibandingkan nilai dari *confusion matrix*. *confusion matrix* merupakan tabel hasil klasifikasi yang berisi nilai benar atau salah. Berikut adalah tabel *confusion matrix* pada Tabel 3.7.

Tabel 3.7 *Confusion Matrix*

Kelas Sebenarnya	Kelas Prediksi	
	Positif	Negatif
Positif	TP	FP
Negatif	FN	TN

Pada tabel di atas dapat dilihat pemetaan dari nilai *confusion matrix* sehingga dari pemetaan tersebut dapat dilakukan perhitungan dengan Persamaan 3.9, 3.10, 3.11, dan 3.12.

$$Akurasi(\%) = \frac{TP + TN}{TP + FP + TN + FN} \times 100\% \quad (3.9)$$

$$Presisi(\%) = \frac{TP}{TP + FP} \times 100\% \quad (3.10)$$

$$Recall(\%) = \frac{TP}{TP + FN} \times 100\% \quad (3.11)$$

$$F - measure(\%) = \frac{2 \times Presisi \times Recall}{Presisi + Recall} \quad (3.12)$$

Berdasarkan penelitian oleh Putra et al. (2023), model SVM yang diterapkan pada data ulasan produk dalam bahasa Indonesia berhasil mencapai akurasi sebesar 68%–72%, dan hasil tersebut dianggap memadai untuk klasifikasi biner. Selain itu, dalam studi oleh Rahmadani & Yusuf (2024), SVM menghasilkan akurasi sebesar 70,4% dalam analisis sentimen pada data media sosial, dan disimpulkan bahwa performa tersebut dapat diterima untuk keperluan

penelitian maupun aplikasi dasar sistem pendukung keputusan. Oleh karena itu, akurasi 70% mencerminkan kemampuan model yang kompeten dalam mengenali pola sentimen, meskipun masih dapat ditingkatkan dengan optimalisasi fitur seperti pemilihan kata kunci, balancing data, atau penggunaan model hibrida.

BAB IV

HASIL DAN PEMBAHASAN

Bab ini menyajikan hasil dari pengujian sistem yang mencakup beberapa tahapan utama, yaitu preprocessing teks, penyamartaan data dengan *Random Undersampling*. Perhitungan *Term Frequency-Inverse Document Frequency* (TF-IDF), serta proses klasifikasi menggunakan *Support Vector Machine* (SVM). Selain itu, evaluasi kinerja dilakukan dengan menghitung akurasi, presisi, *recall*, dan *F-measure* menggunakan confusion matrix.

4.1 Data Penelitian

Seperti telah dijelaskan sebelumnya, data yang digunakan dalam penelitian ini berasal dari komentar Instagram pada unggahan akun yang membahas program makan siang gratis. Dari postingan tahun 2024 terkumpul sebanyak 933 komentar, sedangkan dari postingan tahun 2025 diperoleh 641 komentar. Proses pengambilan data komentar dilakukan dengan memanfaatkan salah satu ekstensi pada browser. Berikut adalah perbandingan jumlah komentar negatif dan positif yang dapat dilihat dalam Tabel 4.1.

Tabel 4.1 Perbandingan Jumlah Data

Data	Label	
	Positif	Negatif
2024	268	665
2025	325	316
Gabungan	593	981

4.2 Preprocessing Data

Preprocessing dalam analisis sentimen merupakan tahapan untuk menghapus kata, tanda baca, emoticon yang tidak penting dalam sebuah sentimen untuk memperoleh kata yang penting dan mempermudah pengklasifikasian. *Output* dari *preprocessing* merupakan data yang bersih yang berisi kata-kata penting. Tahapan dari *Preprocessing* adalah sebagai berikut.

1. *Cleaning*

Cleaning merupakan tahapan untuk menghapus emoticon, huruf, simbol, dan tanda baca yang tidak berpengaruh terhadap proses klasifikasi. Berikut adalah kode program pada proses *cleaning* yang diperlihatkan dalam Gambar 4.1.

```
import re
def clean_text(text):
    text = re.sub(r'@[A-Za-z0-9_]+', '', text)
    text = re.sub(r'[0-9]+', '', text)
    text = re.sub(r'#\w+', '', text)
    text = re.sub(r'https?://\S+', '', text)
    text = re.sub(r'#', '', text)
    text = re.sub(r'^[A-Za-z0-9 ]', '', text)
    text = re.sub(r'\s+', ' ', text).strip()
    return text
df['clean'] = df['komentar'].apply(clean_text)
```

Gambar 4.1 Kode proses *Cleaning*

Gambar di atas adalah kode tahapan *cleaning* yang memiliki input judul dari sebuah teks berita. Kemudian apabila terdeteksi angka dan tanda baca pada kalimat, maka akan dihapus. Hasil tahap *cleaning* ini ditunjukkan pada Tabel 4.3.

Tabel 4.3 Hasil proses *Cleaning*

Sebelum	Sesudah
“@necthophyle112 Sejujurnya banyak anak sekolah yg bahkan gak jajan karena gak bawa uang jajan, makanan gratis ini semoga bisa bermanfaat.”	“Sejujurnya banyak anak sekolah yg bahkan gak jajan karena gak bawa uang jajan makanan gratis ini semoga bisa bermanfaat”

2. Case Folding

Case folding adalah tahapan yang merubah huruf kapital menjadi huruf kecil untuk lebih mengoptimalkan proses klasifikasi. Berikut adalah contoh kode program pada proses *case folding* pada Gambar 4.2.

```
# Mengecilkan huruf teks pada kolom 'clean'
df['lower'] = df['clean'].str.lower()
```

Gambar 4.2 Kode proses *Case Folding*

Kode di atas merupakan tahapan *case folding* yang mempunyai input teks hasil *cleaning*. Teks tersebut akan diubah menjadi format *lowercase*. Berikut adalah contoh hasil dari tahap *case folding* terdapat pada Tabel 4.4.

Tabel 4.4 Hasil Proses *Case Folding*

Sebelum	Sesudah
“Sejujurnya banyak anak sekolah yg bahkan gak jajan karena gak bawa uang jajan makanan gratis ini semoga bisa bermanfaat”	“sejujurnya banyak anak sekolah yg bahkan gak jajan karena gak bawa uang jajan makanan gratis ini semoga bisa bermanfaat”

3. Remove stopwords

Remove stopwords merupakan proses untuk menghapus kata yang tidak berpengaruh pada hasil klasifikasi contohnya adalah kata hubung. Berikut adalah kode program pada proses *remove stopwords* dapat dilihat dalam Gambar 4.3.

```
# Definisikan fungsi untuk menghapus stopwords
def remove_stopwords(text, custom_stopwords=set()):
    stop_words = set(stopwords.words('indonesian'))
    # Gabungkan stopwords dari NLTK dengan stopwords kustom
    stop_words = stop_words.union(custom_stopwords)
    word_tokens = word_tokenize(text)
    filtered_text = [word for word in word_tokens if word.lower() not in
stop_words]
    return ' '.join(filtered_text)
```

Gambar 4.3 Kode proses *Remove Stopwords*

Kode di atas adalah tahapan *stopwords removal* yang mempunyai input data hasil tahapan *case folding*. Jika teks tersebut mengandung *stopwords*, maka akan dihapus. Hasil tahap *stopwords removal* dapat ditunjukkan oleh Tabel 4.5.

