Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 259–273
DOI : 10.47065/bulletinds.v5i3.10323
https://ejurnal.seminar-id.com/index.php/bulletinds

Analisis Sentimen Multikelas Program Makan Bergizi Gratis di Media Sosial X Menggunakan VaderSentiment dan SVM Berbasis Oversampling SMOTE

Mohamad Hegar Sukmana Wibowo*, Muhammad Fatchan, Asep Supriyanto

Fakultas Teknik, Teknik Informatika, Universitas Pelita Bangsa, Bekasi, Indonesia
Email: 1,*hegarsw@gmail.com, 2fatchan@pelitabangsa.ac.id, 3asep.supriyanto@pelitabangsa.ac.id
Email Penulis Korespondensi: hegarsw@gmail.com

Abstrak−Media sosial X (Twitter) menjadi ruang diskusi publik yang menghasilkan data opini masyarakat secara real-time terhadap berbagai kebijakan pemerintah, termasuk Program Makan Bergizi Gratis (MBG). Analisis sentimen terhadap kebijakan publik umumnya masih terbatas pada klasifikasi biner serta belum optimal dalam menangani ketidakseimbangan distribusi data. Penelitian ini bertujuan membangun model analisis sentimen multikelas terhadap opini masyarakat mengenai Program MBG dengan mengintegrasikan VaderSentiment sebagai metode ekstraksi fitur, Support Vector Machine (SVM) sebagai algoritma klasifikasi, dan Synthetic Minority Oversampling Technique (SMOTE) untuk mengatasi imbalanced dataset. Penelitian menggunakan pendekatan kuantitatif dengan metode eksperimental. Data berupa tweet berbahasa Indonesia yang relevan dengan MBG dikumpulkan melalui teknik crawling. Tahapan pengolahan data meliputi preprocessing seperti cleaning, case folding, tokenizing, dan stopword removal. Skor sentimen yang dihasilkan oleh VaderSentiment digunakan sebagai fitur numerik dalam proses klasifikasi SVM. Selanjutnya, teknik SMOTE diterapkan pada data latih untuk menyeimbangkan distribusi kelas sentimen positif, netral, dan negatif. Evaluasi model dilakukan menggunakan confusion matrix serta metrik akurasi, precision, recall, dan F1-score. Hasil penelitian menunjukkan bahwa penerapan kombinasi VaderSentiment, SVM, dan SMOTE mampu meningkatkan kemampuan model dalam mengenali kelas sentimen minoritas dibandingkan kondisi sebelum penyeimbangan data, meskipun terjadi penurunan pada nilai akurasi agregat. Temuan ini menunjukkan bahwa evaluasi performa model klasifikasi sentimen tidak cukup hanya menggunakan metrik accuracy, tetapi juga perlu mempertimbangkan precision, recall, dan F1-score untuk memperoleh interpretasi yang lebih komprehensif terhadap performa klasifikasi multikelas. Secara metodologis, penelitian ini berkontribusi pada pengembangan analisis sentimen berbasis machine learning, serta secara praktis mendukung evaluasi kebijakan publik berbasis data melalui pemanfaatan opini masyarakat di media sosial.

Kata Kunci: Analisis Sentimen; Multikelas; VaderSentiment; SVM; SMOTE; MBG; Media Sosial X

Abstract−Social media platform X (Twitter) has become a public discussion space that generates real-time public opinion data regarding various government policies, including the Free Nutritious Meal Program (MBG). Sentiment analysis of public policies is generally still limited to binary classification and has not optimally addressed the issue of imbalanced data distribution. This study aims to develop a multiclass sentiment analysis model to examine public opinion regarding the MBG program by integrating VaderSentiment as a feature extraction method, Support Vector Machine (SVM) as the classification algorithm, and the Synthetic Minority Oversampling Technique (SMOTE) to handle imbalanced datasets. This research employed a quantitative approach using an experimental method. The data consisted of Indonesian-language tweets related to the MBG program collected through a crawling technique. Data processing stages included preprocessing steps such as cleaning, case folding, tokenizing, and stopword removal. Sentiment scores generated by VaderSentiment were used as numerical features in the SVM classification process. Subsequently, SMOTE was applied to the training data to balance the distribution of positive, neutral, and negative sentiment classes. Model performance was evaluated using a confusion matrix and metrics including accuracy, precision, recall, and F1-score. The results showed that the combination of VaderSentiment, SVM, and SMOTE improved the model's ability to recognize minority sentiment classes compared with conditions prior to data balancing, although a decrease in overall accuracy was observed. These findings indicate that sentiment classification performance should not be evaluated solely based on accuracy but should also consider precision, recall, and F1-score to obtain a more comprehensive assessment of multiclass classification performance. Methodologically, this study contributes to the advancement of machine learning–based sentiment analysis and, practically, supports data-driven public policy evaluation through the utilization of public opinion on social media.

Keywords: Sentiment Analysis; Multiclass; VaderSentiment; SVM; SMOTE; MBG; Social Media X

1. PENDAHULUAN

Media sosial telah berkembang menjadi ruang utama pembentukan dan ekspresi opini publik pada era digital. Platform X (sebelumnya Twitter) tidak lagi berfungsi hanya sebagai media interaksi personal, tetapi juga menjadi sumber data yang relevan untuk memahami respons masyarakat terhadap isu sosial dan kebijakan pemerintah secara real time. Pemanfaatan data media sosial melalui pendekatan analisis sentimen memungkinkan identifikasi pola persepsi publik secara lebih sistematis dan terukur, sehingga semakin banyak digunakan dalam penelitian berbasis text mining dan Natural Language Processing (NLP) [1],[2],[3]. Tingginya volume data yang dihasilkan pengguna serta karakteristik teks yang singkat dan dinamis menjadikan X sebagai sumber informasi yang representatif untuk mengevaluasi perubahan opini masyarakat terhadap suatu kebijakan [2]. Dalam konteks Indonesia, fenomena tersebut menjadi semakin relevan pada pembahasan kebijakan Program Makan Bergizi Gratis (MBG) yang memperoleh perhatian luas di ruang digital [4].

Program Makan Bergizi Gratis (MBG) sebagai salah satu kebijakan strategis pemerintah memunculkan spektrum respons publik yang beragam, mulai dari dukungan, kritik, hingga sikap netral. Namun demikian, sebagian besar pendekatan analisis sentimen pada penelitian terdahulu masih didominasi oleh klasifikasi biner yang hanya membedakan sentimen positif dan negatif. Pendekatan tersebut memiliki keterbatasan dalam merepresentasikan kompleksitas opini masyarakat yang sering kali tidak berada pada posisi ekstrem. Selain itu, penelitian analisis sentimen berbasis media

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 259–273
DOI : 10.47065/bulletinds.v5i3.10323
https://ejurnal.seminar-id.com/index.php/bulletinds

sosial juga menghadapi persoalan ketidakseimbangan distribusi data (imbalanced dataset) yang berpotensi menurunkan kemampuan model dalam mengenali kelas minoritas secara akurat. Beberapa penelitian menunjukkan bahwa penerapan teknik oversampling seperti Synthetic Minority Oversampling Technique (SMOTE) mampu meningkatkan performa klasifikasi pada kondisi data yang tidak seimbang [5],[6], termasuk ketika dikombinasikan dengan algoritma Support Vector Machine (SVM) pada studi sentimen kebijakan publik [7].

Di sisi lain, perkembangan metode analisis sentimen berbasis leksikon menunjukkan bahwa VADER (Valence Aware Dictionary and Sentiment Reasoner) memiliki kemampuan yang baik dalam mengakomodasi karakteristik bahasa informal yang umum ditemukan pada media sosial [8]. Meskipun demikian, kajian terdahulu masih cenderung menggunakan pendekatan secara terpisah antara ekstraksi sentimen, klasifikasi, dan penanganan ketidakseimbangan data. Penelitian yang secara khusus mengintegrasikan VADER sebagai metode ekstraksi sentimen, SVM sebagai algoritma klasifikasi multikelas, serta SMOTE sebagai teknik penyeimbangan data pada konteks kebijakan sosial seperti MBG masih relatif terbatas. Kondisi tersebut menunjukkan adanya kesenjangan penelitian yang perlu diisi untuk menghasilkan pendekatan analitis yang lebih komprehensif dalam memahami persepsi publik [9],[10],[11].

Berdasarkan kondisi tersebut, penelitian ini menjadi penting baik secara ilmiah maupun praktis. Secara ilmiah, pengembangan pendekatan analisis sentimen multikelas berpotensi memperluas kontribusi metodologis dalam bidang NLP dan machine learning, khususnya pada pengolahan data opini publik yang tidak seimbang. Secara praktis, pemahaman yang lebih akurat terhadap persepsi masyarakat dapat dimanfaatkan sebagai dasar evaluasi kebijakan berbasis data serta mendukung perumusan strategi komunikasi publik yang lebih responsif. Oleh karena itu, penelitian ini diarahkan untuk membangun model analisis sentimen multikelas terhadap opini masyarakat mengenai Program Makan Bergizi Gratis di media sosial X dengan memanfaatkan kombinasi metode VADER, SVM, dan SMOTE.

Artikel ini berkontribusi melalui pengembangan kerangka analisis yang mengintegrasikan ekstraksi sentimen berbasis leksikon, klasifikasi machine learning, dan penanganan ketidakseimbangan data dalam satu alur metodologis yang terpadu. Pendekatan tersebut diharapkan mampu menghasilkan representasi sentimen yang lebih proporsional melalui tiga kategori sentimen, yaitu positif, netral, dan negatif. Selain memberikan kontribusi empiris pada kajian analisis sentimen kebijakan publik di Indonesia, penelitian ini juga menawarkan alternatif metodologis yang dapat digunakan pada studi sejenis dengan karakteristik data media sosial yang kompleks dan tidak seimbang.

