VOL. 03. NO. 5 SEPTEMBER 2026

JOURNAL OF COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE (JCSAI)

OPTIMALISASI MODEL KLASIFIKASI MENGGUNAKAN SUPPORT VECTOR MACHINE DENGAN SMOTE PADA DATA SENTIMEN PROGRAM MAKAN BERGIZI GRATIS

Maya Melliyananda[1], Bambang Irawan[2], Ahmad Faqih[3], Mulaywan[4].

Program Studi Teknik Informatika[123]
Program Studi Sistem Informasi[4]

STMIK IKMI Cirebon
https://ikmi.ac.id/page/18/?lang=de
mayamelliyananda18@gmail.com

(*) Corresponding Author : mayamelliyananda18@gmail.com
Published : 15 July 2026

Abstract—*The development of social media, particularly YouTube, has become a platform for the public to express opinions on various government policies, including the Free Nutritious Meal Program. The large volume of comments makes manual analysis ineffective, thus requiring a computational approach to identify public sentiment. This study aims to analyze and optimize the performance of a classification model using the Support Vector Machine (SVM) algorithm with the implementation of the Synthetic Minority Over-sampling Technique (SMOTE). The methods used in this study include data collection, sentiment labeling, reliability testing of the labeling process, text preprocessing, feature weighting using TF-IDF, data splitting with an 80:20 ratio, and model optimization using cross-validation. The study was conducted under two scenarios: SVM without SMOTE and SVM with SMOTE. The results show that the model without SMOTE achieved an accuracy of 0.5724 but tended to be biased toward the majority class. Meanwhile, the model with SMOTE achieved an accuracy of 0.3914, with improvements in precision (0.3097), recall (0.3061), and F1-score (0.3049), indicating a more balanced classification performance. Based on these results, it can be concluded that the application of SMOTE improves the performance of the SVM model in handling imbalanced data, particularly in enhancing the detection capability of minority classes.*

Keywords*: sentiment analysis, Support Vector Machine, TF-IDF, SMOTE, text classification*

Abstrak—Perkembangan media sosial, khususnya *YouTube,* telah menjadi sarana bagi masyarakat untuk menyampaikan opini terhadap berbagai kebijakan pemerintah, termasuk Program Makan Bergizi Gratis. Volume komentar yang besar menjadikan analisis manual menjadi tidak efektif, sehingga diperlukan pendekatan komputasional untuk mengidentifikasi sentimen publik. Penelitian ini bertujuan untuk menganalisis dan mengoptimalkan performa model klasifikasi menggunakan algoritma *Support Vector Machine (SVM)* dengan penerapan metode *Synthetic Minority Over-sampling Technique (SMOTE).* Metode yang digunakan meliputi pengumpulan data, pelabelan sentimen, pengujian reliabilitas pelabelan, *preprocessing teks*, pembobotan fitur menggunakan *TF-IDF,* split data dengan rasio 80:20, serta optimasi model menggunakan cross validation. Penelitian dilakukan dengan dua skenario, yaitu *SVM* tanpa *SMOTE* dan *SVM* dengan *SMOTE*. Hasil penelitian menunjukkan bahwa model tanpa *SMOTE* menghasilkan akurasi sebesar 0.5724 namun cenderung bias terhadap kelas mayoritas. Sementara itu, model dengan *SMOTE* menghasilkan akurasi sebesar 0.3914 dengan peningkatan nilai precision sebesar 0.3097, *recall* sebesar 0.3061, dan *F1-score* sebesar 0.3049, yang menunjukkan performa klasifikasi yang lebih seimbang. Berdasarkan hasil tersebut, dapat disimpulkan bahwa penerapan *SMOTE* mampu meningkatkan performa model *SVM* dalam menangani data tidak seimbang, khususnya dalam meningkatkan kemampuan deteksi pada kelas minoritas.