Tabel 4.5 Hasil Proses *Remove Stopwords*

Sebelum	Sesudah
“sejujurnya banyak anak sekolah yg bahkan gak jajan karena gak bawa uang jajan makanan gratis ini semoga bisa bermanfaat”	“sejujurnya anak sekolah jajan bawa uang jajan makanan gratis semoga bermanfaat”

4. *Tokenizing*

Tokenizing adalah untuk proses memisah kata yang dipisahkan berdasarkan spasi yang berguna untuk pembobotan kata. Berikut adalah gambar kode tahapan *tokenizing* yang diperlihatkan dalam Gambar 4.4.

```
# Fungsi untuk melakukan tokenisasi
```

```
def tokenize_text(text):
    tokens = word_tokenize(text)
    return tokens

# Misalnya kita ingin melakukan tokenisasi pada kolom 'text'
df['tokens'] = df['stopword'].apply(tokenize_text)
```

Gambar 4.4 Kode proses *Tokenizing*

Kode di atas adalah tahapan *tokenizing* yang memiliki input data hasil tahapan *remove stopwords*. Teks tersebut diubah menjadi potongan kata. Berikut hasil tahapan *tokenizing* terdapat pada Tabel 4.6.

Tabel 4.6 Hasil Proses *Tokenizing*

Sebelum	Sesudah
“sejujurnya anak sekolah jajan bawa uang jajan makanan gratis semoga bermanfaat”	“sejujurnya”,”anak”,”sekolah”,”jajan”,”bawa”, ”uang”,”jajan”,”makanan”,”gratis”,”semoga”, ”bermanfaat”

5. *Stemming*

Stemming merupakan proses untuk menghapus imbuhan pada data yang akan dilakukan klasifikasi. Penghapusan imbuhan dilakukan karena tidak berpengaruh terhadap proses klasifikasi. Berikut adalah kode program *stemming* yang dapat dilihat dalam Gambar 4.5.

```
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
from collections import Counter

factory = StemmerFactory()
stemmer = factory.create_stemmer()

def stemmed_wrapper(term):
    return stemmer.stem(term)

term_base = {}

for document in df['tokens']:
```

```

for term in document:
    if term not in term_base:
        term_base[term] = " "
for term in term_base:
    term_base[term] = stemmed_wrapper(term)
def get_stemmed_term(document):
    return [term_base[term] for term in document]
df['stemmed'] = df['tokens'].apply(get_stemmed_term)

```

Gambar 4.5 Kode proses *Stemming*

Kode di atas adalah tahapan proses *stemming* yang mempunyai input berupa data hasil *tokenizing*. Jika teks terdapat *prefix* dan *suffix*, maka akan dihapus. Berikut hasil proses *stemming* dapat dilihat oleh Tabel 4.7.

Tabel 4.7 Hasil Proses *Stemming*

Sebelum	Sesudah
“sejujurnya”, “anak”, “sekolah”, “jajan”, “bawa”, “uang”, “jajan”, “makanan”, “gratis”, “semoga”, “bermanfaat”	“jujur”, “anak”, “sekolah”, “jajan”, “bawa”, “uang”, “jajan”, “makan”, “gratis”, “moga”, “manfaat”

4.3 Random Undersampling

Random Undersampling merupakan metode yang digunakan untuk menyeimbangkan jumlah data antara kelas mayoritas dan minoritas dengan cara mengurangi data dari kelas mayoritas secara acak hingga jumlahnya sama dengan kelas minoritas. Teknik ini bertujuan agar model tidak terlalu condong pada kelas mayoritas sehingga dapat meningkatkan kinerja model secara menyeluruh, khususnya dalam mengenali kelas minoritas. Berikut adalah kode program *random undersampling* yang dapat dilihat dalam Gambar 4.6.

```

from imblearn.under_sampling import RandomUnderSampler

print("Distribusi label sebelum undersampling:")

print(df['label'].value_counts())

# Ambil data label

y = df['label']

# Konversi ke format DataFrame agar sesuai dengan `fit_resample`

X = df['filtered_stemmed'].apply(lambda tokens: ' '.join(tokens)).to_frame()

# Lakukan Random Undersampling

undersampler = RandomUnderSampler(random_state=42)

X_resampled, y_resampled = undersampler.fit_resample(X, y)

# Buat DataFrame hasil undersampling

df_resampled = pd.DataFrame({'processed_text_str': X_resampled.squeeze(), 'label': y_resampled})

```

Gambar 4.6 Kode proses *Random Undersampling*

Kode di atas menggunakan input dari hasil preprocessing data yang telah dilakukan. Berikut adalah hasil random undersampling yang dapat dilihat pada Tabel 4.8.

Tabel 4.8 Hasil *Random Undersampling*

Data	Random Undersampling			
	Sebelum		Sesudah	
	Positif	Negatif	Positif	Negatif
2024	268	665	268	268
2025	325	316	316	316
Gabungan	593	981	593	593

4.4 TF-IDF

TF-IDF merupakan proses untuk melakukan pembobotan kata yang digunakan untuk proses klasifikasi SVM. Tujuan dari TF-IDF adalah untuk merubah kata menjadi sebuah angka yang memiliki bobot. Berikut adalah kode proses TF-IDF yang dapat dilihat pada Gambar 4.7.

```
# Fungsi untuk menghitung TF-IDF

def calculate_tfidf(tf_df, idf_values):

    tfidf_table = tf_df.copy() # Salin TF DataFrame untuk dihitung

    # Kalikan TF dengan IDF untuk setiap kata

    for word in tfidf_table.columns:

        tfidf_table[word] = tfidf_table[word] * idf_values[word]

    return tfidf_table

# Hitung TF-IDF

tfidf_df = calculate_tfidf(tf_df, idf_values)
```

Gambar 4.7 Kode Proses TF-IDF

Kode di atas adalah proses TF-IDF yang dapat dilakukan dengan cara melakukan perkalian antara proses TF dan IDF. Berikut adalah hasil dari TF-IDF yang dapat dilihat pada Tabel 4.9.