2. METODOLOGI PENELITIAN

2.1 Jenis dan Pendekatan Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan desain penelitian eksplanatori dan metode eksperimen berbasis klasifikasi data. Pendekatan kuantitatif dipilih karena penelitian berfokus pada pengukuran variabel secara objektif melalui representasi numerik, pemodelan komputasional, dan evaluasi statistik terhadap performa model klasifikasi [12],[13],[14],[15]. Dalam penelitian ini, proses analisis tidak diarahkan hanya untuk menggambarkan fenomena opini publik, tetapi juga untuk menguji pengaruh penerapan kombinasi metode ekstraksi sentimen, penyeimbangan data, dan algoritma klasifikasi terhadap hasil prediksi sentimen multikelas. Pendekatan eksplanatori digunakan karena penelitian berupaya menjelaskan hubungan sebab–akibat antara variabel independen berupa fitur sentimen hasil ekstraksi dan teknik penanganan ketidakseimbangan data terhadap variabel dependen berupa hasil klasifikasi sentimen [7]. Untuk memenuhi tujuan tersebut, metode eksperimen dipilih karena memungkinkan pengendalian tahapan pemrosesan data, konfigurasi model, serta evaluasi hasil secara sistematis sehingga pengaruh setiap komponen terhadap performa model dapat diamati secara terukur [16],[17].

2.2 Pengumpulan Data

Penelitian dilaksanakan menggunakan sumber data sekunder yang berasal dari platform media sosial X (Twitter). Data dikumpulkan melalui proses crawling terhadap unggahan publik yang relevan dengan Program Makan Bergizi Gratis (MBG). Pengumpulan data dilakukan dalam rentang waktu Desember 2024 hingga September 2025 karena periode tersebut merepresentasikan fase meningkatnya perhatian publik terhadap implementasi dan diskursus kebijakan MBG di ruang digital. Penggunaan media sosial sebagai lokasi observasi didasarkan pada karakteristiknya yang mampu menyediakan data opini masyarakat secara real time, spontan, dan merefleksikan dinamika persepsi publik secara luas [9]. Seluruh proses penelitian dilakukan secara bertahap mulai dari perumusan masalah, pengumpulan data, preprocessing, ekstraksi fitur sentimen, penyeimbangan distribusi kelas, pembangunan model klasifikasi, evaluasi performa, hingga penyusunan hasil penelitian.

Teknik pengumpulan data dilakukan menggunakan metode crawling atau scraping berbasis API dan perangkat lunak berbasis Python [10]. Proses pengambilan data memanfaatkan library Pandas serta Command-Line Tool Tweet Harvest yang telah terintegrasi dengan mekanisme autentikasi platform X. Pengumpulan dilakukan menggunakan sejumlah kata kunci yang secara langsung berhubungan dengan topik penelitian, antara lain “Makan Bergizi Gratis”, “Program MBG”, “MBG pemerintah”, dan kombinasi hashtag yang relevan. Strategi pemilihan kata kunci diterapkan untuk meningkatkan relevansi data sekaligus mengurangi masuknya data yang tidak berkaitan (noise).

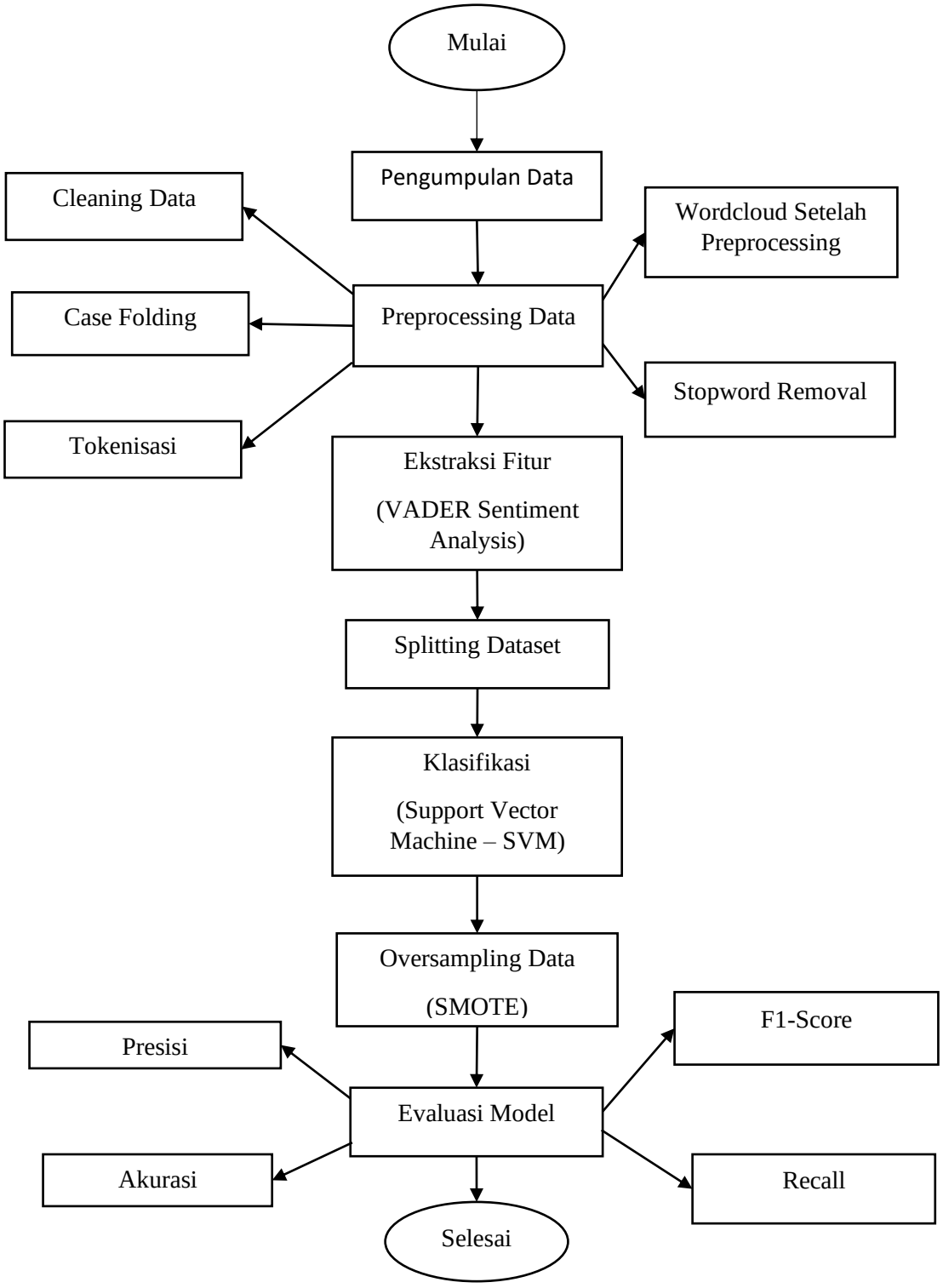
Data yang diambil hanya berupa unggahan publik tanpa retweet untuk mengurangi duplikasi informasi. Kriteria inklusi meliputi tweet yang memiliki konten tekstual dan relevan dengan topik penelitian, sedangkan data duplikat, unggahan kosong, serta konten nonbahasa Indonesia dieliminasi. Dari total 2.623 tweet awal, dilakukan proses validasi

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 259–273
DOI : 10.47065/bulletinds.v5i3.10323
https://ejurnal.seminar-id.com/index.php/bulletinds

dan pembersihan sehingga diperoleh dataset akhir yang digunakan pada tahap eksperimen. Seluruh proses crawling dan pengolahan data dijalankan menggunakan Python versi 3.11 dengan pustaka utama *pandas*, *scikit-learn*, *nltk*, *sastrawi*, *vaderSentiment*, dan *imbalanced-learn*.

2.3 Alur Penelitian

Gambar 1 pada diagram alir penelitian disusun untuk memvisualisasikan rangkaian tahapan penelitian secara menyeluruh, mulai dari data tweet mentah hingga proses evaluasi hasil klasifikasi sentimen. Keberadaan diagram ini membantu memastikan bahwa setiap langkah penelitian saling berkaitan secara logis dan sejalan dengan tujuan penelitian kuantitatif, khususnya dalam menguji efektivitas integrasi metode VADER, SVM, dan SMOTE pada analisis sentimen multikelas dalam tiga kategori, yaitu positif, netral, dan negatif.



Gambar 1. Diagram Alir Penelitian

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 259–273
DOI : 10.47065/bulletinds.v5i3.10323
https://ejurnal.seminar-id.com/index.php/bulletinds

2.4 Populasi dan Sample Penelitian

Populasi penelitian mencakup seluruh tweet publik pada platform X yang mengandung kata kunci atau hashtag yang berkaitan dengan Program Makan Bergizi Gratis selama periode pengumpulan data. Populasi tidak dibatasi berdasarkan karakteristik demografis pengguna karena unit analisis penelitian adalah isi teks, bukan identitas akun. Pendekatan tersebut didasarkan pada asumsi bahwa opini yang disampaikan melalui media sosial dapat merepresentasikan persepsi publik secara aktual terhadap isu kebijakan yang sedang berkembang. Dari populasi tersebut diperoleh dataset awal berupa 2.623 tweet yang memenuhi kriteria relevansi terhadap topik penelitian.

2.5 Metode Pengambilan Sample

Penentuan sampel dilakukan menggunakan pendekatan Slovin dengan tingkat kesalahan (error rate) sebesar 5% [18],[19],. Pendekatan ini dipilih karena sesuai digunakan pada populasi yang besar dan bersifat dinamis sebagaimana karakteristik data media sosial. Formula yang digunakan adalah:

$$n = \frac{N}{1+N(e)^2} \quad (1)$$

Keterangan n untuk ukuran sampel, N untuk ukuran populasi, dan e untuk margin of error (0,05). Rumus tersebut digunakan untuk memperoleh jumlah data yang representatif tanpa mengurangi kemampuan model dalam mempelajari pola distribusi sentimen. Setelah proses seleksi dan validasi data dilakukan, dataset akhir digunakan sebagai bahan analisis dalam proses pelatihan dan pengujian model. Selanjutnya, data dipisahkan menggunakan rasio 70:30, yaitu 70% sebagai data pelatihan dan 30% sebagai data pengujian. Strategi pembagian ini digunakan untuk memastikan kemampuan generalisasi model terhadap data baru yang tidak dilibatkan selama proses pembelajaran.

