

Analisis Sentimen Ulasan *Instagram* Menggunakan Algoritma *Support Vector Machine* dan *Random Forest* (Studi Kasus: Universitas Dian Nusantara)

Giri Purnama

Fakultas Teknik dan Informatika, Universitas Dian Nusantara, DKI Jakarta, Indonesia
giri.purnama@undira.ac.id

Article Info

Article history:

Received, 2026-06-23

Revised, 2026-07-08

Accepted, 2026-07-13

Kata Kunci:

analisis sentiment;
NLP;
Support Vector Machine;
Random Forest;
media social;

Keywords:

sentiment analysis;
NLP;
Support Vector Machine;
Random Forest;
social media;

ABSTRAK

Meningkatnya penggunaan media sosial, khususnya *Instagram* dengan lebih dari 1 miliar pengguna aktif pada tahun 2023, membuka peluang bagi Universitas Dian Nusantara (UNDIRA) untuk memahami persepsi publik melalui analisis ulasan pengguna. Penelitian ini bertujuan mengembangkan model analisis sentimen berbasis pembelajaran mesin guna mengkategorikan ulasan pengguna *Instagram* UNDIRA menjadi sentimen positif, negatif, dan netral. Kebaruan penelitian terletak pada penggabungan fitur TF-IDF dan Word2Vec secara bersamaan serta perbandingan sistematis SVM dan Random Forest pada data ulasan berbahasa Indonesia yang belum banyak dikaji pada konteks institusi pendidikan tinggi. Penelitian dilakukan melalui beberapa tahapan: pengumpulan data ulasan dari *Instagram*, pra-pemrosesan data seperti penghapusan stop words dan stemming, serta penerapan tiga model pembelajaran mesin. Model yang digunakan adalah Support Vector Machine (SVM) dengan ekstraksi fitur TF-IDF, Random Forest (RF) dengan Word2Vec, dan RF dengan TF-IDF untuk membandingkan kinerja masing-masing metode. Hasil penelitian menunjukkan bahwa model SVM dengan TF-IDF memberikan performa terbaik dengan akurasi 99.11% dan F1-Score 99.13%, melampaui Random Forest yang mencapai akurasi 96.43%.

ABSTRACT

The increasing use of social media, particularly *Instagram* with over 1 billion active users in 2023, creates opportunities for Universitas Dian Nusantara (UNDIRA) to understand public perception through user review analysis. This study aims to develop a machine learning-based sentiment analysis model to categorize UNDIRA *Instagram* user reviews into positive, negative, and neutral sentiments. The novelty of this study lies in the combined use of TF-IDF and Word2Vec features together with a systematic comparison of SVM and Random Forest on Indonesian-language review data, a context still rarely examined for higher-education institutions. The research involved data collection through *Instagram* scraping, data preprocessing including stop word removal and stemming, and the application of three machine learning models: Support Vector Machine (SVM) with TF-IDF feature extraction, Random Forest (RF) with Word2Vec, and RF with TF-IDF. Results indicate that SVM with TF-IDF achieved the best performance with 99.11% accuracy and 99.13% F1-Score, outperforming Random Forest at 96.43% accuracy.

This is an open access article under the CC BY-SA license.



Penulis Korespondensi:

Giri Purnama,
Fakultas Teknik dan Informatika
Universitas Dian Nusantara, DKI Jakarta, Indonesia
Email: *giri.purnama@undira.ac.id*

1. PENDAHULUAN

Dalam era digital, media sosial telah menjadi bagian penting kehidupan dan memberikan dampak kepada cara komunikasi manusia [1], [2], [3], [4], [5]. Ulasan di *Instagram* dapat memberikan wawasan tentang persepsi masyarakat terhadap kualitas pendidikan dan layanan UNDIRA. Namun, dengan banyaknya data,

analisis manual menjadi tidak praktis. Di sinilah peran *Natural Language Processing* (NLP) menjadi penting, untuk menganalisis sentimen ulasan secara otomatis. Dalam dunia bisnis, NLP diterapkan dalam pengembangan chatbot untuk memberikan informasi produk. NLP juga berperan penting dalam koreksi tata bahasa, identifikasi bahasa pada teks campuran di media sosial, serta klasifikasi opini pengguna pada berbagai aplikasi layanan digital [6].