Tabel 4.9 Hasil Perhitungan TF-IDF

Kata	TF-IDF				
	D1	D2	D3	D4	D5
tuju	0.1452923	0.0	0.0	0.0	0.0
makan	0.0083479	0.02318868	0.00993800	0.052174531	0.020869812
siang	0.0126958	0.03526623	0.01511409	0.079349022	0.031739608
gratis	0.0111560	0.03098915	0.01328106	0.069725596	0.027890238
baik	0.1781315	0.0	0.0	0.0	0.0

gizi	0.1307066	0.0	0.0	0.0	0.0
cegah	0.3924219	0.0	0.0	0.0	0.0
stunting	0.3125396	0.0	0.0	0.0	0.390674594
sekolah	0.0790395	0.0	0.09409464	0.0	0.0
suka	0.1746511	0.0	0.0	0.0	0.0
jls	0.2740473	0.0	0.0	0.0	0.0
efektif	0.2023770	0.0	0.0	0.0	0.0
prio	0.2740473	0.0	0.0	0.0	0.0
anak	0.2527503	0.0	0.0	0.0	0.0
alergi	0.2463215	0.0	0.0	0.0	0.0

4.5 Perhitungan Algoritma *Support Vector Machine*

Tahapan ini merupakan proses analisis sentimen algoritma SVM yang melakukan prediksi untuk melabeli data menjadi positif atau negatif. SVM dilakukan dengan menggunakan dua data dalam pengklasifikasian yaitu data *training* dan data *testing*. Berikut adalah kode klasifikasi SVM yang dilakukan pada penelitian ini yang dapat dilihat pada Gambar 4.8.

```
for ratio in split_ratios:
    print(f"\n===== Pembagian Data {int((1 - ratio) * 10)}:{int(ratio * 10)} =====")
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=ratio, random_state=42,
    stratify=y)
    svm_model = SVC(kernel='poly', degree=3, coef0=2, gamma='scale', C=1, random_state=42)
    svm_model.fit(X_train, y_train)
    y_pred = svm_model.predict(X_test)
    accuracy = accuracy_score(y_test, y_pred)
    report = classification_report(y_test, y_pred)
```

Gambar 4.8 Kode proses SVM

Kode di atas adalah tahapan klasifikasi SVM dengan menggunakan SVM dengan cara memanggil *library* yang sudah tersedia. Berikut adalah beberapa sampel hasil pengujian yang telah dilakukan yang dapat dilihat dalam Tabel 4.10.

Tabel 4.10 Hasil Klasifikasi SVM

No	Kalimat	Label sebenarnya	Label prediksi
1	deddy.hartadiDengan budget super irit kualitas gizi macam apa yang bisa diharapkan 2w11 likesReplySee translationComment OptionsLike	N	P
2	baguspras34sbyKalau untuk menunya mungkin bisa nasi dengan suwiran ayam dan irisan telur dadar tahu atau tempe goreng di tambah tumis sayur seperti wortel buncis di tambah dengan buah dan air mineral gelas sepertinya masih masuk akal karena jika memaksakan dengan susu kotak maka akan semakin tipis keuntungan pihak yang menyiapkan makanan tersebut atau bahkan bisa" tidak mendapatkan keuntungan jika dengan anggaran 10 ribu per porsi .12wReplySee translationComment OptionsLike	P	P
3	ridwan124521 Setuju tidak monoton 2wReplySee translationComment OptionsLike	P	N
4	arafly200Keroco nilep lagi 11wReplyComment OptionsLike	N	N
5	mjnaymataraFaktor didikan orang tua sangat berpengaruh pada anaknya soal adab dan rasa bersyukur 10wReplySee translationComment OptionsLike	P	P
6	milie.alexander Alhamdulillah walaupun kebijakan ini penuh drama semoga dapat konsisten dan bermanfaat bagi anak Indonesia. Walaupun belum sempurna semoga ini awal yg baik untuk pengentasan gizi buruk & kecerdasan generasi penerus. Satu lagi tidak dikorupsi tentunya 12wReplySee translationComment OptionsLike	P	P
...
1573	randy_liverpoolDoktrin politisasi sejak dini . Terimakasih terimakasih terimakasih terimakasih BERTERIMA KASIH LAH PADA RAKYAT DARI DUIT PAJAK RAKYAT BISA BIKIN KALIAN	N	P

	MAKAN. KALO SAMA SI OMON ² ENTAR ADA KEPENTINGAN UTANG BUDI KALIAN SEMUA. kasihan10wReplySee translationComment OptionsLike		
1574	sulis.tiawati15Anak 90an yg uang sakunya di SMA 3rb pingin ngerasain makan siang gratis 9wReplySee translationComment OptionsLike	P	P

Berdasarkan keseluruhan data yang diujicobakan terdapat 1574 data yang sama antara label sebenarnya dan label prediksi. SVM berhasil mendapatkan hasil akurasi sebesar 78 % .

4.6 Skenario Pengujian

Pengujian dilakukan untuk memperoleh besar nilai akurasi, presisi, *recall*, dan *f-measure* menggunakan algoritma SVM. Pengujian dilakukan dengan 3 variasi jumlah data yaitu data tahun 2024, data tahun 2025, dan gabungan dari tahun 2024 dan 2025 dengan perbandingan jumlah data 933, 641, dan 1574. Selain itu, pengujian juga akan dilakukan dengan 2 skenario jumlah data latih yaitu 8:2, dan 9:1. Kemudian dilakukan dengan 2 skenario kernel yaitu linear dan *polynomial* sehingga menghasilkan total 12 skenario pengujian. Berikut adalah skenario pengujian yang dapat dilihat dalam Tabel 4.2.

Tabel 4.2 Skenario Pengujian

Kernel	Data	Variasi Data		Nilai Performa			
		Data Latih	Data Uji	Akurasi	Presisi	Recall	F-measure
Linear	2024	80%	20%				
		70%	30%				
	2025	80%	20%				
		70%	30%				

	Gabung	80%	20%				
		70%	30%				
Polynomial	2024	80%	20%				
		70%	30%				
	2025	80%	20%				
		70%	30%				
	Gabung	80%	20%				
		70%	30%				

4.7 Hasil Uji Coba

Data penelitian yang sudah didapat kemudian dilakukan pengujian sistem yang mencakup beberapa tahapan utama, yaitu preprocessing teks, perhitungan TF-IDF. Kemudian dilakukan penyamarataan data dengan *random undersampling*. Setelah itu, dilakukan proses klasifikasi menggunakan SVM berdasarkan skenario pengujian. Seperti pada skenario pengujian yang telah dijelaskan pada bab sebelumnya, pengujian akan dilakukan dengan 3 dataset berbeda yaitu dataset 2024, 2025, dan gabungan keduanya, serta menggunakan 2 rasio data latih dan data uji yang berbeda dan 2 kernel yang berbeda yaitu *Linear* dan *Polynomial*.

4.5.1 Hasil Uji Coba Dataset 2024

Dalam uji coba dataset 2024 berjumlah 933 data yang mempunyai 268 data berlabel positif dan 665 data berlabel negatif. Data tersebut kemudian dilakukan penyamarataan data dengan menggunakan *random undersampling* sehingga data menjadi 268 data positif dan 268 data negatif. Berikut adalah hasil *confusion matrix* yang dapat dilihat dalam Tabel 4.11.

Tabel 4.11 Hasil *Confusion Matrix* Dataset 2024

Kernel	Rasio Data	Kelas Sebenarnya	Kelas Prediksi	
			Positif	Negatif
Linear	8:2	Positif	35	19
		Negatif	11	43
	7:3	Positif	56	24
		Negatif	36	45
Polynomial	8:2	Positif	35	19
		Negatif	10	44
	7:3	Positif	57	23
		Negatif	25	56

Berdasarkan tabel di atas maka didapatkan hasil akurasi, presisi, *recall*, dan *f-measure* yang dapat dilihat dalam Tabel 4.12.