Kata Kunci: analisis sentimen, *Support Vector Machine, TF-IDF, SMOTE*, klasifikasi teks

INTRODUCTION

Perkembangan teknologi informasi dan komunikasi telah mengubah cara masyarakat berinteraksi, memperoleh informasi, serta menyampaikan pendapat. Media sosial menjadi ruang publik baru tempat masyarakat dapat mengekspresikan pandangan mereka secara cepat

dan terbuka. Melalui platform seperti *YouTube*, diskusi mengenai berbagai isu dapat berkembang dengan luas dan melibatkan banyak pengguna dalam waktu singkat. Kondisi ini memberikan peluang bagi peneliti untuk memanfaatkan komentar pengguna sebagai sumber data dalam memahami opini publik terhadap kebijakan pemerintah (Harwenda et al., 2025).

Salah satu kebijakan pemerintah yang banyak memperoleh perhatian ialah Program Makan Bergizi Gratis, sebuah program nasional yang berfokus pada peningkatan kualitas gizi anak dan penurunan angka stunting. Meskipun membawa tujuan positif, pelaksanaannya memicu beragam tanggapan dari masyarakat, mulai dari dukungan hingga kritik. *YouTube* menjadi salah satu media yang aktif digunakan untuk menyampaikan pendapat mengenai program tersebut. Komentar pengguna di platform ini mencerminkan sikap dan persepsi publik yang dapat dianalisis untuk mengetahui pola sentimen yang berkembang, baik positif, negatif, maupun netral

Namun, volume komentar yang sangat besar menimbulkan tantangan tersendiri dalam proses analisis secara manual. Oleh karena itu, dibutuhkan pendekatan komputasional yang mampu mengolah teks dalam jumlah besar secara efisien. Metode berbasis *machine learning* menjadi salah satu solusi yang banyak digunakan karena kemampuannya mengidentifikasi pola dalam data tidak terstruktur, termasuk teks komentar [2]. Berbagai penelitian sebelumnya menunjukkan keberhasilan pendekatan ini dalam menganalisis sentimen di sejumlah platform media sosial seperti *Twitter, Facebook*, dan *Instagram.* Meski demikian, kajian yang secara khusus membahas *respons* publik terhadap Program Makan Bergizi Gratis melalui komentar *YouTube* masih sangat terbatas.

Selain itu, permasalahan lain yang sering muncul dalam analisis sentimen adalah ketidakseimbangan distribusi data (*class imbalance*), di mana jumlah data pada masing-masing kelas tidak seimbang. Kondisi ini dapat menyebabkan model klasifikasi cenderung bias terhadap kelas mayoritas sehingga menurunkan kemampuan model dalam mengenali kelas minoritas. Untuk mengatasi permasalahan tersebut, salah satu teknik yang dapat digunakan adalah *Synthetic Minority Over-sampling Technique (SMOTE),* yaitu metode *oversampling* yang menghasilkan data sintetis pada kelas minoritas. Permasalahan ketidakseimbangan data dalam analisis sentimen juga telah banyak dibahas pada penelitian sebelumnya, di mana penggunaan teknik *SMOTE* dapat membantu meningkatkan distribusi data dan performa klasifikasi, meskipun hasilnya sangat bergantung pada karakteristik dataset yang digunakan. Namun demikian, beberapa penelitian menunjukkan bahwa penerapan *SMOTE* tidak selalu menghasilkan peningkatan performa yang signifikan, terutama pada data teks yang kompleks.