Analisis sentimen adalah cabang NLP yang bertujuan mengidentifikasi dan mengklasifikasikan opini atau emosi dalam teks ke dalam kategori positif, negatif, atau netral [7], [8], [9]. Analisis sentimen umumnya dibagi menjadi tiga pendekatan utama, yaitu pendekatan berbasis kamus (*lexicon-based*), pendekatan berbasis pembelajaran mesin (*machine learning*), dan pendekatan berbasis pembelajaran mendalam (*deep learning*). Pendekatan berbasis kamus memanfaatkan kamus kata yang telah diberi label sentimen, seperti SentiWordNet dan VADER, untuk menentukan polaritas suatu teks [10], [11]. Sementara itu, pendekatan berbasis pembelajaran mesin menggunakan data berlabel untuk melatih model klasifikasi dengan algoritma seperti Naive Bayes, Support Vector Machine (SVM), dan Random Forest. Adapun pendekatan berbasis pembelajaran mendalam menerapkan jaringan saraf, seperti Long Short-Term Memory (LSTM) dan Convolutional Neural Network (CNN) [12], [13].

Penelitian yang dilakukan oleh Sari et al. (2024) bertujuan untuk menganalisis sentimen masyarakat terhadap Program Kampus Merdeka berdasarkan data unggahan di platform X (Twitter). Dataset yang digunakan terdiri dari 500 tweet yang mengandung kata kunci "*Kampus Merdeka*". Penelitian ini menerapkan algoritma Support Vector Machine (SVM) sebagai metode klasifikasi sentimen dengan seleksi fitur Chi-Square untuk memilih fitur-fitur yang paling relevan. Hasil penelitian menunjukkan bahwa penggunaan seleksi fitur Chi-Square mampu meningkatkan performa klasifikasi SVM sehingga menghasilkan akurasi yang lebih baik dalam membedakan sentimen positif, negatif, dan netral terhadap Program Kampus Merdeka [14].

Penelitian oleh Fitri et al. (2025) berfokus pada analisis sentimen publik terhadap program Kampus Mengajar melalui komentar pengguna di platform X. Penelitian ini menggunakan algoritma Naïve Bayes untuk mengklasifikasikan opini masyarakat ke dalam kategori sentimen positif, negatif, dan netral. Tujuan penelitian adalah mengetahui persepsi publik terhadap pelaksanaan Program Kampus Mengajar sekaligus mengukur kinerja algoritma Naïve Bayes dalam tugas klasifikasi sentimen [15].

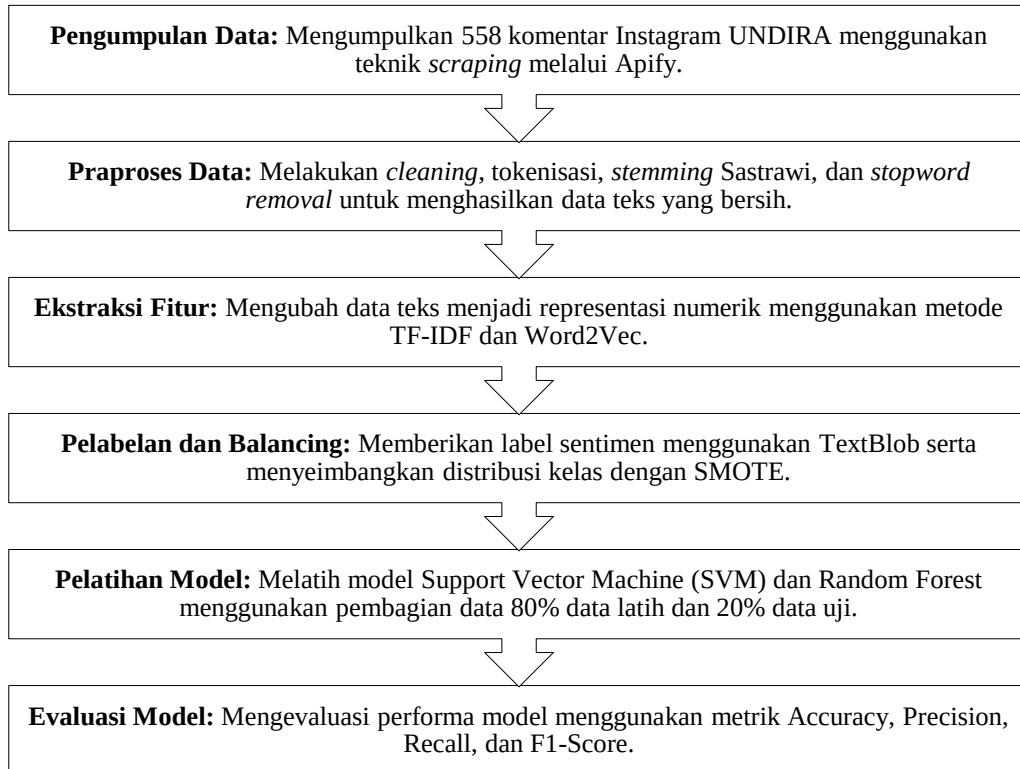
Penelitian yang dilakukan oleh Sapitri dan Costaner (2025) membandingkan kinerja algoritma Naïve Bayes Classifier (NBC) dan Support Vector Machine (SVM) dalam menganalisis sentimen masyarakat terhadap Program Kampus Merdeka di Twitter. Data yang digunakan berupa tweet yang membahas Program Kampus Merdeka, kemudian melalui tahapan prapemrosesan teks dan klasifikasi sentimen. Hasil penelitian menunjukkan adanya perbedaan performa antara kedua algoritma, di mana Support Vector Machine memberikan tingkat akurasi yang lebih tinggi dibandingkan Naïve Bayes dalam mengklasifikasikan sentimen pengguna Twitter [16].