2.6 Instrumen Penelitian

Instrumen penelitian dalam studi ini bersifat komputasional dan tidak menggunakan kuesioner konvensional. Instrumen utama terdiri atas skrip pengolahan data yang dibangun menggunakan bahasa pemrograman Python serta algoritma analisis sentimen yang digunakan dalam pipeline penelitian. Instrumen tersebut mencakup modul crawling data, preprocessing, ekstraksi fitur menggunakan VADER, penanganan ketidakseimbangan data dengan SMOTE, klasifikasi menggunakan Support Vector Machine (SVM), serta evaluasi performa menggunakan confusion matrix [12]. Variabel pengukuran direpresentasikan melalui skor sentimen yang dihasilkan VADER berupa nilai positive, neutral, negative, dan compound serta label kelas sentimen hasil klasifikasi.

Validitas instrumen dijaga melalui penggunaan algoritma, pustaka, dan prosedur yang telah digunakan secara luas dalam penelitian analisis sentimen sebelumnya. Reliabilitas dijamin dengan menerapkan parameter pemrosesan yang konsisten sehingga proses analisis dapat direplikasi dan menghasilkan keluaran yang stabil. Sebelum analisis akhir dilakukan, seluruh pipeline diuji menggunakan sebagian dataset untuk memastikan tidak terdapat kesalahan pemrosesan, bias model, maupun overfitting [20].

2.7 Preprocessing Data

Tabel 1. Cleaning Dataset

No	full_text	Cleaning
0	Menteri Bappenas: Program Makan Bergizi Gratis...	Menteri Bappenas Program Makan Bergizi Gratis ...
1	Program Makan Bergizi Gratis (MBG) Mulai 6 Jan...	Program Makan Bergizi Gratis MBG Mulai Januar...
2	43 pemilik katering menuntut uang jaminan yg t...	pemilik katering menuntut uang jaminan yg tlh...
3	Program Makan Bergizi Gratis akan Dorong Permi...	Program Makan Bergizi Gratis akan Dorong Permi...
4	@KKusumawardani Persiapan Makan Bergizi Gratis...	Persiapan Makan Bergizi Gratis Mbak

Berdasarkan Tabel 1, tahap *cleaning* merupakan proses eliminasi unsur-unsur teks yang tidak memberikan kontribusi langsung terhadap analisis sentimen. Unsur yang dihapus meliputi tautan web, *mention* pengguna, tagar yang tidak relevan, emoji, angka, karakter khusus, serta spasi berlebih. Hasil pada tabel menunjukkan adanya perubahan dari kolom *full_text* menjadi *cleaning*, di mana teks dipertahankan hanya pada bagian yang memiliki informasi penting untuk analisis. Tahap ini bertujuan mengurangi *noise* sehingga representasi teks menjadi lebih bersih, konsisten, dan fokus sebelum diproses pada tahapan berikutnya.

Tabel 2. Case Folding

No	full_text	cleaning	case_folding
0	Menteri Bappenas: Program Makan Bergizi Gratis...	Menteri Bappenas Program Makan Bergizi Gratis ...	menteri bappenas program makan bergizi gratis ...
1	Program Makan Bergizi Gratis (MBG) Mulai 6 Jan...	Program Makan Bergizi Gratis MBG Mulai Januar...	program makan bergizi gratis mbg mulai januar...
2	43 pemilik katering menuntut uang jaminan yg t...	pemilik katering menuntut uang jaminan yg tlh...	pemilik katering menuntut uang jaminan yg tlh...

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 259–273
DOI : 10.47065/bulletinds.v5i3.10323
https://ejurnal.seminar-id.com/index.php/bulletinds

No	full_text	cleaning	case_folding
3	Program Makan Bergizi Gratis akan Dorong Permi...	Program Makan Bergizi Gratis akan Dorong Permi...	program makan bergizi gratis akan dorong permi...
4	@KKusumawardani Persiapan Makan Bergizi Gratis...	Persiapan Makan Bergizi Gratis Mbak	persiapan makan bergizi gratis mbak

Berdasarkan Tabel 2, tahap *case folding* dilakukan dengan mengubah seluruh karakter teks menjadi huruf kecil menggunakan fungsi transformasi *string* pada Python. Proses ini bertujuan menyeragamkan format penulisan sehingga perbedaan penggunaan huruf besar dan huruf kecil tidak menghasilkan token yang berbeda pada tahap analisis berikutnya. Hasil pada tabel menunjukkan perubahan teks dari kolom *cleaning* menjadi *case_folding*, di mana seluruh karakter dikonversi ke bentuk huruf kecil tanpa mengubah makna utama dari isi teks. Tahap ini penting untuk mengurangi duplikasi token dan meningkatkan konsistensi representasi data sebelum proses tokenisasi dilakukan.

Tabel 3. Tokenizing

No	full_text	cleaning	case_folding	tokenize
0	Menteri Bappenas: Program Makan Bergizi Gratis...	Menteri Bappenas Program Makan Bergizi Gratis ...	menteri bappenas program makan bergizi gratis ...	[menteri, bappenas, program, makan, bergizi, g...
1	Program Makan Bergizi Gratis (MBG) Mulai 6 Jan...	Program Makan Bergizi Gratis MBG Mulai Januar...	program makan bergizi gratis mbg mulai januar...	[program, makan, bergizi, gratis, mbg, mulai, ...
2	43 pemilik katering menuntut uang jaminan yg t...	pemilik katering menuntut uang jaminan yg tlh...	pemilik katering menuntut uang jaminan yg tlh...	[pemilik, katering, menuntut, uang, jaminan, y...
3	Program Makan Bergizi Gratis akan Dorong Permi...	Program Makan Bergizi Gratis akan Dorong Permi...	program makan bergizi gratis akan dorong permi...	[program, makan, bergizi, gratis, akan, dorong...
4	@KKusumawardani Persiapan Makan Bergizi Gratis...	Persiapan Makan Bergizi Gratis Mbak	persiapan makan bergizi gratis mbak	[persiapan, makan, bergizi, gratis, mbak]

Berdasarkan Tabel 3, tahap *tokenizing* dilakukan menggunakan library *Natural Language Toolkit (NLTK)* untuk memecah teks menjadi unit-unit kata (*token*) yang dapat diproses secara komputasional. Proses ini mengubah teks hasil *case folding* menjadi daftar kata terpisah sehingga setiap elemen teks dapat dianalisis secara lebih terstruktur pada tahapan berikutnya. Hasil pada tabel menunjukkan bahwa kalimat yang semula berbentuk teks utuh diubah menjadi kumpulan token dalam bentuk daftar kata, seperti kata *program*, *makan*, *bergizi*, dan *gratis*. Tahap *tokenizing* berperan penting dalam mempersiapkan data agar proses penghapusan *stopword*, ekstraksi fitur, dan analisis sentimen dapat dilakukan secara lebih efektif dan akurat.

Tabel 4. Stopword Removal

No	full_text	cleaning	case_folding	tokenize	stopword removal
0	Menteri Bappenas: Program Makan Bergizi Gratis...	Menteri Bappenas Program Makan Bergizi Gratis ...	menteri bappenas program makan bergizi gratis ...	[menteri, bappenas, program, makan, bergizi, g...	menteri bappenas program makan bergizi gratis ...
1	Program Makan Bergizi Gratis (MBG) Mulai 6 Jan...	Program Makan Bergizi Gratis MBG Mulai Januar...	program makan bergizi gratis mbg mulai januar...	[program, makan, bergizi, gratis, mbg, mulai, ...	program makan bergizi gratis mbg januari targe...
2	43 pemilik katering menuntut uang jaminan yg t...	pemilik katering menuntut uang jaminan yg tlh...	pemilik katering menuntut uang jaminan yg tlh...	[pemilik, katering, menuntut, uang, jaminan, y...	pemilik katering menuntut uang jaminan yg tlh ...
3	Program Makan Bergizi Gratis akan Dorong Permi...	Program Makan Bergizi Gratis akan Dorong Permi...	program makan bergizi gratis akan dorong permi...	[program, makan, bergizi, gratis, akan, dorong...	program makan bergizi gratis dorong permintaan...

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 259–273
DOI : 10.47065/bulletinds.v5i3.10323
https://ejurnal.seminar-id.com/index.php/bulletinds

No	full_text	cleaning	case_folding	tokenize	stopword removal
4	@KKusumawardani Persiapan Makan Bergizi Gratis...	Persiapan Makan Bergizi Gratis Mbak	persiapan makan bergizi gratis mbak	[persiapan, makan, bergizi, gratis, mbak]	persiapan makan bergizi gratis mbak

Berdasarkan Tabel 4, tahap *stopword removal* dilakukan menggunakan pustaka *Sastrawi* dan *Natural Language Toolkit (NLTK)* untuk menghapus kata-kata umum yang tidak memberikan kontribusi signifikan terhadap analisis sentimen. Proses ini bertujuan mengurangi kata yang sering muncul namun memiliki nilai informasi rendah sehingga teks menjadi lebih fokus pada kata-kata yang merepresentasikan makna utama. Hasil pada tabel menunjukkan bahwa beberapa kata yang dianggap tidak relevan dihilangkan dari hasil *tokenizing* sehingga menghasilkan representasi teks yang lebih ringkas dan informatif. Tahap *stopword removal* menjadi penting karena dapat meningkatkan kualitas data sebelum dilakukan ekstraksi fitur dan proses klasifikasi sentimen.

2.8 Penerapan Vadersentiment

Tahapan ekstraksi fitur dilakukan menggunakan metode VADER (*Valence Aware Dictionary and Sentiment Reasoner*). Metode ini digunakan untuk mengubah data teks menjadi representasi numerik sentimen berdasarkan leksikon dan aturan linguistik yang telah ditentukan sebelumnya. Setiap tweet menghasilkan skor sentimen yang terdiri atas dimensi positif, netral, negatif, dan compound. Skor tersebut selanjutnya digunakan sebagai fitur masukan pada model klasifikasi. Pendekatan ini dipilih karena dinilai efektif dalam menangani karakteristik bahasa informal yang umum ditemukan pada media sosial.

Nilai compound telah dinormalisasi pada rentang −1 hingga +1, di mana nilai di atas nol menunjukkan dominasi sentimen positif, nilai mendekati nol merefleksikan sentimen netral, dan nilai di bawah nol menandakan kecenderungan sentimen negatif [14]. Secara matematis, perhitungan skor agregat ini dirumuskan sebagai berikut:

$$\text{Compound} = \frac{\sum_{i=1}^{n} s_i}{\sqrt{\sum_{i=1}^{n} s_i^2 + \alpha}} \quad (2)$$

dengan s_i sebagai skor leksikon setiap kata dan α sebagai konstanta normalisasi. Formulasi ini berperan dalam mentransformasikan variabel independen berupa fitur teks menjadi nilai numerik yang selanjutnya digunakan untuk penentuan variabel dependen dalam proses klasifikasi sentimen.