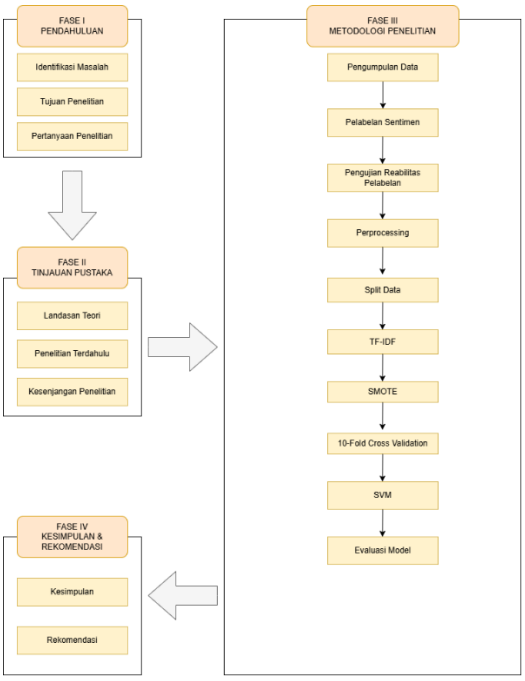
Meskipun algoritma *Support Vector Machine (SVM)* telah banyak digunakan dalam analisis sentimen dan menunjukkan performa yang baik, hasil klasifikasi sangat dipengaruhi oleh karakteristik dataset yang digunakan. Pada data teks seperti komentar media sosial, performa model seringkali mengalami penurunan akibat adanya ambiguitas bahasa, penggunaan bahasa tidak baku, serta keberadaan noise dalam data. Selain itu, penerapan teknik *oversampling* seperti *SMOTE* pada data teks berbasis *TF-IDF* tidak selalu memberikan peningkatan performa yang signifikan, karena data sintetis yang dihasilkan belum tentu mampu merepresentasikan distribusi data asli secara optimal.

Berdasarkan permasalahan tersebut, penelitian ini berfokus pada analisis sentimen komentar *YouTube* untuk mengetahui kecenderungan opini masyarakat terhadap Program Makan Bergizi Gratis. Komentar diklasifikasikan ke dalam tiga kategori utama, yaitu positif, negatif, dan netral, menggunakan algoritma *Support Vector Machine*. Selain itu, penelitian ini juga menerapkan teknik *SMOTE* untuk mengatasi ketidakseimbangan data serta menganalisis perbandingan performa model sebelum dan sesudah penerapan teknik tersebut.

MATERIALS AND METHODS

Desain penelitian berfungsi sebagai kerangka sistematis yang menjelaskan alur penelitian dan merangkum satu sama lain, dari mengidentifikasi masalah hingga pengambilan kesimpulan dan membuat rekomendasi rekomendasi. Studi ini menggunakan pendekatan kuantitatif dan metode eksperimen karena fokus pada bagaimana algoritma *Support Vector Machine (SVM)* berfungsi untuk mengklasifikasikan komentar publik pada platform *YouTube* terkait Program Makan Bergizi Gratis.

Penelitian ini dirancang dalam empat fase utama yang saling berhubungan secara logis dan berurutan. Setiap langkah, mulai dari pengumpulan data hingga evaluasi hasil mode, diperjelas. Hubungan antarproses tersebut divisualisasikan pada Gambar 1 Desain Penelitian, yang menggambarkan hubungan logis antar setiap tahap penelitian.



Gambar 1 Desain Penelitian

Gambar tersebut menggambarkan alur penelitian yang terbagi ke dalam empat fase utama, yaitu pendahuluan, tinjauan pustaka, metodologi penelitian, serta kesimpulan dan rekomendasi. Pada Fase I (Pendahuluan), penelitian diawali dengan identifikasi masalah, kemudian dilanjutkan dengan penetapan tujuan penelitian dan perumusan pertanyaan penelitian sebagai dasar arah kajian. Selanjutnya, Fase II (Tinjauan Pustaka) berisi kajian teoritis yang relevan, analisis penelitian terdahulu, serta identifikasi kesenjangan penelitian untuk menunjukkan kontribusi penelitian yang dilakukan.

Tahapan inti terdapat pada Fase III (Metodologi Penelitian), yang dimulai dari pengumpulan data, kemudian pelabelan sentimen, serta pengujian reliabilitas pelabelan untuk memastikan konsistensi data. Setelah itu dilakukan preprocessing data, pembagian data (split data), dan transformasi fitur menggunakan TF-IDF. Untuk mengatasi ketidakseimbangan data, diterapkan metode SMOTE, kemudian model divalidasi menggunakan teknik 10-fold cross validation. Model klasifikasi yang digunakan adalah Support Vector Machine (SVM), yang selanjutnya dievaluasi untuk mengukur kinerjanya. Terakhir, pada Fase IV (Kesimpulan dan Rekomendasi), hasil penelitian dirangkum dalam bentuk kesimpulan dan diberikan rekomendasi sebagai tindak lanjut dari temuan penelitian.