Tabel 4.12 Hasil Akurasi, Presisi, *Recall*, dan *F-measure* Dataset 2024

Kernel	Variasi Data		Nilai Performa			
	Data Latih	Data Uji	Akurasi	Presisi	Recall	F-measure
Linear	80%	20%	0.722	0.73	0.72	0.72
	70%	30%	0.62	0.63	0.63	0.63
Polynomial	80%	20%	0.7315	0.74	0.73	0.73
	70%	30%	0.7019	0.70	0.70	0.70

Dari hasil diatas dapat dilihat pada kernel *Linear*, akurasi yang diperoleh dengan *split* 80:20 adalah 72,2%, dengan presisi 73%, *recall* 72%, dan *f1-score* 72%. Namun, saat proporsi data latih dikurangi menjadi 70%, performa *Linear* kernel menurun cukup signifikan, dengan akurasi hanya mencapai 62%, serta presisi, *recall*, dan *f1-score* yang masing-masing berada di angka 63%. Penurunan

performa ini menunjukkan bahwa *Linear* kernel cukup sensitif terhadap jumlah data latih.

Sementara itu, kernel *Polynomial* menunjukkan performa yang lebih unggul dan stabil. Dengan split 80:20, diperoleh akurasi sebesar 73,15%, presisi 74%, *recall* 73%, dan *f1-score* 73%. Pada split 70:30, meskipun terjadi sedikit penurunan, performanya tetap tinggi dengan akurasi 70,19% dan metrik lainnya berada di kisaran 70%. Hal ini menunjukkan bahwa *Polynomial* kernel lebih mampu menangkap pola *non-linear* dalam data dan lebih stabil meskipun jumlah data latih berkurang. Selain itu, penurunan performa pada proporsi data latih yang lebih kecil juga menandakan bahwa penambahan data latih berpotensi meningkatkan kinerja model secara signifikan.

4.5.2 Hasil Uji Coba Dataset 2025

Dalam uji coba dataset 2025 berjumlah 641 data yang mempunyai 325 data berlabel positif dan 316 data berlabel negatif. Data tersebut kemudian dilakukan peyamarataan data dengan menggunakan *random undersampling* sehingga data menjadi 316 data positif dan 316 data negatif. Berikut adalah hasil *confusion matrix* yang dapat dilihat dalam Tabel 4.13.

Tabel 4.13 Hasil *Confusion Matrix* Dataset 2025

Kernel	Rasio Data	Kelas Sebenarnya	Kelas Prediksi	
			Positif	Negatif
Linear	8:2	Positif	48	20
		Negatif	12	52

Polynomial	7:3	Positif	75	20
		Negatif	24	71
	8:2	Positif	51	12
		Negatif	15	49
	7:3	Positif	75	23
		Negatif	20	78

Berdasarkan tabel di atas maka didapatkan hasil akurasi, presisi, *recall*, dan *f-measure* yang dapat dilihat dalam Tabel 4.12.

Tabel 4.14 Hasil Akurasi, Presisi, *Recall*, dan *F-measure* Dataset 2025

Kernel	Variasi Data		Nilai Performa			
	Data Latih	Data Uji	Akurasi	Presisi	Recall	F-measure
Linear	80%	20%	0.7874	0.79	0.79	0.79
	70%	30%	0.7648	0.77	0.77	0.77
Polynomial	80%	20%	0.7874	0.79	0.79	0.79
	70%	30%	0.7832	0.77	0.76	0.76

Dalam tabel diatas pada kernel Linear, performa model cukup konsisten. Dengan pembagian data 80:20, diperoleh akurasi sebesar 78,74% dengan presisi, *recall*, dan *f1-score* masing-masing sebesar 79%. Saat proporsi data latih dikurangi menjadi 70%, akurasi sedikit menurun menjadi 76,48%, dengan presisi, *recall*, dan *f1-score* berada di angka 77%. Penurunan performa ini relatif kecil, menunjukkan bahwa kernel *Linear* cukup stabil dan tidak terlalu sensitif terhadap perubahan jumlah data latih dalam dataset ini.

Sementara itu, kernel *Polynomial* pada split 80:20, hasil pengujian sama persis dengan *Linear* kernel, yaitu akurasi 78,74% dan metrik lainnya sebesar

79%. Namun, saat jumlah data latih dikurangi menjadi 70%, akurasi *Polynomial* kernel hanya turun sedikit menjadi 78,32%, yang masih lebih tinggi dibandingkan Linear kernel pada proporsi yang sama. Meskipun presisi, *recall*, dan *f1-score* sedikit menurun menjadi 77%, 76%, dan 76%, stabilitas performa *Polynomial* kernel tetap terlihat.

Secara keseluruhan, baik *Linear* maupun *Polynomial* kernel memberikan hasil yang sangat baik pada dataset ini, dengan performa yang hampir setara saat data dilatih dengan porsi 80%. Namun, *Polynomial* kernel cenderung lebih stabil pada pembagian data 70:30, sehingga dapat dianggap sedikit lebih unggul dalam hal ketahanan terhadap perubahan proporsi data latih.

4.5.3 Hasil Uji Coba Dataset Gabung (2024 dan 2025)

Dalam uji coba dataset Gabungan berjumlah 1574 data yang mempunyai 593 data berlabel positif dan 981 data berlabel negatif. Data tersebut kemudian dilakukan peyamarataan data dengan menggunakan *random undersampling* sehingga data menjadi 593 data positif dan 593 data negatif. Berikut adalah hasil *confusion matrix* yang dapat dilihat dalam Tabel 4.15.

Tabel 4.15 Hasil *Confusion Matrix* Dataset Gabungan

Kernel	Rasio Data	Kelas Sebenarnya	Kelas Prediksi	
			Positif	Negatif
Linear	8:2	Positif	91	28
		Negatif	22	97
	7:3	Positif	122	56

		Negatif	43	136
Polynomial	8:2	Positif	89	30
		Negatif	26	93
	7:3	Positif	124	54
		Negatif	37	142

Berdasarkan tabel di atas maka didapatkan hasil akurasi, presisi, *recall*, dan *f-measure* yang dapat dilihat dalam Tabel 4.16.

Tabel 4.16 Hasil Akurasi, Presisi, *Recall*, dan *F-measure* Dataset Gabungan

Kernel	Variasi Data		Nilai Performa			
	Data Latih	Data Uji	Akurasi	Presisi	Recall	F-measure
Linear	80%	20%	0.7857	0.79	0.79	0.79
	70%	30%	0.7227	0.72	0.72	0.72
Polynomial	80%	20%	0.7899	0.79	0.79	0.79
	70%	30%	0.7451	0.75	0.75	0.74

Pada kernel *Linear*, performa model cukup baik ketika data dibagi dengan rasio 80:20. Akurasi yang diperoleh adalah sebesar 78,57%, dengan presisi, *recall*, dan *f1-score* masing-masing sebesar 79%. Namun, ketika proporsi data latih dikurangi menjadi 70%, performa mengalami penurunan yang cukup signifikan, dengan akurasi turun menjadi 72,27%. Hal ini mengindikasikan bahwa kernel *Linear* masih cukup sensitif terhadap pengurangan data latih pada dataset gabungan yang lebih besar ini.