Implementasi metode VADER dilakukan dengan dukungan sejumlah library Python. Library vaderSentiment berperan sebagai komponen inti dalam analisis sentimen berbasis leksikon, khususnya melalui modul SentimentIntensityAnalyzer untuk menghitung skor sentimen setiap teks. Selain itu, pandas digunakan untuk mengelola data hasil preprocessing dan menyimpan skor sentimen ke dalam bentuk DataFrame, sementara numpy dimanfaatkan untuk mendukung operasi numerik serta konversi tipe data hasil ekstraksi

2.9 Synthetic Minority Over-sampling Technique (SMOTE)

Distribusi kelas sentimen kemudian dianalisis untuk mengidentifikasi adanya kondisi data tidak seimbang (imbalanced dataset). Ketidakseimbangan distribusi kelas berpotensi menyebabkan model cenderung mempelajari kelas mayoritas dan mengabaikan kelas minoritas [15]. Untuk mengatasi kondisi tersebut digunakan metode Synthetic Minority Oversampling Technique (SMOTE), yaitu teknik oversampling yang menghasilkan data sintetis berdasarkan pendekatan k-nearest neighbors [16]. Persamaan pembentukan data sintetis pada SMOTE dinyatakan sebagai:

$$x_{new} = x_i + \delta(x_{zi} - x_i) \quad (3)$$

dengan δ merupakan nilai acak pada rentang $[0, 1]$yang mengatur posisi sampel sintetis di sepanjang garis penghubung kedua titik tersebut. Proses ini bertujuan memperluas sebaran kelas minoritas sehingga representasi variabel independen menjadi lebih proporsional. Beberapa indikator operasional dalam penerapan SMOTE meliputi jumlah tetangga terdekat (k) yang digunakan, nilai δ sebagai pengendali variasi lokal data sintetis, serta rasio *oversampling* yang menentukan tingkat penambahan data minoritas dibandingkan kelas mayoritas.

Tahapan penyeimbangan data dalam penelitian ini berfungsi untuk mengatasi ketidakseimbangan distribusi kelas pada data sentimen multikelas yang dapat menurunkan kemampuan model dalam mengenali kelas minoritas. Untuk mengatasi kondisi tersebut digunakan metode Synthetic Minority Over-sampling Technique (SMOTE), yaitu teknik oversampling yang menghasilkan data sintetis berdasarkan kedekatan antar data pada ruang fitur. Implementasi penyeimbangan data dilakukan melalui analisis distribusi kelas sebelum SMOTE, proses pembentukan data sintetis menggunakan SMOTE, serta evaluasi distribusi data setelah penerapan metode untuk meningkatkan kualitas klasifikasi sentimen.

2.9.1 Distribusi Data Sebelum SMOTE

Distribusi data sebelum SMOTE menggambarkan kondisi awal sebaran jumlah data pada setiap kelas sentimen sebelum dilakukan proses penyeimbangan. Tahap ini bertujuan untuk mengidentifikasi tingkat ketimpangan kelas pada dataset hasil ekstraksi fitur VADER sehingga dapat diketahui dominasi kelas mayoritas maupun keterwakilan kelas minoritas.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 259–273
DOI : 10.47065/bulletinds.v5i3.10323
https://ejurnal.seminar-id.com/index.php/bulletinds

Analisis dilakukan dengan menghitung jumlah sampel pada masing-masing kategori sentimen, yaitu positif, netral, dan negatif, sebagai dasar penerapan teknik oversampling.

Secara operasional, distribusi awal dianalisis menggunakan pustaka *pandas* melalui fungsi *value_counts()* untuk menghitung frekuensi kelas dan mengelola data dalam bentuk *DataFrame*. Visualisasi distribusi selanjutnya dapat disajikan menggunakan *matplotlib* dan *seaborn* dalam bentuk grafik batang agar pola ketidakseimbangan data lebih mudah diamati. Indikator tahap ini meliputi perbedaan jumlah data antar kelas, identifikasi kelas mayoritas dan minoritas, serta visualisasi yang menunjukkan tingkat ketimpangan dataset.

2.9.2 Distribusi Data Setelah SMOTE

Distribusi data setelah SMOTE menggambarkan kondisi sebaran kelas sentimen setelah proses oversampling dilakukan untuk mengurangi ketimpangan data. Tahap ini bertujuan mengevaluasi efektivitas SMOTE dalam menghasilkan distribusi kelas yang lebih seimbang dengan membandingkan jumlah data setiap kelas sebelum dan sesudah proses penyeimbangan. Dataset dinilai berhasil diseimbangkan apabila proporsi antar kelas menjadi lebih merata sehingga lebih representatif untuk proses klasifikasi. Secara operasional, analisis distribusi pasca-SMOTE dilakukan menggunakan *pandas* untuk menghitung frekuensi kelas, sedangkan *matplotlib* dan *seaborn* digunakan untuk memvisualisasikan perubahan distribusi data. Proses penyeimbangan didukung oleh pustaka *numpy*, *imbalanced-learn*, dan *scikit-learn* sehingga menghasilkan dataset yang lebih proporsional dan siap digunakan pada model SVM. Tahap ini berkontribusi dalam meningkatkan kemampuan klasifikasi terhadap seluruh kelas sentimen serta memperkuat validitas hasil penelitian.

2.10 Support Vector Machine (SVM)

Tahap klasifikasi dilakukan menggunakan Support Vector Machine (SVM). Algoritma ini dipilih karena memiliki kemampuan yang baik dalam menangani data berdimensi tinggi dan telah banyak digunakan dalam klasifikasi teks [17]. Model dibangun menggunakan fitur sentimen hasil ekstraksi VADER yang telah diseimbangkan menggunakan SMOTE. SVM bekerja dengan membentuk hyperplane optimal yang memaksimalkan margin antar kelas sehingga menghasilkan pemisahan kelas yang lebih stabil. Secara matematis, fungsi keputusan SVM dirumuskan sebagai:

$$f(x) = \sum_{i=1}^{m} a_i \, y_i K(x, x_i) + b \tag{4}$$

dengan x merepresentasikan vektor fitur teks, a_i sebagai bobot atau *Lagrange multipliers*, y_i sebagai label kelas, $K(x, x_i)$ sebagai fungsi kernel, dan b sebagai bias. Karakteristik kernel ini menjadi indikator operasional penting yang memengaruhi kemampuan model dalam membedakan kelas sentimen pada data teks yang berdimensi tinggi dan bersifat kompleks.

Tahapan klasifikasi menggunakan Support Vector Machine (SVM) menjadi bagian utama penelitian karena menghasilkan prediksi sentimen multikelas berupa positif, netral, dan negatif berdasarkan fitur numerik hasil ekstraksi VADER dan data yang telah diseimbangkan menggunakan SMOTE. Pemilihan SVM didasarkan pada kemampuannya dalam mengolah data berdimensi tinggi, menangani pola data yang kompleks, serta menghasilkan generalisasi yang baik melalui konsep maximum margin hyperplane. Implementasi klasifikasi dilakukan melalui pembagian data latih dan data uji, pemilihan kernel, serta proses pelatihan dan pengujian model untuk memperoleh hasil klasifikasi yang optimal.

Berbagai penelitian menunjukkan bahwa Support Vector Machine (SVM) merupakan algoritma yang konsisten dan andal untuk klasifikasi teks, termasuk analisis opini dan sentimen pada media sosial. Kinerja SVM dipengaruhi oleh representasi fitur yang digunakan, seperti TF-IDF atau *word embedding*, serta pemilihan kernel yang sesuai agar model mampu menangkap pola sentimen dan menghasilkan prediksi yang akurat pada data tweet

2.11 Pembagian Data Latih dan Data Uji

Pembagian data latih dan data uji merupakan proses pemisahan dataset menjadi dua kelompok dengan fungsi yang berbeda, yaitu data latih (*training set*) untuk membangun model dan data uji (*testing set*) untuk mengevaluasi kemampuan model dalam melakukan generalisasi terhadap data yang belum pernah dipelajari sebelumnya. Dalam penelitian ini, dataset dibagi menggunakan rasio 70% data latih dan 30% data uji. Proses pemisahan dilakukan setelah tahapan ekstraksi fitur dan penyeimbangan data pada data latih agar evaluasi model tetap objektif serta menghindari terjadinya *data leakage* yang dapat memengaruhi validitas hasil klasifikasi. Secara operasional, proses pembagian data dilakukan menggunakan pustaka *scikit-learn* melalui fungsi *train_test_split* dan didukung *numpy* untuk menjaga kompatibilitas struktur data numerik. Indikator tahap ini meliputi terbentuknya dua subset dengan proporsi 70:30, tidak adanya duplikasi data antara data latih dan data uji, serta konsistensi fitur dan label pada kedua kelompok data.