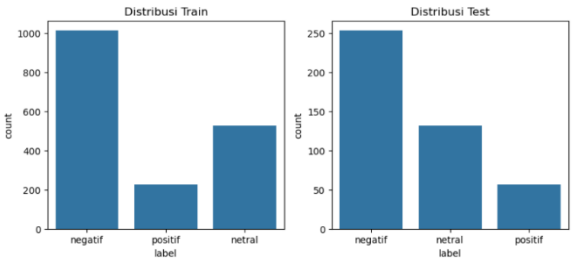
RESULTS AND DISCUSSION

Dataset yang telah melalui proses preprocessing selanjutnya dibagi menjadi dua bagian, yaitu data *training* 80% dan data *testing* 20%. Proses pembagian data dilakukan menggunakan metode *stratified sampling* untuk menjaga proporsi distribusi kelas sentimen pada kedua subset data tetap seimbang dan representatif terhadap dataset awal. Hal ini bertujuan agar model yang dilatih dapat mempelajari pola data secara optimal dan hasil evaluasi yang diperoleh lebih mencerminkan performa model pada data sebenarnya. Distribusi jumlah data pada masing-masing kelas untuk data training dan data testing terlihat pada 1 distribusi split data.

Tabel 1 distribusi split data

Label	Data training (80%)	Data testing (20%)
Negatif	1011	253
Netral	528	132
Positif	225	57

Berdasarkan tabel 1 distribusi split data, dapat diketahui bahwa distribusi data pada masing-masing kelas antara data *training* dan data *testing* memiliki proporsi yang relatif seimbang. Hal ini menunjukkan bahwa proses *stratified sampling* berhasil mempertahankan karakteristik distribusi data asli, sehingga tidak terjadi bias yang signifikan pada salah satu subset data.



Gambar 2 distribusi split data

Gambar 2 distribusi split data, menunjukkan visualisasi distribusi data pada masing-masing kelas sentimen untuk data training dan data testing. Terlihat bahwa pola distribusi antara kedua dataset memiliki kecenderungan yang serupa, sehingga dapat disimpulkan bahwa pembagian data telah dilakukan secara proporsional. Keseimbangan distribusi ini sangat penting dalam proses pelatihan model klasifikasi, karena dapat membantu model dalam mempelajari pola dari setiap kelas secara lebih adil serta mengurangi potensi bias terhadap kelas tertentu.

Term frequency-inverse document frequency (tf-idf) Setelah melalui tahap *preprocessing,* data teks diubah menjadi bentuk numerik menggunkan metode term *frequency-iverse document frequency (tf-idf).* Pada penelitian ini, ekstraksi fitur dilakukan

menggunakna *tfidfvectorizer* dengan jumlah fitur maksimum sebanyak 5000 kata.

Hasil dari proses *tf-idf* berupa matriks fitur numerik yang merepresentasikan bobot setiap kata dalam dokumen. Setiap dokumen direpresentasikan sebagai vektor fitur yang kemudian digunakan sebagai input dalam proses klasifikasi.

Berdasarkan hasil transformasi, diperoleh data dalam bentuk metriks berdimensi sejumlah dokumen dan hingga 5000 fitur kata. Representasi ini menunjukkan bahwa setiap komentar telah berhasil dikonversi ke dalam bentuk numerik yang siap digunakan pada model klasifikasi.

Proses ekstraksi fitur dilakukan pada setiap iterasi *10-fold cross validation*, di mana *tf-idf* di-fit pada data training dan digunakan untuk mentransformasikan data testing. Dengan demikian, hasil ekstraksi fitur yang diperoleh dapat digunakan secara konsisten dalam proses eksperimen tanpa menyebabkan *data leakage.*

Synthetic minority over-sampling technique (smote)

Pada tahap ini dilakukan penanganan terhadap permasalahan ketidakseimbangan data *(imbalanced dataset)* menggunakan metode *synthetic minority over-sampling technique (smote).* Ketidakeimbangan data terjadi ketika jumlah data pada masing-masing kelas sentimen tidak proporsional, sehingga berpotensi menyebabkan model klasifikasi cenderung bias terhadap kelas mayoritas.