Meskipun penelitian analisis sentimen di Indonesia terus berkembang, sebagian besar studi masih berfokus pada platform Twitter/X atau ulasan aplikasi di marketplace. Penelitian yang secara khusus mengkaji ulasan Instagram pada institusi pendidikan tinggi, terutama dengan membandingkan representasi fitur berbasis frekuensi (TF-IDF) dan representasi berbasis makna semantik (Word2Vec), masih relatif terbatas. Oleh karena itu, penelitian ini memberikan tiga kontribusi utama, yaitu menyediakan dataset ulasan Instagram UNDIRA yang telah melalui tahapan praproses *Natural Language Processing* (NLP) sesuai kaidah bahasa Indonesia, meliputi *cleaning*, tokenisasi, *stemming* menggunakan Sastrawi, dan *stopword removal*, sehingga dapat menjadi referensi bagi penelitian serupa pada institusi pendidikan lainnya.

Berdasarkan latar belakang diatas, penelitian ini melakukan evaluasi komparatif antara kombinasi algoritma Support Vector Machine (SVM) dengan TF-IDF serta Random Forest dengan Word2Vec untuk memperoleh bukti empiris mengenai pendekatan yang paling sesuai dalam mengklasifikasikan data ulasan berbahasa Indonesia berskala kecil hingga menengah. Hasil analisis sentimen selanjutnya diterjemahkan menjadi rekomendasi strategis yang dapat dimanfaatkan oleh UNDIRA dalam meningkatkan kualitas layanan, mengelola reputasi, dan memperkuat citra institusi di media sosial.