2.12 Confusion Matrix dan Metrik Evaluasi

Evaluasi performa model dilakukan menggunakan confusion matrix sebagai dasar perhitungan metrik akurasi, precision, recall, dan F1-score [18]. Akurasi digunakan untuk mengukur proporsi prediksi yang benar terhadap keseluruhan data [19], Akurasi dirumuskan sebagai:

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \tag{5}$$

Precision mengukur tingkat ketepatan prediksi dengan menghitung rasio prediksi benar terhadap seluruh prediksi positif:

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 259–273
DOI : 10.47065/bulletinds.v5i3.10323
https://ejurnal.seminar-id.com/index.php/bulletinds

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

Recall atau sensitivitas menilai kemampuan model dalam mengenali seluruh data yang benar pada suatu kelas:

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

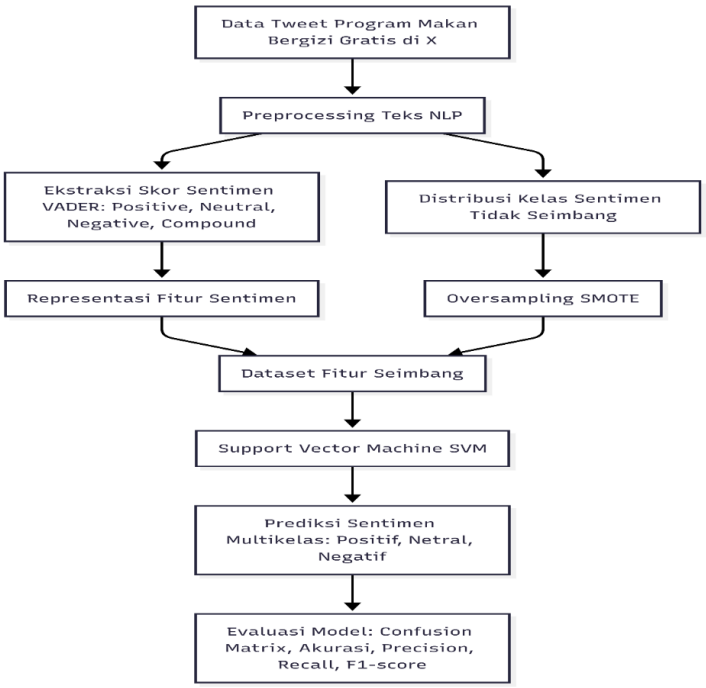
Sementara itu, F1-score dirumuskan sebagai rata-rata harmonis antara precision dan recall guna menyeimbangkan keduanya [20]:

$$F1\text{-}score = 2 \cdot \frac{Precision \cdot Recall}{Precision+Recall} \quad (8)$$

Kombinasi ketiga metrik ini menjadi krusial dalam analisis sentimen multikelas karena mampu menunjukkan sejauh mana fitur teks yang dihasilkan dari proses NLP dapat dipetakan secara konsisten menjadi label sentimen yang akurat.

2.13 Kerangka Berpikir

Pada Gambar 2 melalui rangkaian metode tersebut, penelitian ini mengembangkan pendekatan analisis sentimen multikelas yang mengintegrasikan preprocessing berbasis NLP, ekstraksi sentimen menggunakan VADER, penanganan ketidakseimbangan data melalui SMOTE, serta klasifikasi menggunakan SVM dalam satu kerangka analitis yang terukur dan dapat direplikasi.



Gambar 2. Kerangka Berpikir

3. HASIL DAN PEMBAHASAN

3.1 Deskripsi Data Penelitian

Penelitian ini bertujuan menganalisis sentimen masyarakat terhadap Program Makan Bergizi Gratis (MBG) pada media sosial X menggunakan pendekatan berbasis leksikon dan machine learning melalui integrasi metode VADER Sentiment, Support Vector Machine (SVM), dan teknik penyeimbangan data Synthetic Minority Over-sampling Technique (SMOTE). Hasil penelitian disajikan secara bertahap mulai dari karakteristik data, hasil ekstraksi sentimen, performa model sebelum dan sesudah penyeimbangan data, hingga evaluasi komparatif untuk mengidentifikasi dampak penerapan SMOTE terhadap klasifikasi sentimen multikelas.

Dataset penelitian diperoleh melalui proses crawling data dari platform X menggunakan kata kunci yang berkaitan dengan Program Makan Bergizi Gratis. Setelah melalui tahapan preprocessing yang mencakup cleaning, case folding, tokenizing, dan stopword removal, diperoleh dataset yang siap dianalisis. Tahapan preprocessing dilakukan untuk meningkatkan kualitas data teks dengan mengurangi noise dan memastikan konsistensi representasi linguistik sehingga dapat diproses secara optimal oleh model klasifikasi.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 259–273
DOI : 10.47065/bulletinds.v5i3.10323
https://ejurnal.seminar-id.com/index.php/bulletinds

	conversation_id_str	created_at	favorite_count	full_text	id_str	image_url	in_reply_to_screen_name	lang	location	quote_count	reply_count	retweet_count	tweet_url	user_id_str	username
0	1873748415148179781	Mon Dec 30 15:10:09 +0000 2024	0	Center of Economic and Law Studies (Celios) ba...	1873748415148179781	https://pbs.twimg.com/media/GgDkPozakAEdlVi.jpg	NaN	in	NaN	0	0	0	https://x.com/undefined/status/1873748415148179...	1718264577534906368	NaN
1	1873699007006130315	Mon Dec 30 14:46:37 +0000 2024	0	@03__nakula Jamaah pasti ikhlas dan ridho kalo...	1873742493478723830	NaN	03__nakula	in	NaN	0	0	0	https://x.com/undefined/status/1873742493478729...	1627825267220824065	NaN

Gambar 3. Data Mentah Sebelum Preprocessing

Berdasarkan Gambar 3, data mentah sebelum *preprocessing* masih mengandung berbagai unsur yang berpotensi mengganggu proses analisis, seperti URL, karakter khusus, *hashtag* yang tidak relevan, simbol, serta variasi penulisan teks. Setelah melalui tahapan *preprocessing*, gangguan tersebut berhasil diminimalkan sehingga menghasilkan dataset yang lebih bersih dan homogen. Transformasi ini memungkinkan teks menjadi lebih terstruktur dan siap digunakan pada tahap ekstraksi fitur sentimen. Tahap *preprocessing* memiliki peran penting karena kualitas data masukan akan memengaruhi kemampuan model dalam mengenali pola dan meningkatkan akurasi proses klasifikasi sentimen.

3.2 Analisis Pola Kata Dominan dan Implikasinya terhadap Fitur Sentimen

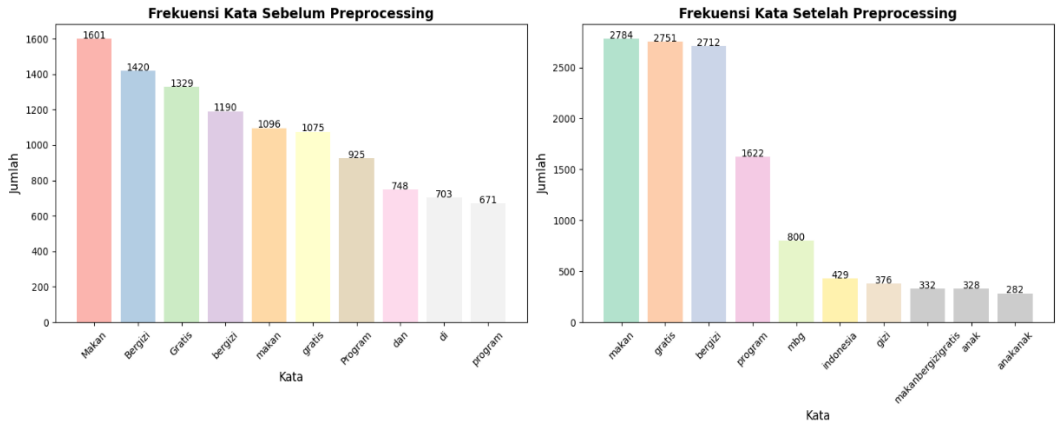
3.2.1 Wordcloud Setelah Preprocessing

Gambar 4. Wordcloud Perbandingan Setelah dan Sebelum Preprocessing

Gambar 4, menampilkan visualisasi wordcloud yang digunakan untuk membandingkan kondisi data sebelum dan setelah proses preprocessing. Visualisasi ini berfungsi sebagai alat evaluasi awal untuk mengamati perubahan distribusi kata setelah dilakukan tahap pembersihan data. Pada kondisi sebelum preprocessing, wordcloud cenderung masih didominasi oleh kata-kata tidak informatif seperti kata umum, tanda baca, atau kata penghubung yang belum tersaring dengan baik. Setelah preprocessing dilakukan, terlihat bahwa kata-kata yang muncul menjadi lebih terfokus pada istilah yang lebih bermakna dan relevan terhadap konteks dataset. Hal ini menunjukkan bahwa proses pembersihan data berhasil mengurangi noise sehingga informasi utama lebih mudah diidentifikasi. Dengan demikian, Gambar 4 memberikan gambaran visual yang jelas mengenai efektivitas tahap preprocessing dalam meningkatkan kualitas data teks untuk analisis lebih lanjut.

Tahapan ini dioperasionalkan menggunakan daftar stopword dari library Sastrawi StopWordRemover, yang kemudian dikombinasikan dengan stopword tambahan yang disesuaikan dengan konteks bahasa media sosial. Library NLTK juga digunakan sebagai referensi pendukung dalam penyusunan daftar stopword.

3.2.2 Analisis Pola Kata Dominan



Gambar 5. Frekuensi Kata Perbandingan Setelah dan Sebelum Preprocessing

Gambar 5 menunjukkan perbandingan frekuensi kata sebelum dan sesudah preprocessing, di mana kata-kata seperti *makan*, *gratis*, *bergizi*, dan *program* memiliki tingkat kemunculan yang tinggi. Setelah dilakukan preprocessing,

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 259–273
DOI : 10.47065/bulletinds.v5i3.10323
https://ejurnal.seminar-id.com/index.php/bulletinds

terlihat adanya perubahan frekuensi yang menandakan bahwa proses pembersihan data berhasil menghilangkan kata tidak relevan serta menyatukan bentuk kata yang serupa sehingga menghasilkan representasi yang lebih konsisten. Proses ini dilakukan menggunakan beberapa library Python, yaitu pandas untuk pengelolaan dataset dan perhitungan frekuensi kata melalui *value_counts()*, re untuk pembersihan teks seperti penghapusan URL dan karakter non-teks, serta NLTK atau Sastrawi untuk tokenisasi dan penghapusan stopword. Selain itu, matplotlib digunakan sebagai dasar visualisasi dan seaborn untuk meningkatkan estetika grafik. Secara keseluruhan, hasil pada Gambar 5 menunjukkan bahwa preprocessing mampu meningkatkan kualitas fitur teks yang akan digunakan pada tahap analisis selanjutnya.