Sementara itu, kernel Polynomial menunjukkan performa yang sedikit lebih baik secara keseluruhan. Pada rasio data 80:20, diperoleh akurasi sebesar

78,99%, yang sedikit lebih tinggi dibanding *Linear* kernel, dengan presisi, *recall*, dan *f1-score* tetap stabil di angka 79%. Saat jumlah data latih dikurangi menjadi 70%, performa *Polynomial* kernel tetap lebih baik dibandingkan *Linear* kernel dengan akurasi 74,51%, presisi dan *recall* 75%, serta *f1-score* 74%. Penurunan performa memang terjadi, tetapi tidak sedrastis *Linear* kernel, menandakan bahwa *Polynomial* kernel lebih stabil terhadap pengurangan jumlah data latih, bahkan ketika ukuran dataset meningkat.

Secara keseluruhan, hasil ini menunjukkan bahwa penambahan jumlah data dari penggabungan dua dataset sebelumnya mampu meningkatkan kestabilan performa model, terutama pada kernel *Polynomial*. Performa kedua kernel terlihat semakin berbeda pada rasio 80:20, tetapi *Polynomial* kernel tetap mempertahankan keunggulannya dalam menghadapi perubahan proporsi data latih.

4.8 Pembahasan

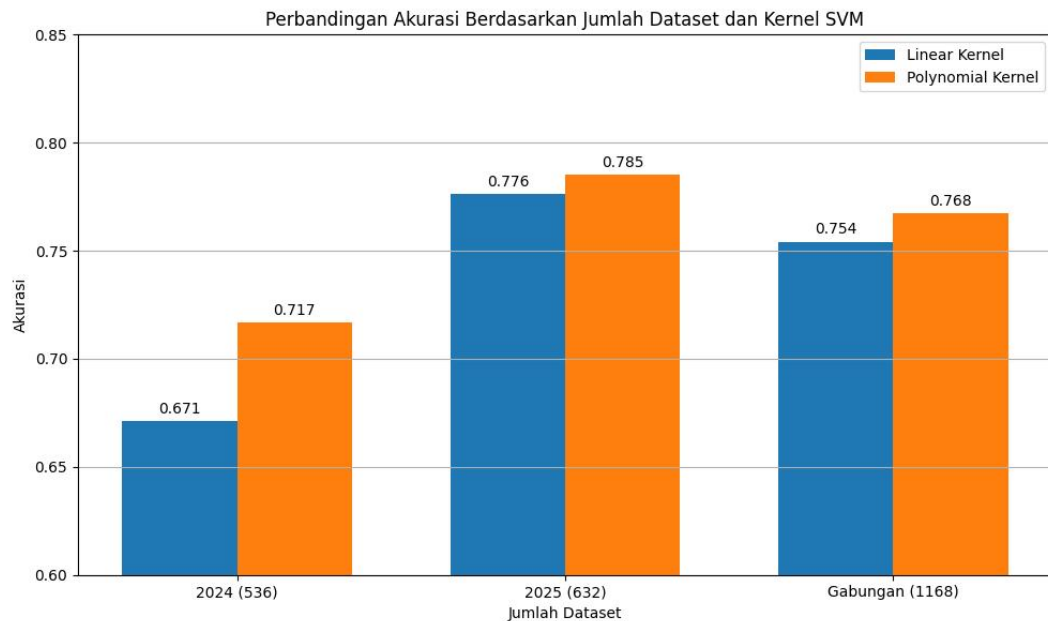
Hasil uji coba yang telah dilakukan melalui skenario 3 dataset yang berbeda dari segi jumlah dan waktu pengambilan, serta menggunakan 2 rasio data uji dan data latih yang berbeda dan 2 jenis kernel yang berbeda yaitu *Linear* dan *Polynomial*. Skenario pengujian menggunakan 6 skenario yang hasilnya dapat dilihat pada Tabel 4.17.

Tabel 4.17 Hasil Akurasi, Presisi, *Recall*, dan *F-measure* Semua Dataset

Kernel	Data	Variasi Data		Nilai Performa			
		Data Latih	Data Uji	Akurasi	Presisi	Recall	F-measure

Linear	2024	80%	20%	0.722	0.73	0.72	0.72
		70%	30%	0.62	0.63	0.63	0.63
	2025	80%	20%	0.7874	0.79	0.79	0.79
		70%	30%	0.7648	0.77	0.77	0.77
	Gabung	80%	20%	0.7857	0.79	0.79	0.79
		70%	30%	0.7227	0.72	0.72	0.72
Polynomial	2024	80%	20%	0.7315	0.74	0.73	0.73
		70%	30%	0.7019	0.70	0.70	0.70
	2025	80%	20%	0.7874	0.79	0.79	0.79
		70%	30%	0.7832	0.77	0.76	0.76
	Gabung	80%	20%	0.7899	0.79	0.79	0.79
		70%	30%	0.7451	0.75	0.75	0.74

Dari tabel di atas dapat kita lihat pada dataset 2024, kernel *Linear* menunjukkan performa paling rendah dengan akurasi hanya 62% saat data latih 70%. Sedangkan performa tertinggi ditunjukkan pada dataset 2025, kernel *Polynomial* dengan akurasi 78,99% pada data latih 80%. Berikut adalah grafik perbandingan jumlah data latih terhadap hasil akurasi yang digambarkan dalam Gambar 4.9.



Gambar 4.9 Perbandingan Akurasi dengan Jumlah Dataset dan Kernel SVM

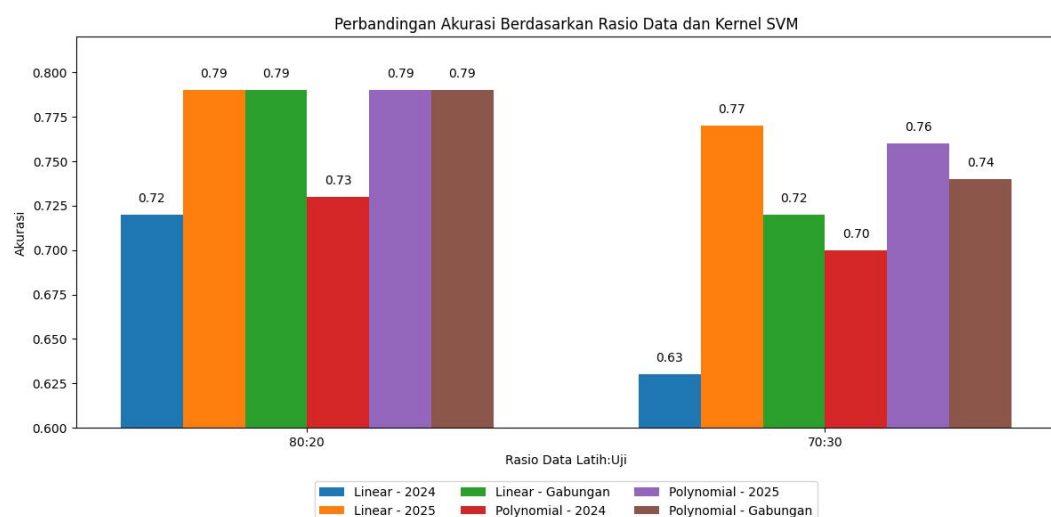
Berdasarkan hasil visualisasi diagram batang, dapat disimpulkan bahwa semakin banyak jumlah data yang digunakan dalam analisis sentimen, umumnya akurasi model meningkat. Hal ini terlihat dari perbandingan dataset tahun 2024 (536 data), 2025 (632 data), dan gabungan keduanya (1168 data). Akurasi model pada dataset tahun 2025 meningkat secara signifikan dibandingkan tahun 2024 dengan peningkatan sekitar 9%, menunjukkan bahwa jumlah data berpengaruh terhadap performa model. Namun, meskipun dataset gabungan memiliki jumlah data paling banyak, akurasinya tidak lebih tinggi dibandingkan dataset tahun 2025 dengan penurunan sekitar 2%, yang mengindikasikan bahwa peningkatan jumlah data tidak selalu menjamin peningkatan akurasi jika tidak diikuti oleh keseragaman atau kualitas data yang baik. Akurasi dataset 2024 mendapat akurasi yang paling kecil disebabkan oleh kualitas data yang kurang baik, yang dapat dilihat pada Tabel 3.4 frekuensi kemunculan kata dataset 2024. Pada frekuensi