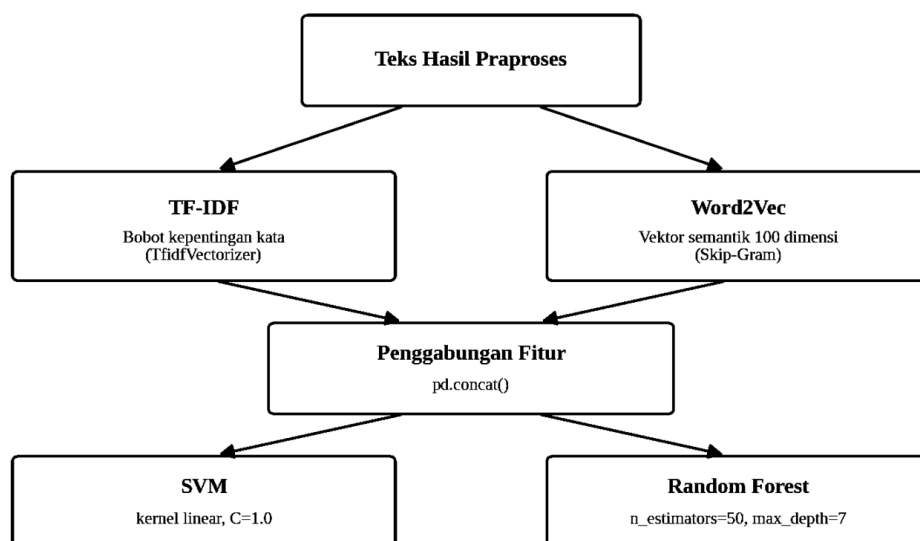
2. METODE PENELITIAN

Penelitian ini menggunakan jenis penelitian kuantitatif dengan pendekatan eksperimen. Desain penelitian berbasis data sekunder mengumpulkan data ulasan pengguna dari komentar dan unggahan terkait Instagram UNDIRA. Penelitian bertujuan mengembangkan dan menguji model analisis sentimen yang mampu mengklasifikasikan sentimen (positif, negatif, netral) secara otomatis. Proses analisis dilakukan melalui tahapan preprocessing data, implementasi algoritma NLP, serta evaluasi performa model menggunakan metrik akurasi, presisi, dan recall. Alur tahapan penelitian, mulai dari pengumpulan data hingga evaluasi model dapat dilihat pada **Gambar 1**.



Gambar 1 Alur Tahapan Penelitian

Populasi penelitian ini mencakup seluruh komentar dan ulasan pengguna pada akun Instagram resmi UNDIRA, sedangkan sampel dipilih menggunakan teknik *purposive sampling* dengan kriteria komentar berbahasa Indonesia, relevan dengan topik penelitian, dan dipublikasikan dalam rentang waktu yang telah ditentukan. Pengumpulan data dilakukan melalui teknik *scraping* menggunakan layanan Apify.com yang mematuhi regulasi yang berlaku, sehingga diperoleh sebanyak 558 komentar dari akun @undiraofficial. Data yang terkumpul kemudian melalui tahapan praproses yang meliputi *cleaning* untuk menghapus URL, tanda baca, angka, dan *whitespace* berlebih, tokenisasi menggunakan pustaka NLTK, normalisasi melalui *stemming* dengan Sastrawi, serta penghapusan *stopwords* bahasa Indonesia untuk menghasilkan data yang siap dianalisis. Tahap berikutnya adalah ekstraksi fitur menggunakan dua pendekatan, yaitu TF-IDF dan Word2Vec, seperti yang terlihat pada **Gambar 2**.



Gambar 2 Ekstraksi Fitur (TF-IDF dan Word2Vec)

TF-IDF digunakan untuk merepresentasikan bobot setiap kata berdasarkan tingkat kepentingannya dalam dokumen terhadap keseluruhan korpus, sedangkan Word2Vec digunakan untuk menghasilkan representasi vektor semantik dengan parameter *vector_size* = 100, *window* = 5, dan *min_count* = 1. Selanjutnya, data diberi label sentimen menggunakan TextBlob dan distribusi kelas diseimbangkan menggunakan SMOTE sebelum dilakukan pelatihan model. Model Support Vector Machine (SVM) dan Random Forest dilatih menggunakan pembagian data 80% untuk pelatihan dan 20% untuk pengujian, kemudian dievaluasi berdasarkan metrik accuracy, precision, recall, dan F1-score untuk menentukan performa klasifikasi.