3.3 Hasil Pelabelan Sentimen Multikelas Menggunakan VaderSentiment

3.3.1 Distribusi Skor Sentimen VADER

Tabel 5. Output Distribusi Skor Sentimen VADER 5 Dataset Teratas

No	full_text	cleaning	case_folding	tokenize	stopword removal	Compound_ Score	Sentiments
0	Menteri Bappenas: Program Makan Bergizi Gratis...	Menteri Bappenas Program Makan Bergizi Gratis ...	menteri bappenas program makan bergizi gratis ...	['menteri', 'bappenas', 'program', 'makan', 'b...	menteri bappenas program makan bergizi gratis ...	0.0516	Positif
1	Program Makan Bergizi Gratis (MBG) Mulai 6 Jan...	Program Makan Bergizi Gratis MBG Mulai Januar...	program makan bergizi gratis mbg mulai januar...	['program', 'makan', 'bergizi', 'gratis', 'mbg...	program makan bergizi gratis mbg januari targe...	0.0516	Positif
2	43 pemilik katering menuntut uang jaminan yg t...	pemilik katering menuntut uang jaminan yg tlh...	pemilik katering menuntut uang jaminan yg tlh...	['pemilik', 'katering', 'menuntut', 'uang', 'j...	pemilik katering menuntut uang jaminan yg tlh ...	0.0516	Positif
3	Program Makan Bergizi Gratis akan Dorong Permi...	Program Makan Bergizi Gratis akan Dorong Permi...	program makan bergizi gratis akan dorong permi...	['program', 'makan', 'bergizi', 'gratis', 'aka...	program makan bergizi gratis dorong permintaan...	0.0516	Positif
4	Waduh gmn ini Pak Wo.? Padahal programnya blm ...	Waduh gmn ini Pak Wo Padahal programnya blm be...	waduh gmn ini pak wo padahal programnya blm be...	['waduh', 'gmn', 'ini', 'pak', 'wo', 'padahal'...	gmn wo programnya blm berjalan loh budget udah...	0.0000	Netral

Tabel 5 tersebut menampilkan sebagian dataset yang telah melalui tahapan preprocessing dan analisis sentimen menggunakan VADER. Setiap baris merepresentasikan satu tweet, sedangkan setiap kolom menunjukkan hasil transformasi teks, mulai dari *full_text* (teks asli), *cleaning* (pembersihan karakter tidak relevan), *case_folding* (konversi huruf kecil), *tokenize* (pemecahan menjadi token), hingga *stopword_removal* (penghapusan kata umum). Pada tahap akhir, kolom *Compound_Score* menampilkan skor sentimen numerik hasil VADER dan kolom *Sentiments* menunjukkan label sentimen (Positif, Netral, atau Negatif), di mana contoh nilai *compound_score* sebesar 0,0516 diklasifikasikan sebagai sentimen positif karena berada di atas ambang batas yang ditentukan.

3.3.2 Aturan Threshold Pelabelan Multikelas

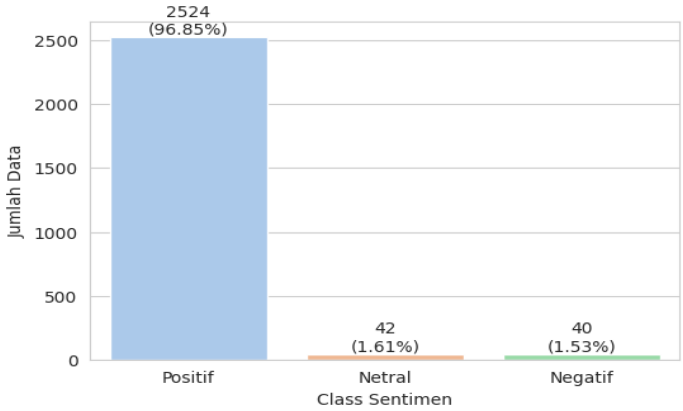
Pada penelitian ini, pelabelan sentimen dilakukan menggunakan nilai *compound score* dari metode VADER karena mampu merepresentasikan kecenderungan sentimen keseluruhan dalam rentang −1 hingga +1. Untuk mengubah skor numerik menjadi kategori sentimen multikelas, diterapkan aturan ambang batas (*threshold*) sehingga proses klasifikasi dapat dilakukan secara konsisten dan objektif, yaitu *compound* $\geq$ 0,05 dikategorikan sebagai positif, *compound* $\leq$ −0,05 sebagai negatif, dan nilai di antara −0,05 hingga 0,05 sebagai netral.

Pemilihan batas ±0,05 mengacu pada rekomendasi standar penggunaan VADER dalam penelitian analisis sentimen dan bertujuan mengurangi kesalahan klasifikasi akibat perbedaan skor yang kecil. Dengan penerapan ambang tersebut, proses pelabelan sentimen menjadi lebih terstruktur, proporsional, serta memungkinkan hasil penelitian direplikasi pada studi selanjutnya.

3.4 Hasil Pelabelan Sentimen Multikelas Menggunakan VaderSentiment

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 259–273
DOI : 10.47065/bulletinds.v5i3.10323
https://ejurnal.seminar-id.com/index.php/bulletinds

Tahap berikutnya adalah ekstraksi sentimen menggunakan metode VADER. Metode ini menghasilkan empat komponen skor sentimen, yaitu positive, neutral, negative, dan compound. Skor compound digunakan sebagai dasar penentuan label sentimen multikelas sehingga setiap tweet dikategorikan menjadi sentimen positif, netral, atau negatif.



Gambar 6. Labelling VaderSentimen

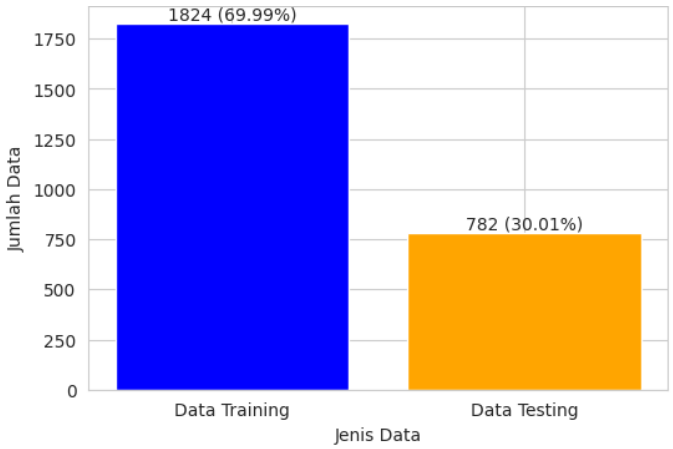
Gambar 6 menunjukkan distribusi label sentimen hasil pelabelan menggunakan metode VADER Sentiment terhadap 2.606 tweet, yang terbagi ke dalam tiga kategori utama yaitu positif, netral, dan negatif. Berdasarkan visualisasi tersebut, terlihat bahwa sentimen positif mendominasi dengan 2.524 tweet (96,85%), yang mengindikasikan bahwa mayoritas opini publik di media sosial X/Twitter terhadap Program Makan Bergizi Gratis (MBG) cenderung bersifat mendukung dan memberikan persepsi yang baik. Sementara itu, proporsi sentimen netral dan negatif relatif jauh lebih kecil, sehingga menunjukkan bahwa respons publik didominasi oleh pandangan yang positif terhadap program tersebut. Secara keseluruhan, hasil pada Gambar 6 menggambarkan kecenderungan sentimen yang sangat condong ke arah positif dalam data yang dianalisis.

Di sisi lain, sentimen netral berjumlah 42 tweet (1,61%) dan umumnya berupa informasi atau opini deskriptif tanpa kecenderungan emosi yang kuat, sedangkan sentimen negatif tercatat sebanyak 40 tweet (1,53%), yang mengindikasikan bahwa respons bernada kritik terhadap program MBG relatif rendah dalam dataset yang dianalisis. Perbedaan proporsi tersebut menunjukkan adanya ketidakseimbangan distribusi kelas sentimen.

Proses pelabelan dilakukan menggunakan library *vaderSentiment* melalui modul *SentimentIntensityAnalyzer* yang menghasilkan skor *positive*, *neutral*, *negative*, dan *compound*, dengan nilai *compound* digunakan sebagai dasar klasifikasi. Pengolahan dan penyimpanan data didukung oleh *numpy* dan *pandas*, sedangkan visualisasi dilakukan menggunakan *matplotlib*. Ketimpangan distribusi yang ditemukan menjadi dasar penerapan metode SMOTE pada tahap berikutnya untuk meningkatkan kemampuan model dalam mengenali kelas minoritas.

3.5 Kinerja Support Vector Machine (SVM) pada Dataset Imbalanced

3.5.1 Pembagian Data Latih dan Uji



Gambar 7. Jumlah Data Latih dan Data Uji

Gambar 7 menunjukkan proses pembagian data menjadi data latih dan data uji yang digunakan dalam pembangunan model klasifikasi SVM untuk memastikan kemampuan generalisasi model terhadap data baru serta meminimalkan risiko overfitting. Pembagian ini dilakukan dengan memisahkan dataset menjadi data pelatihan untuk proses pembelajaran model dan data pengujian untuk evaluasi kinerja secara objektif. Implementasi proses ini menggunakan library scikit-learn melalui fungsi *train_test_split*, serta didukung oleh numpy untuk menjaga konsistensi

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 259–273
DOI : 10.47065/bulletinds.v5i3.10323
https://ejurnal.seminar-id.com/index.php/bulletinds

struktur data numerik dan pandas dalam pengelolaan dataset. Dari total 2.606 data tweet, sebanyak 1.824 data (69,99%) digunakan sebagai data latih dan 782 data (30,01%) digunakan sebagai data uji, sehingga proporsi ini dianggap cukup seimbang untuk mendukung proses pembelajaran dan evaluasi model secara optimal.

Data latih digunakan oleh SVM untuk mempelajari hubungan antara fitur sentimen hasil metode VADER dan label sentimen, sedangkan data uji digunakan untuk mengukur kemampuan prediksi model terhadap data yang belum pernah dipelajari. Pemilihan rasio 70:30 didasarkan pada praktik umum dalam *machine learning* karena mampu menjaga keseimbangan antara kualitas pelatihan dan representativitas evaluasi sehingga mendukung pembentukan model dengan kemampuan generalisasi yang baik.