Smote bekerja dengan cara menghasilkan data sintesis pada kelas minoritas berdasarkan kedekatan antar data dalam ruang fitur. Dengan pendekatan ini, distribusi data menjadi lebih seimbang tanpa menghilangkan informasi dari data asli. Penerapan *smote* pada penelitian ini dilakukan hanya pada data *training*. Hal ini bertujuan untuk menjaga objektivitas dalam proses evaluasi model. Distribusi data sebelum dan sesudah penerapan *smote* pada data pelatihan dapat dilihat pada tabel 2 distribusi data sebelum dan sesudah *smote*.

Tabel 2 distribusi data sebelum dan sesudah *smote*

Kelas	Sebelum smote	Sesudah smote
Negatif	1011	1011
Netral	528	1011
Positif	225	1011

Berdasarkan tabel 2 distribusi data sebelum dan sesudah *smote*, terlihat bahwa sebelum penerapan *smote*, distribusi data antar kelas masih tidak seimbang, di mana terdapat kelas dengan jumlah data yang lebih dominan dibandingkan kelas lainnya. Kondisi ini berpotensi menyebabkan model kurang optimal dalam mengenali pola pada kelas minoritas.

Setelah penerapan *smote*, jumlah data pada setiap kelas menjadi seimbang, sehingga tidak terdapat perbedaan yang signifikan antar kelas. Dengan distribusi yang lebih merata, model diharapkan mampu mempelajari pola dari setiap kelas secara lebih optimal dan mengurangi bias terhadap kelas tertentu. Oleh karena itu, penerapan *smote* dalam penelitian ini diharapkan dapat meningkatkan performa model klasifikasi, khususnya dalam hal kemampuan mendeteksi kelas minoritas.

10-fold cross validation Pada tahap ini dilakukan proses validasi model menggunakan metode *k-fold cross validation* dengan nilai k = 10. Metode ini digunakan untuk mengukur performa model secara lebih akurat dan stabil dengan cara membagi data *training* menjadi 10 bagian *(fold)*. Dalam setiap iterasi, model dilatih menggunakan 9 *fold* sebagai data validasi. Proses ini dilakukan secara berulang sebanyak 10 kali hingga seluruh data *training* pernah digunakan sebagai data validasi.

Penerapan *10-fold cross validation* bertujuan untuk mengurangi bias yang mungkin terjadi akibat pembagian data serta memberikan estimasi performa model yang lebih representatif. Hasil pengujian moenggunakan *10-fold cross validation* untuk model *support vector machine* tanpa *smote* dan dengan *smote* disajikan pada tabel 1 hasil *10-fold cross validation*.

Tabel 1 hasil *10-fold cross validation*

Fold	Tanpa smote	Dengan smote
1	0.5762	0.7302
2	0.5706	0.6842
3	0.5706	0.7302
4	0.5706	0.7788
5	0.5738	0.7029
6	0.5738	0.7359
7	0.5738	0.7326
8	0.5738	0.7293
9	0.5738	0.7689
10	0.5738	0.7458
Rata-rata	0.5731	0.7339

Tabel 1 hasil *10-fold cross validation*, dapat dilihat bahwa nilai akurasi pada setiap *fold* mengalami variasi, namun tidak menunjukkan perbedaan yang signifikan. Hal ini menunjukkan bahwa model memiliki tingkat stabilitas yang baik dalam proses pelatihan.

Support vector machine (svm)

Pada tahap ini dilakukan proses pelatihan model menggunakan algoritma *support vector machine (svm).* Model yang digunakan dalam penelitian ini menerapkan *kernel linear* yang umum digunakan dalam permasalahan klasifikasi teks. Untuk memperoleh model yang optimal, dilakukan proses *hyperparameter tuning* menggunakan metode *gridsearchcv*. Parameter yang diuji dalam proses tuning adalah nilai parameter c dengan variasi nilai 0.1, 1, dan 10. Proses ini dilakukan pada data *training* yang digunakan dalam setiap iterasi *10-fold cross validation.*