kemunculan kata dataset 2024 banyak kata yang muncul pada dua komentar positif dan negatif, hal ini akan membuat banyak noise yang menjadikan kualitas data menjadi kurang baik.

Dilihat dari perbandingan dataset berdasarkan tahun, model yang dilatih dengan data tahun 2025 menghasilkan akurasi tertinggi dibandingkan data tahun 2024 maupun dataset gabungan. Hal ini dapat disebabkan oleh kualitas data tahun 2025 yang lebih baik, lebih relevan, atau lebih konsisten secara distribusi. Kualitas data tersebut dapat dilihat pada frekuensi kemunculan kata dataset 2025 pada Tabel 3.5 yang menunjukkan lebih sedikit kata yang muncul pada dua label dibandingkan dengan dataset 2024. Sementara itu, gabungan data dari dua tahun justru berpotensi menambah keragaman dalam gaya bahasa, topik, atau konteks, yang bisa memperbesar noise dan mengurangi ketepatan model dalam mengenali pola sentimen secara konsisten. Perbedaan tersebut dapat dilihat pada tabel frekuensi kemunculan kata yang menunjukkan banyak kata yang semakin beragam. Dengan demikian, penggabungan data dari berbagai sumber atau waktu perlu diimbangi dengan proses normalisasi dan penyelarasan karakteristik data agar model tetap optimal.

Dari sisi pemilihan kernel, hasil menunjukkan bahwa kernel Polynomial secara konsisten memberikan akurasi yang lebih tinggi dibandingkan kernel Linear, baik pada dataset 2024, 2025, maupun dataset gabungan dengan total penurunan mencapai 2%. Perbedaan paling mencolok terjadi pada dataset yang lebih kecil (2024) dengan perbedaan sampai 4%, di mana kernel *Polynomial*

menunjukkan keunggulan dalam menangkap pola-pola *non-linear* dalam data. Namun, perbedaan performa antara kedua kernel mulai mengecil dari 4% menjadi 1% seiring dengan bertambahnya jumlah data, yang mengindikasikan bahwa kernel *Linear* menjadi lebih kompetitif saat data cukup besar dan distribusinya stabil. Oleh karena itu, pemilihan kernel harus mempertimbangkan karakteristik data, baik dari segi jumlah maupun kompleksitas pola yang ingin ditangkap oleh model. Selain itu, penelitian ini juga terdapat perbandingan dengan rasio data perbandingannya pada Gambar 4.10.



Gambar 4.10 Perbandingan Akurasi dengan Rasio Data

Dapat dilihat pada diagram diatas rasio 80:20 menghasilkan akurasi yang lebih tinggi dengan rata-rata 0.768 dibandingkan dengan rasio 70:30 dengan rata-rata 0.72, baik pada kernel linear maupun polynomial. Hal ini tampak dari nilai akurasi yang lebih tinggi di semua kategori dataset (2024, 2025, dan gabungan) saat menggunakan rasio 80:20. Misalnya, pada dataset tahun 2024 dengan kernel *Linear*, akurasi pada rasio 80:20 adalah 0.72, sedangkan pada rasio 70:30 turun

menjadi 0.63. Hal yang sama terjadi pada kernel *Polynomial* di dataset gabungan, di mana akurasi pada rasio 80:20 mencapai 0.79 dan menurun menjadi 0.74 saat rasio berubah menjadi 70:30.

Pada rasio 80:20, jumlah data latih yang lebih besar memberikan model lebih banyak informasi untuk belajar, sehingga menghasilkan prediksi yang lebih akurat. Sebaliknya, rasio 70:30 memberikan lebih banyak data untuk pengujian, tetapi mengurangi kapasitas pembelajaran model karena data latih lebih sedikit. Oleh karena itu, rasio 80:20 cenderung lebih optimal dalam konteks dataset dan algoritma ini, khususnya untuk meningkatkan akurasi klasifikasi pada model SVM baik dengan kernel linear maupun *Polynomial*.

Proses klasifikasi atau pengelompokan komentar ke dalam kategori positif dan negatif, melalui sistem otomatis, selalu berlandaskan pada perhitungan logis dan aturan yang terstruktur. Setiap keputusan klasifikasi bukanlah hasil acak, melainkan berdasarkan pola dan ukuran yang telah ditentukan. Menariknya, prinsip ini memiliki kemiripan dengan pandangan dalam Islam, di mana segala sesuatu yang diciptakan Allah tidak pernah lepas dari aturan dan ukuran yang pasti. Seperti disebutkan dalam Al-Qur'an, seluruh ciptaan Allah berjalan dalam harmoni berdasarkan takaran yang telah ditetapkan. Ini menunjukkan bahwa konsep keteraturan, baik dalam dunia teknologi maupun spiritual, merupakan fondasi penting dalam memahami dan mengelola kehidupan. Hal ini berdasarkan Al-Qur'an surat Al-Furqan Ayat 2:

الَّذِي لَهُ مُلْكُ السَّمَاوَاتِ وَالْأَرْضِ وَلَمْ يَتَّخِذْ وَلَدًا وَلَمْ يَكُنْ لَهُ شَرِيكٌ فِي الْمُلْكِ وَخَلَقَ كُلَّ شَيْءٍ
فَقَدَرَهُ تَقْدِيرًا

“yang kepunyaan-Nya-lah kerajaan langit dan bumi, dan Dia tidak mempunyai anak, dan tidak ada sekutu bagi-Nya dalam kekuasaan(Nya), dan dia telah menciptakan segala sesuatu, dan Dia menetapkan ukuran-ukurannya dengan serapi-rapinya”. (Q.S Al-Furqon:2)

Menurut tafsir Ibnu Katsir, dalam ayat ini Allah *Subhanahu Wa Ta'ala* membersihkan diri-Nya dari beranak dan sekutu. Kemudian dalam firman berikutnya disebutkan: Dia menciptakan segala sesuatu, dan Dia menetapkan ukuran-ukurannya dengan serapi-rapinya. (Al Furqaan:2) Yakni segala sesuatu selain Dia adalah makhluk lagi dimiliki, sedangkan Dialah Yang Menciptakan segala sesuatu, Yang Menguasai, Yang Memiliki dan Tuhannya, segala sesuatu berada di bawah kekuasaan-Nya, diatur oleh-Nya, tunduk kepada-Nya dan kepada takdir-Nya.