3.5.2 Akurasi SVM Sebelum SMOTE

```
=== AKURASI SVM SEBELUM SMOTE ===
0.9833759590792839

=== CLASSIFICATION REPORT SEBELUM SMOTE ===
              precision    recall  f1-score   support

     Negatif       0.67      0.17      0.27        12
      Netral       0.92      0.92      0.92        13
     Positif       0.99      1.00      0.99       757

    accuracy                           0.98       782
   macro avg       0.86      0.70      0.73       782
weighted avg       0.98      0.98      0.98       782
```

Gambar 8. Akurasi SVM Sebelum SMOTE

Gambar 8 menunjukkan hasil evaluasi performa model Support Vector Machine (SVM) sebelum penerapan SMOTE menggunakan metrik accuracy, precision, recall, dan F1-score pada 782 data uji. Berdasarkan hasil pengujian, model memperoleh nilai accuracy sebesar 98,34% yang menunjukkan bahwa sebagian besar data berhasil diklasifikasikan dengan benar oleh model. Meskipun demikian, tingginya nilai akurasi tersebut belum sepenuhnya mencerminkan kualitas model secara menyeluruh karena distribusi kelas pada dataset masih tidak seimbang. Kondisi ketidakseimbangan data ini dapat menyebabkan bias pada model terhadap kelas mayoritas, sehingga diperlukan evaluasi lebih lanjut menggunakan metrik lain untuk memastikan kinerja model yang lebih representatif.

Berdasarkan evaluasi per kelas, sentimen positif menunjukkan performa terbaik dengan *precision* 0,99, *recall* 1,00, dan *F1-score* 0,99, sedangkan kelas netral memperoleh nilai masing-masing 0,92. Sebaliknya, kelas negatif hanya mencapai *precision* 0,67, *recall* 0,17, dan *F1-score* 0,27, yang menunjukkan kemampuan deteksi kelas minoritas masih rendah. Hasil ini mengindikasikan bahwa model cenderung mengenali kelas mayoritas sehingga diperlukan penerapan SMOTE untuk menyeimbangkan distribusi data dan meningkatkan performa klasifikasi pada seluruh kelas sentimen.

3.5.3 Confusion Matrix Sebelum SMOTE

```
=== CONFUSION MATRIX SEBELUM SMOTE ===
[[  2   0  10]
 [  0  12   1]
 [  1   1 755]]
```

Gambar 9. Confusion Matrix Sebelum SMOTE

Gambar 9 menunjukkan confusion matrix dari model Support Vector Machine (SVM) sebelum penerapan SMOTE, yang digunakan untuk mengevaluasi secara lebih rinci performa klasifikasi. Meskipun hasil akurasi secara keseluruhan tergolong sangat tinggi, visualisasi confusion matrix menunjukkan bahwa hasil tersebut tidak sepenuhnya merepresentasikan kualitas klasifikasi yang sebenarnya. Hal ini terlihat dari kecenderungan model yang lebih banyak mengarahkan prediksi pada kelas mayoritas dibandingkan kelas lainnya. Kondisi tersebut mengindikasikan adanya ketidakseimbangan data yang memengaruhi kinerja model, sehingga diperlukan pendekatan tambahan untuk meningkatkan kemampuan model dalam mengenali seluruh kelas secara lebih seimbang.

Confusion matrix menunjukkan bahwa sebanyak 755 data positif berhasil diprediksi secara benar. Namun, pada kelas negatif hanya 2 data yang berhasil diklasifikasikan dengan benar, sementara sebagian besar lainnya diprediksi sebagai sentimen positif. Kelas netral menunjukkan performa yang lebih baik dengan 12 prediksi benar. Temuan ini mengindikasikan bahwa tingginya akurasi terutama dipengaruhi dominasi kelas positif dalam dataset.

3.6 Penerapan SMOTE dan Peningkatan Kinerja SVM

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 259–273
DOI : 10.47065/bulletinds.v5i3.10323
https://ejurnal.seminar-id.com/index.php/bulletinds

3.6.1 Akurasi SVM Sesudah SMOTE

```
Jumlah data sebelum SMOTE: 1824
Jumlah data sesudah SMOTE: 5301

=== AKURASI SVM SESUDAH SMOTE ===
0.8823529411764706

=== CLASSIFICATION REPORT SESUDAH SMOTE ===
              precision    recall  f1-score   support

     Negatif       0.04      0.25      0.06        12
      Netral       0.93      1.00      0.96        13
     Positif       0.99      0.89      0.94       757

    accuracy                           0.88       782
   macro avg       0.65      0.71      0.65       782
weighted avg       0.97      0.88      0.92       782
```

Gambar 10. Akurasi SVM Sesudah SMOTE

Gambar 10 menunjukkan hasil evaluasi model Support Vector Machine (SVM) setelah penerapan SMOTE menggunakan metrik yang sama seperti sebelumnya. Berdasarkan hasil pengujian, diperoleh nilai accuracy sebesar 88,24%. Meskipun terjadi penurunan nilai akurasi dibandingkan kondisi sebelum SMOTE, terdapat perubahan pola performa yang lebih signifikan pada tingkat kelas. Hal ini menunjukkan bahwa model menjadi lebih seimbang dalam melakukan klasifikasi terhadap setiap kelas sentimen, sehingga tidak lagi terlalu didominasi oleh kelas mayoritas. Dengan demikian, penerapan SMOTE memberikan dampak pada peningkatan keseimbangan kinerja model meskipun nilai akurasi secara keseluruhan mengalami penurunan.

Precision sesudah SMOTE mencapai 0,99 pada kelas positif dan 0,93 pada kelas netral. Recall kelas negatif meningkat menjadi 0,25, sedangkan F1-score kelas negatif meningkat dibanding kondisi awal meskipun masih relatif rendah.

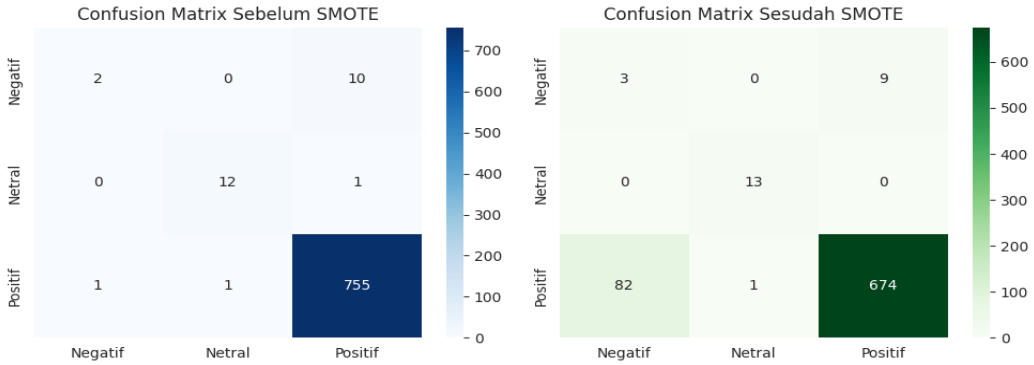
3.6.2 Confusion Matrix Sesudah SMOTE

```
=== CONFUSION MATRIX SESUDAH SMOTE ===
[[  3   0   9]
 [  0  13   0]
 [ 82   1 674]]
```

Gambar 11. Confusion Matrix Sesudah SMOTE

Gambar 11 menunjukkan perbandingan antara kelas sentimen aktual dengan kelas sentimen hasil prediksi model. Baris pada tabel menunjukkan label sentimen yang sebenarnya (actual), sedangkan kolom menunjukkan hasil prediksi model. Berdasarkan gambar tersebut, terlihat bahwa model mampu mengklasifikasikan 674 data positif dengan benar, 13 data netral dengan benar, serta 3 data negatif dengan benar.

3.6.3 Distribusi Data Sebelum dan Sesudah SMOTE



Gambar 12. Distribusi Data Sebelum dan Sesudah SMOTE

Gambar 12 menunjukkan perbandingan distribusi data sebelum dan sesudah penerapan SMOTE (Synthetic Minority Over-sampling Technique) yang digunakan untuk mengatasi ketidakseimbangan jumlah data pada setiap kelas sentimen. Analisis ini penting karena pada kondisi awal terdapat ketimpangan yang cukup signifikan antara kelas mayoritas dan kelas minoritas. Setelah penerapan SMOTE pada data latih, jumlah sampel pada kelas minoritas ditingkatkan secara sintetis sehingga distribusi data menjadi lebih seimbang. Perubahan ini dapat diamati melalui tabel jumlah data per kelas serta visualisasi grafik yang memperlihatkan kondisi distribusi sebelum dan sesudah proses SMOTE, yang menunjukkan bahwa teknik ini berhasil memperbaiki keseimbangan dataset.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 259–273
DOI : 10.47065/bulletinds.v5i3.10323
https://ejurnal.seminar-id.com/index.php/bulletinds

3.7 Pembahasan

Kesamaan mendasar dari berbagai penelitian yang telah dibahas terletak pada pemanfaatan Twitter/X sebagai sumber utama data opini publik serta penggunaan teknik machine learning untuk melakukan klasifikasi sentimen. Sejumlah studi menerapkan algoritma seperti Support Vector Machine (SVM), Naïve Bayes, maupun kombinasi beberapa metode dalam menganalisis pandangan masyarakat terhadap isu sosial, politik, dan kebijakan publik. Selain itu, beberapa penelitian juga menggunakan pendekatan pelabelan otomatis atau metode berbasis leksikon, termasuk VADER dan model berbasis transformer seperti BERT, untuk membantu proses penentuan sentimen.

Perbedaan yang menonjol antara penelitian-penelitian terdahulu dan penelitian ini terletak pada ruang lingkup serta pendekatan analisis yang digunakan. Mayoritas studi sebelumnya masih berfokus pada klasifikasi sentimen biner yang hanya membedakan antara sentimen positif dan negatif, atau sebatas membandingkan dua kelas sentimen. Sebaliknya, penelitian ini secara tegas mengadopsi pendekatan multikelas yang mencakup sentimen positif, netral, dan negatif, sehingga mampu merepresentasikan spektrum opini publik secara lebih utuh.

Selain itu, penerapan teknik penyeimbangan data seperti SMOTE dalam konteks klasifikasi sentimen multikelas masih relatif jarang dilakukan secara konsisten pada penelitian sebelumnya. Beberapa studi memang memanfaatkan metode embedding atau model transformator yang kompleks seperti BERT, namun pendekatan tersebut umumnya berdiri sendiri dan belum dikombinasikan dengan teknik penilaian sentimen berbasis leksikon seperti VADER serta algoritma klasik yang efisien seperti SVM, khususnya pada data Twitter berbahasa Indonesia.