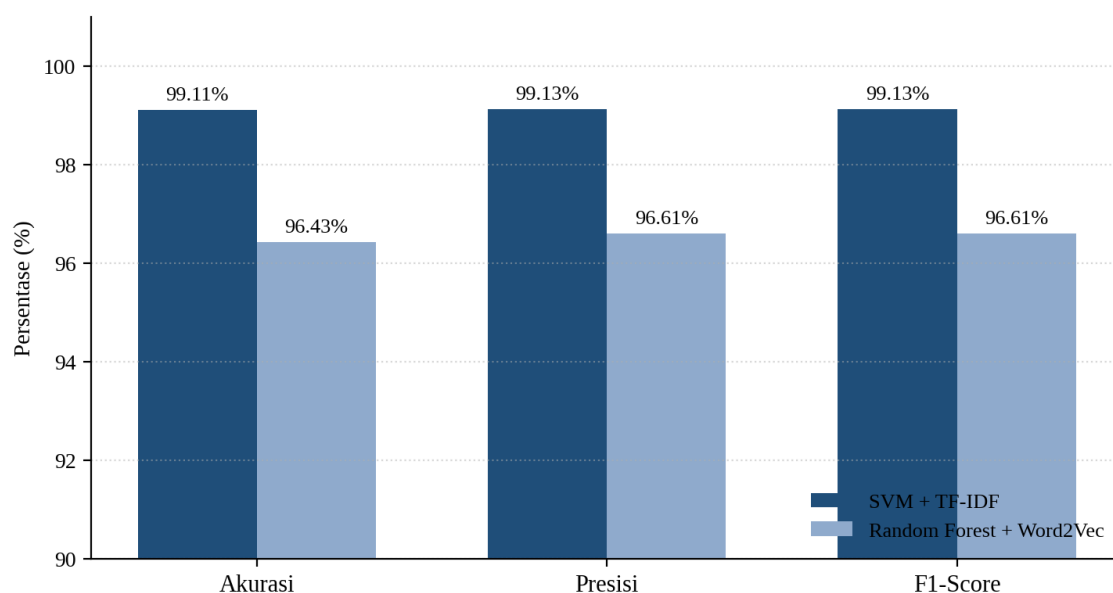
Pelabelan sentimen menggunakan TextBlob untuk menghitung polarity score. Jika nilai polarity > 0, teks dikategorikan sebagai Positive (1); jika polarity < 0, sebagai Negative (-1); dan jika polarity = 0, sebagai Neutral (0). Ketidakseimbangan data diatasi menggunakan SMOTE (Synthetic Minority Over-sampling Technique) dengan parameter *k_neighbors*=2. Dataset selanjutnya dibagi menjadi data training (80%) dan testing (20%) menggunakan stratified train-test split. Dua model dilatih: SVM dengan kernel linear (*C*=1.0, *gamma*='scale') dan Random Forest dengan parameter *n_estimators*=50, *max_depth*=7, *min_samples_split*=10, *min_samples_leaf*=4. Evaluasi model menggunakan metrik Accuracy, Precision, Recall, dan F1-Score.

3. HASIL DAN ANALISIS

Proses web scraping menggunakan Apify.com berhasil mengumpulkan 558 komentar dari akun @undiraofficial. Pada tahun 2025, Instagram memperketat kebijakan akses data mereka dengan membatasi penggunaan skrip otomatis, sehingga penggunaan Apify.com menjadi solusi yang lebih efisien dan legal. Platform ini menawarkan fitur otomatisasi, skalabilitas, serta kemudahan dalam mengekstrak data tanpa mengembangkan skrip sendiri.

Tahapan preprocessing berhasil membersihkan dan menstrukturkan 558 komentar. Setelah melalui proses cleaning, tokenisasi, stemming menggunakan Sastrawi, dan penghapusan stopwords, setiap komentar diubah menjadi daftar kata-kata yang lebih bersih dan bermakna. Data teks yang telah diproses ini siap digunakan untuk tahap ekstraksi fitur dan analisis sentimen. Ekstraksi fitur TF-IDF menghasilkan matriks numerik berbasis frekuensi kata yang mencerminkan kepentingan relatif setiap kata dalam konteks dataset. Sementara itu, Word2Vec menghasilkan vektor dimensi 100 untuk setiap kata, menangkap hubungan semantik antar kata. Penggabungan kedua fitur menghasilkan representasi teks yang lebih komprehensif, memungkinkan model menangkap baik bobot kata penting maupun konteks semantik.

Dari 558 komentar yang dianalisis menggunakan TextBlob, distribusi sentimen menunjukkan dominasi sentimen positif, diikuti netral, dan negatif. Ketidakseimbangan distribusi ini diatasi dengan penerapan SMOTE yang secara sintesis menambah sampel pada kelas minoritas melalui interpolasi dari titik data yang ada. Setelah penerapan SMOTE, dataset menjadi lebih seimbang dan memungkinkan model mempelajari pola dari semua kelas secara lebih baik. Perbandingan hasil evaluasi kedua model yang diuji dapat dilihat pada Gambar 3.