Pada data yang akan diuji pada penelitian ini, masih ditemukan banyak masyarakat yang kurang baik dalam menyampaikan pendapatnya. Terdapat penggunaan bahasa yang kasar serta pernyataan yang terkesan menyudutkan pihak tertentu. Fenomena ini tidak selaras dengan jati diri negara Indonesia, yang kebanyakan penduduknya memeluk agama Islam. Padahal, dalam ajaran Islam, Allah telah menyuruh untuk mengutarakan pendapat dengan tutur kata yang santun dan penuh etika. Hal ini berdasarkan Al-Qur'an surat An-Nisa ayat 9:

وَلْيَخْشَ الَّذِينَ لَوْ تَرَكَوا مِنْ خَلْفِهِمْ ذُرِّيَّةً ضِعَافًا خَافُوا عَلَيْهِمْ فَلْيَتَّقُوا اللَّهَ وَلْيَقُولُوا قَوْلًا سَدِيدًا

“Dan hendaklah takut (kepada Allah) orang-orang yang sekiranya mereka meninggalkan keturunan yang lemah di belakang mereka, yang mereka merasa khawatir terhadap (kesejahteraan)nya. Oleh karena itu, hendaklah mereka bertakwa kepada Allah, dan hendaklah mereka berbicara dengan tutur kata yang benar”. (Q.S An-Nisa: 9)

Kandungan ayat di atas menurut tafsir Kementrian Agama menjelaskan untuk berbagi sebagian dari harta warisan yang didapat kepada kerabat yang tidak mendapatkan bagian, ayat ini memberi anjuran untuk memperhatikan nasib anak-anak mereka apabila menjadi yatim. Dan hendaklah takut kepada Allah orang-orang yang sekiranya mereka meninggalkan keturunan di kemudian hari anak-anak yang lemah dalam keadaan yatim yang belum mampu mandiri di belakang mereka yang mereka khawatir terhadap kesejahteraan-nya lantaran mereka tidak terurus, lemah, dan hidup dalam kemiskinan. Oleh sebab itu, hendaklah mereka para wali bertakwa kepada Allah dengan mengindahkan perintah-Nya dan menjauhi larangan-Nya, dan hendaklah mereka berbicara dengan tutur kata yang benar, penuh perhatian dan kasih sayang terhadap anak-anak yatim dalam asuhannya.

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Pada penelitian ini hasil klasifikasi pada dataset 2024 (536 data) adalah yang terburuk dengan akurasi rata-rata 70 %, setelah itu dataset gabungan dengan rata-rata 76 %, dan yang terbaik dataset 2025 dengan rata-rata akurasi 78%. Sedangkan untuk perbandingan kernel hasil klasifikasi pada kernel *Polynomial* unggul dari hampir semua dataset dan variasi rasio data latih dan uji. Perbedaan paling signifikan terjadi pada dataset 2024 rasio 70 : 30 dengan akurasi pada kernel *Linear* 62% menjadi 70% pada kernel *Polynomial*. Sedangkan untuk perbandingan rasio data uji dan data latih hasil klasifikasi pada rasio 80 : 20 unggul pada semua skenario pengujian pada rasio 70: 30.

Oleh karena itu, dari penelitian ini dapat disimpulkan bahwa hasil klasifikasi pada analisis sentimen sangat bergantung pada kualitas data yang baik, serta banyaknya data tidak selalu menjamin akurasi yang tinggi, terutama jika datanya beragam dari segi waktu dan konteks berita. Penggabungan dapat berpotensi menambah keragaman dalam topik, konteks, dan kata, yang bisa memperbesar *noise*. Kernel *Linear* jauh lebih dapat bersaing dengan kernel *Polynomial* apabila data latih lebih banyak dengan memperhatikan karakteristik data yang diuji. Perbedaan rasio data uji dan data latih pada rasio 8: 2 umumnya

akan jauh lebih baik dibandingkan 7: 3 dikarenakan jumlah data latih yang lebih banyak dapat memberikan sistem lebih banyak informasi untuk dipelajari, sehingga menghasilkan akurasi yang lebih baik.

5.2 Saran

Berdasarkan hasil uji coba yang telah dilakukan, penulis menyadari bahwa penelitian ini masih memiliki ruang untuk penyempurnaan guna mengoptimalkan performa sistem. Oleh karena itu, penulis memberikan beberapa saran sebagai langkah strategis untuk pengembangan dan penyempurnaan penelitian di masa mendatang:

1. Lakukan dengan lebih banyak data sentimen agar model dapat agar perbedaan trend, konteks berita, dan gaya bahasa tidak mempersulit model untuk melakukan klasifikasi.
2. Apabila ingin menggabungkan data antar tahun, pertimbangkan menambahkan fitur "tahun" sebagai bagian dari input agar model mengetahui dari konteks waktu mana data itu berasal.
3. Pertimbangkan menggunakan model lain seperti *Naive bayes* dan *Random Forest*.
4. Coba untuk menghapus kata makan, siang, dan gratis karena ketiga kata tidak memberikan nilai pada label positif dan negatif.

DAFTAR PUSTAKA

- Eliza, F., Gistituati, N., Rusdinal, R., & Fadli, R. (2024). Analisis SWOT Kebijakan Makan Siang Gratis di Sekolah Menengah Kejuruan. *Juwara Jurnal Wawasan dan Aksara*, 4(1), 121–129. <https://doi.org/10.58740/juwara.v4i1.91>.
- Fanny, I. S., Nadia, R., & Alisa, A. (2024). Dampak Makan Siang Gratis Pada Kondisi Keuangan Negara Dan Peningkatan Mutu Pendidikan. *Jurnal Sistem Informasi, Teknologi Informasi dan Komputer*, 1, 192–196. <https://journalwbl.com/index.php/jupensal/article/view/176/42>.
- Fasha, S. S., & Tesniyadi, D. (2024). Analisis Wacana Kritis Pada Artikel Tempo.co Yang Berjudul "Dana BOS Untuk Program Makan Siang Gratis". *INNOVATIVE: Journal Of Social Science Research*, 4, 15077–15089. <https://doi.org/https://doi.org/10.31004/innovative.v4i3.12362>.
- Giovani, A. P., Ardiansyah, A., Haryanti, T., Kurniawati, L., & Gata, W. (2020). ANALISIS SENTIMEN APLIKASI RUANG GURU DI TWITTER MENGGUNAKAN ALGORITMA KLASIFIKASI. *Jurnal Teknoinfo*, 14(2), 115. <https://doi.org/10.33365/jti.v14i2.679>.
- Ipmawati, J., Saifulloh, S., & Kusnawi, K. (2024). Analisis Sentimen Tempat Wisata Berdasarkan Ulasan pada Google Maps Menggunakan Algoritma Support Vector Machine. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, 4(1), 247–256. <https://doi.org/10.57152/malcom.v4i1.1066>.
- Kaharudin, A., Agus Supriyadi, A., Baitika, H., & Derryanur, M. (2023). OKTAL : Jurnal Ilmu Komputer dan Science Analisis Sentimen pada Media Sosial dengan Teknik Kecerdasan Buatan Naïve Bayes: Kajian Literatur Review. 2(6). <https://harzing.com/resources/publish-or-perish>.
- Karimah, A., & Dwilestari, G. (2024). ANALISIS SENTIMEN KOMENTAR VIDEO MOBIL LISTRIK DI PLATFORM YOUTUBE DENGAN METODE NAIVE BAYES. Dalam *Jurnal Mahasiswa Teknik Informatika* (Vol. 8, Nomor 1). <https://www.kaggle.com/datasets/billycemerson/analisis>.
- Khatib Sulaiman, J., Baehaqi, F., Cahyono, N., & Amikom Yogyakarta, U. (t.t.). Analisis Sentimen Terhadap Cyberbullying Pada Komentar di Instagram Menggunakan Algoritma Naïve Bayes. *Indonesian Journal of Computer Science*.