Berdasarkan kondisi tersebut, terlihat adanya kesenjangan penelitian terkait pengembangan model sentimen multikelas yang mampu menangani permasalahan ketidakseimbangan data secara efektif. Penelitian ini hadir untuk menjawab kekosongan tersebut melalui penerapan klasifikasi multikelas yang lebih sensitif terhadap nuansa opini publik, integrasi VADER dan SVM dalam satu kerangka analisis, serta penggunaan SMOTE untuk meningkatkan keseimbangan dan performa model. Dengan fokus pada kebijakan Program Makan Bergizi Gratis (MBG) yang masih minim kajian sistematis, penelitian ini diharapkan memberikan kontribusi baru baik dari sisi metodologis maupun penerapan praktis dalam analisis sentimen kebijakan publik berbasis media sosial.

4. KESIMPULAN

Penelitian ini bertujuan menganalisis sentimen masyarakat terhadap Program Makan Bergizi Gratis (MBG) pada media sosial X melalui integrasi metode VADER Sentiment, Support Vector Machine (SVM), dan Synthetic Minority Over-sampling Technique (SMOTE). Hasil penelitian menunjukkan bahwa pendekatan yang digunakan mampu mengidentifikasi pola persepsi publik secara sistematis melalui klasifikasi sentimen multikelas yang mencakup kategori positif, netral, dan negatif. Berdasarkan hasil analisis, distribusi sentimen masyarakat cenderung didominasi oleh sentimen positif yang mengindikasikan adanya respons publik yang relatif mendukung terhadap implementasi program tersebut pada periode pengamatan. Pengujian model menunjukkan bahwa penggunaan SVM pada dataset yang belum diseimbangkan menghasilkan nilai accuracy sebesar 98,34%. Namun, performa tersebut belum mencerminkan kemampuan klasifikasi yang optimal terhadap kelas minoritas, yang ditunjukkan oleh nilai recall kelas negatif yang hanya mencapai 0,17 dan F1-score sebesar 0,27. Setelah penerapan SMOTE, nilai accuracy menurun menjadi 88,24%, tetapi kemampuan model dalam mengenali kelas minoritas mengalami peningkatan, ditunjukkan oleh kenaikan recall kelas negatif menjadi 0,25 serta peningkatan distribusi prediksi yang lebih proporsional pada confusion matrix. Temuan ini menegaskan bahwa evaluasi model klasifikasi sentimen tidak dapat hanya didasarkan pada accuracy, tetapi perlu mempertimbangkan precision, recall, dan F1-score agar interpretasi performa model menjadi lebih komprehensif dan proporsional. Dari sisi metodologis, penelitian ini menunjukkan bahwa integrasi ekstraksi sentimen berbasis leksikon dengan teknik penyeimbangan data dan algoritma machine learning dapat menjadi pendekatan yang efektif untuk analisis opini publik berbasis media sosial. Kontribusi penelitian tidak hanya terletak pada pengembangan model klasifikasi sentimen multikelas, tetapi juga pada penyediaan kerangka analisis yang dapat direplikasi untuk kajian kebijakan publik berbasis data digital. Secara praktis, hasil penelitian dapat dimanfaatkan sebagai dasar pendukung evaluasi kebijakan dan pengambilan keputusan yang lebih adaptif terhadap respons publik melalui pemanfaatan informasi distribusi dan karakteristik sentimen masyarakat. Meskipun demikian, penelitian ini masih memiliki keterbatasan pada penggunaan satu platform media sosial dan pendekatan berbasis leksikon yang belum sepenuhnya menangkap kompleksitas konteks linguistik. Oleh karena itu, penelitian selanjutnya disarankan untuk memperluas sumber data, menerapkan pendekatan berbasis deep learning atau transformer seperti BERT/IndoBERT, serta mengeksplorasi kombinasi metode penyeimbangan data dan klasifikasi yang lebih adaptif untuk meningkatkan kemampuan deteksi sentimen pada kondisi data tidak seimbang.

REFERENCES

[1] A. S. Manah, “Analisis Sentimen Publik Terhadap Kasus Keracunan Makanan MBG Berbasis Data Media Sosial Menggunakan Model Indobert,” *J. Inform. dan Tek. Elektro Terap.*, vol. 14, no. 2, 2026.

[2] T. Yuniarti and B. Andriza, “Analisis Sentimen Dan Opini Publik: Evaluasi Komunikasi Media Sosial Instagram Tim Kepresidenan@ Pco. Ri Dalam Menangani Kontroversi Program Makan Bergizi Gratis (MBG),” *J. Ilmu Komun. UHO J. Penelit. Kaji. Ilmu Komun. dan Inf.*, vol. 11, no. 1, pp. 200–215, 2026.

[3] M. S. Arum and A. Hendrawan, "Analisis Sentimen Komentar YouTube Terhadap Program Makan Bergizi Gratis (MBG) Menggunakan Long Short-Term Memory (LSTM)," *KETIK J. Inform.*, vol. 3, no. 03, pp. 121–129, 2026.
[4] M. D. Anggraeni, A. C. Yudiananta, A. H. Arifin, and W. Arifin, "Analisis Sentimen Masyarakat Terhadap Permasalahan Keracunan Program Makan Bergizi Gratis (MBG) pada Sosial Media 'X,'" *TRANSFORMASI*, vol. 21, no. 2, 2025.
[5] D. A.-T. Mukarom, G. P. Pangestu, M. F. Margadipraja, and H. M. Winata, "ANALISIS SENTIMEN BERBASIS ASPEK TERHADAP PROGRAM 'MAKAN BERGIZI GRATIS (MBG)' DI TWITTER/X DENGAN PENDEKATAN EFEKTIVITAS DUNCAN," *J. Pendidik. Sos. dan Hum.*, vol. 5, no. 1, pp. 1890–1907, 2026.
[6] M. E. Zainuri, "Analisis Sentimen Program Makan Bergizi Gratis (MBG) Pada Platform X Dengan Menggunakan Metode Naïve Bayes," *J. Univ. Din.*, 2026.
[7] F. A. Hakim, I. W. D. Prastya, and J. R. Budiani, "Sentiment Analysis of the Free Nutritious Meal Program (MBG) on Social Media X (Twitter) Using K-Nearest Neighbor and Artificial Neural Network," *J. Appl. Informatics Comput.*, vol. 10, no. 1, pp. 865–876, 2026.
[8] R. Hidayat and D. J. Ratnaningsih, "Analisis Sentimen Program Mbg Menggunakan Algoritma Random Forest Dan Naive Bayes," *J. Comput. Informatics Res.*, vol. 5, no. 1, pp. 395–400, 2025.
[9] S. Chairani Siregar, R. T. Adek, and Z. Fitri, "Sentiment Analysis of Comments on Youtube Channel Beauty Vlogger in Indonesian Language Using Support Vector Machine Method," *J. Akuntansi, Audit dan Sist. Inf. Akunt.*, vol. 2, no. 3, pp. 1–6, 2024, doi: 10.29103/icomden.v2.xxxx.
[10] Pahrul, "Skripsi perbandingan metode klasifikasi random forest, xgboost dan svm pada analisis sentimen aplikasi kredit dan pinjaman online," *J. Ilk. Hasanuddin*, vol. 3, no. 2, pp. 60–69, 2023.
[11] F. Saefulloh, "Analisis Sentimen Masyarakat Terhadap E-Commerce Pada Media Sosial Twitter Menggunakan Metode Support Vector Machine (Svm) (Studi Kasus Shopee Dan Tokopedia)," *S1 thesis, Univ. Muhammadiyah Purwokerto*, vol. 8, no. 2502, pp. 79–88, 2023.
[12] W. R. Muntaza, "Analisis Sentimen Tanggapan Masyarakat Terhadap Program Makan Bergizi Gratis Pada Platform YouTube Menggunakan SVM," in *Seminar Nasional Teknologi & Sains*, 2026, pp. 297–302.
[13] D. Triully Prasetyo and Atiqah Meutia Hilda, "Metode Support Vector Machine Untuk Analisis Sentimen Aplikasi Threads di Google Play Store," *Indones. J. Comput. Sci.*, vol. 13, no. 5, pp. 76–83, 2024, doi: 10.33022/ijcs.v13i5.4446.
[14] H. Farman, M. A. Hussain, S. Hassan, S. Shaikh, and K. Ali, "Activation function impact on rainfall prediction: comparative insights across ML and DL architectures," *Model. Earth Syst. Environ.*, vol. 11, no. 6, 2025, doi: 10.1007/s40808-025-02630-6.
[15] H. Hidayat, F. Santoso, and L. F. Lidimillah, "Analisis Sentimen Pengguna YouTube Tentang Rohingya Menggunakan Algoritma SVM (Support Vector Machine)," *G-Tech J. Teknol. Terap.*, vol. 8, no. 3, pp. 1729–1738, 2024, doi: 10.33379/gtech.v8i3.4497.
[16] S. Howay and S. Suhirman, "Comparison of SVM, NBC, and KNN Classification Methods in Determining Students' Majors at SMK N02 Manokwari," *J. Comput. Sci. Technol. Stud.*, vol. 5, no. 1, pp. 15–23, 2023, doi: 10.32996/jcsts.2023.5.1.3.
[17] E. Damayanti, A. V. Vitianingsih, S. Kacung, and D. Cahyono, "Sentiment Analysis of Alfagift Application User Reviews Using Long Short-Term Memory (LSTM) and Support Vector Machine (SVM) Methods," *J. Decod.*, vol. 4, no. 2, pp. 509–521, 2024.
[18] T. Z. Jasman, M. A. Fadhlullah, A. L. Pratama, and R. Rismayani, "Analisis Algoritma Gradient Boosting, Adaboost dan Catboost dalam Klasifikasi Kualitas Air," *J. Tek. Inform. dan Sist. Inf.*, vol. 8, no. 2, pp. 392–402, 2022, doi: 10.28932/jutisi.v8i2.4906.
[19] A. Widodo, B. Agus Herlambang, and R. Renaldy, "Optimizing Support Vector Machine (SVM) for Sentiment Analysis of Blu by BCA Reviews with Chi-Square," 2025.
[20] A. Nur, A. Saputra, R. E. Saputro, and S. Saputra, "Enhancing Sentiment Analysis Accuracy Using SVM and Slang Word Normalization on YouTube Comments," *J. dan Penelit. Tek. Inform.*, vol. 9, no. 2, 2025, doi: 10.33395/sinkron.v9i2.14513.