Pada setiap *fold*, model dilatih menggunakan data latih dan dilakukan pencarian parameter terbaik berdasarkan kombinasi nilai yang telah ditentukan parameter terbaik yang diperoleh dari proses ini digunakan untuk membangun model pada masing-masing *fold.* Proses tuning dilakukan pada dua skenario, yaitu model *svm* tanpa *smote* dan model *svm* dengan *smote*. Hasil parameter terbaik yang diperoleh dari masing-masing skenario ditampilkan padan tabel 3 parameter terbaik model *svm.*

Tabel 3 parameter terbaik model *svm*

Metode	Parameter c terbaik
Svm* tanpa *smote	0.1
Svm* dengan *smote	10

Berdasarkan tabel 3 parameter terbaik model *svm,* terlihat bahwa parameter terbaik yang dihasilkan pada model *svm* tanpa *smote* adalah c = 0.1, sedangkan pada model *svm* dengan *smote* diperoleh parameter terbiak c = 10. Perbedaan nilai parameter ini menunjukkan bahwa karakteristik data mempengaruhi kompleksitas model yang dibutuhkan.

Pada model tanpa *smote*, nilai c yang lebih kecil menunjukkan bahwa model cenderung menggunakan regularisasi yang lebih kuat untuk menghindari *overfitting* terhadap data yang tidak seimbang. Sementara itu, pada model dengan *smote*, nilai c yang lebih besar menunjukkan bahwa model memerlukan fleksibilitas yang lebih tinggi untuk mempelajari pola dari data yang telah diseimbangkan.

CONCLUSION

Berdasarkan hasil penelitian yang telah dilakukan mengenai optimalisasi model klasifikasi menggunkan *Support Vector Machine* dengan penerapan *SMOTE* pada data sentimen terhadap Program Makan Bergizi Gratis, maka dapat ditarik beberapa kesimpulan sebagai berikut

Model *Support Vector Machine (SVM)* tanpa penerapan *SMOTE* menunjukkan performa yang kurang optimal dalam mengklasifikasikan sentimen. Hal ini disebabkan oleh ketidakseimbangan distribusi data, sehingga model cenderung bias terhadap kelas mayoritas, yaitu sentimen negatif. Meskipun menghasilkan nilai akurasi sebesar 0.5724, model tidak mampu mengenali kelas netral dan positif dengan baik.

Model *Support Vector Machine (SVM)* dengan penerapan *SMOTE* menunjukkan peningkatan dalam kemampuan klasifikasi, terutama dalam mengenali seluruh kelas sentimen, yaitu negatif, netral, dan positif. Penerapan *SMOTE* berhasil mengatasi ketidakseimbangan data dengan menambahkan data sintesis pada kelas minoritas, sehingga model mampu mempelajari pola data secara lebih merata. Hal ini tercermin dari nilai *precision, recall*, dan *F1-score* yang lebih seimbang.

Berdasarkan perbandingan performa model, model *SVM* dengan *SMOTE* dapat dianggap lebih baik dibandingkan model tanpa *SMOTE* dalam konteks data yang tidak seimbang. Meskipun nilai akurasi pada model dengan *SMOTE* lebih rendah, model ini mampu memberikan hasil klasifikasi yang adil dan seimbang pada setiap kelas sentimen.

REFERENCE

[1] R. W. Harwenda, M. D. Angelo, I. Budi, A. B. Santoso, and P. K. Putra, "Sentiment Analysis on Government Public Policies: A Systematic Literature Review," *Dinasti International Journal of Education Management And Social Science*, vol. 6, no. 5, pp. 4192–4211, Jul. 2025, doi: 10.38035/dijemss.v6i5.4699.

[2] T. K. Deo, R. K. Deshmukh, and G. Sharma, "Comparative Study among Term Frequency-Inverse Document Frequency and Count Vectorizer towards K Nearest Neighbor and Decision Tree Classifiers for Text Dataset," *Nepal Journal of Multidisciplinary Research*, vol. 7, no. 2, pp. 1–11, Jul. 2024, doi: 10.3126/njmr.v7i2.68189.