- Khusnul, A., Manajemen, K., Keimigrasian, T., & Imigrasi, P. (2024). ANALISIS SENTIMEN TERHADAP KUALITAS PELAYANAN (TINJAUAN LITERATUR). Dalam Jurnal Mahasiswa Teknik Informatika (Vol. 8, Nomor 3).
- Lailita, N. A., & Khoirunnisa, N. V. (2024). Sosial Media pada Pendidikan Indonesia: Pengaruh Inovasi Media Sosial Terhadap Kualitas Pelajar di Indonesia. *Indo-MathEdu Intellectuals Journal*, 5(3), 3136–3143. <https://doi.org/10.54373/imeij.v5i3.1139>.
- Lubis, A. Y., & Setyawan, M. Y. H. (2024). Analisis Sentimen Terhadap Aplikasi Pospay Menggunakan Algoritma Support Vector Machine dan Naive Bayes. *Jurnal Teknologi Dan Sistem Informasi Bisnis*, 6(3), 514–521. <https://doi.org/10.47233/jteksis.v6i3.1310>.
- Lubis, K. A., Theo, M., Bangsa, A., & Yudertha, A. (2024). ANALISIS SENTIMEN OPINI MASYARAKAT TERHADAP PINDAHNYA IBU KOTA INDONESIA DENGAN MENGGUNAKAN KLASIFIKASI NAÏVE BAYES (Vol. 18, Nomor 1). <https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index>.
- Nufairi, F., Pratiwi, N., & Herlando, F. (2024). ANALISIS SENTIMEN PADA ULASAN APLIKASI THREADS DI GOOGLE PLAY STORE MENGGUNAKAN ALGORITMA SUPPORT VECTOR MACHINE. *JIPi (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, 9(1), 339–348. <https://doi.org/10.29100/jipi.v9i1.4929>.
- Nurian, A., Ma'arif, M. S., Amalia, I. N., & Rozikin, C. (2024). ANALISIS SENTIMEN PENGGUNA APLIKASI SHOPEE PADA SITUS GOOGLE PLAY MENGGUNAKAN NAIVE BAYES CLASSIFIER. *Jurnal Informatika dan Teknik Elektro Terapan*, 12(1). <https://doi.org/10.23960/jitet.v12i1.3631>.
- Pamungkas, A. S., & Cahyono, N. (2024). Edumatic: Jurnal Pendidikan Informatika Analisis Sentimen Review ChatGPT di Play Store menggunakan Support Vector Machine dan K-Nearest Neighbor. 8(1), 1–10. <https://doi.org/10.29408/edumatic.v8i1.24114>.
- Putra, D. A., Nugroho, A. S., & Maulana, H. R. (2023). Penerapan Support Vector Machine untuk Analisis Sentimen Ulasan Produk Berbahasa Indonesia. *Jurnal Teknologi Informasi dan Komputer*, 9(2), 112–118.
- Putri Husain, N., Febriana Syam, A., & Mustikosari, R. (2024). Analisis Sentimen Ulasan Pengguna Tiktok pada Google Play Store Berbasis TF-IDF dan

Support Vector Machine. Dalam Journal of System and Computer Engineering (JSCE) ISSN (Vol. 5, Nomor 1). <https://images.app.goo.gl/hC6494uW637VmYVW9>.

Rahmadani, N., & Yusuf, M. (2024). Analisis Sentimen Komentar Pengguna Media Sosial Menggunakan SVM dan TF-IDF. Jurnal Informatika dan Sistem Informasi, 11(1), 45–53..

Ramadhani, B., Suryono, R. R., & Kunci, K. (2024). JURNAL MEDIA INFORMATIKA BUDIDARMA Komparasi Algoritma Naïve Bayes dan Logistic Regression Untuk Analisis Sentimen Metaverse. <https://doi.org/10.30865/mib.v8i2.7458>.

Riyadiiban, S., & Riyadi, S. (t.t.). Analisis Sentimen Opini Masyarakat Terhadap Stadion Jakarta Internasional Stadium (Jis) Pada Twitter ... Analisis Sentimen Opini Masyarakat Terhadap Stadion Jakarta Internasional Stadium (Jis) Pada Twitter Dengan Perbandingan Metode Naive Bayes Dan Support Vector Machine. Jurnal Sains dan Teknologi, 5(3), 2024. <https://doi.org/10.55338/saintek.v5i3.2962>.

Setiawan, A., & Suryono, R. R. (2024). Edumatic: Jurnal Pendidikan Informatika Analisis Sentimen Ibu Kota Nusantara menggunakan Algoritma Support Vector Machine dan Naïve Bayes. 8(1), 183–192. <https://doi.org/10.29408/edumatic.v8i1.25667>.

Viriya, A. A., Sartika, I., Maghfiroh, E., & Setiawan, N. Y. (2024). Analisis Sentimen Ulasan Pengguna Aplikasi Mobile Gapura UB pada Google Play Store Menggunakan Algoritma Support Vector Machine (Vol. 1, Nomor 1). <http://j-ptiik.ub.ac.id>.

Zaman, F. N., Fadhilah, M. A., Ulinuha, M. A., & Umam, K. (2024). MENGANALISIS RESPON NETIZEN TWITTER TERHADAP PROGRAM MAKAN SIANG GRATIS MENERAPKAN NLP METODE NAÏVE BAYES (Vol. 14, Nomor 3). <https://jurnal.umj.ac.id/index.php/just-it/index>.

Zena Lusi, Ayu Eka Saputri, & Tri Basuki Kurniawan. (2024). Identifikasi Komentar Spam Pada Sosial Media. Neptunus: Jurnal Ilmu Komputer Dan Teknologi Informasi, 2(2), 71–76. <https://doi.org/10.61132/neptunus.v2i2.100>.

LAMPIRAN-LAMPIRAN