Gambar 3 Perbandingan Kinerja Model SVM+TF-IDF vs Random Forest+Word2Vec

Model Support Vector Machine (SVM) dengan ekstraksi fitur TF-IDF menunjukkan kinerja terbaik dengan memperoleh akurasi sebesar 99,11% dan F1-Score 99,13%, melampaui model Random Forest yang menghasilkan akurasi 96,43% dan F1-Score 96,61%. Keunggulan SVM didukung oleh kemampuannya dalam menemukan *hyperplane* optimal untuk memisahkan kelas sentimen pada ruang fitur berdimensi tinggi yang dihasilkan oleh TF-IDF. Meskipun Random Forest juga memberikan performa yang sangat baik, hasilnya sedikit lebih rendah, yang kemungkinan dipengaruhi oleh kompleksitas model dan potensi *overfitting*, meskipun telah dilakukan pengaturan parameter seperti *max_depth* = 7 dan *min_samples_leaf* = 4 untuk meminimalkan risiko tersebut. Selain itu, penerapan SMOTE terbukti efektif dalam meningkatkan kemampuan model mengenali kelas sentimen negatif dan netral yang sebelumnya memiliki jumlah data lebih sedikit. Temuan ini memperkuat bukti empiris bahwa kombinasi representasi berbasis frekuensi kata (TF-IDF) dengan algoritma SVM masih lebih efektif dibandingkan representasi semantik (Word2Vec) pada dataset ulasan berbahasa Indonesia berskala kecil hingga menengah.

4. KESIMPULAN

Berdasarkan hasil penelitian, dapat disimpulkan bahwa Support Vector Machine (SVM) merupakan model terbaik untuk klasifikasi sentimen ulasan Instagram UNDIRA dengan akurasi 99,11% dan F1-Score 99,13%, lebih tinggi dibandingkan Random Forest yang memperoleh akurasi 96,43% dan F1-Score 96,61%. Kombinasi ekstraksi fitur TF-IDF dan Word2Vec terbukti efektif dalam merepresentasikan data teks, di mana TF-IDF mampu memberikan bobot pada kata-kata penting, sedangkan Word2Vec menangkap hubungan semantik antar kata. Selain itu, penerapan SMOTE berhasil menyeimbangkan distribusi kelas sehingga model tidak bias terhadap kelas mayoritas dan mampu meningkatkan kemampuan klasifikasi pada sentimen negatif maupun netral. Hasil analisis sentimen yang diperoleh juga dapat dimanfaatkan oleh UNDIRA sebagai dasar untuk mengidentifikasi aspek layanan yang perlu ditingkatkan sekaligus mempertahankan keunggulan yang telah diapresiasi oleh pengguna, sehingga mendukung pengambilan keputusan berbasis data dalam meningkatkan kualitas layanan dan strategi branding institusi.

Penelitian ini memberikan kontribusi ilmiah berupa bukti empiris mengenai perbandingan efektivitas representasi fitur berbasis frekuensi (TF-IDF) dan berbasis semantik (Word2Vec) pada analisis sentimen ulasan Instagram berbahasa Indonesia dalam konteks institusi pendidikan tinggi. Kerangka metodologi yang digunakan, mulai dari praproses teks berbahasa Indonesia, ekstraksi fitur, penyeimbangan data menggunakan SMOTE, hingga evaluasi menggunakan berbagai metrik kinerja, dapat dijadikan acuan bagi penelitian serupa pada institusi maupun platform media sosial lainnya. Untuk penelitian selanjutnya, disarankan memperluas sumber data ke platform seperti Twitter/X, Facebook, dan YouTube, mengeksplorasi metode berbasis deep learning seperti LSTM dan BERT, menerapkan teknik ensemble learning serta hyperparameter tuning, memperbesar ukuran dataset, dan melibatkan pelabelan manual oleh beberapa annotator guna meningkatkan validitas hasil dibandingkan pelabelan otomatis menggunakan TextBlob.

UCAPAN TERIMA KASIH

Tim peneliti mengucapkan terima kasih kepada Universitas Dian Nusantara (UNDIRA) atas dukungan yang diberikan selama pelaksanaan penelitian ini. Ucapan terima kasih juga disampaikan kepada seluruh pihak yang telah berkontribusi dalam proses pengumpulan data, penyusunan, serta penyelesaian penelitian ini.

REFERENSI

- [1] B. Y. Geni, A. Supriyadi, H. Khotimah, and W. I. Yanti, "Rancang Bangun Company Profile Berbasis Web Menggunakan Metode Waterfall (Studi Kasus: APM Frozen Food)," *J. RESTIKOM Ris. Tek. Inform. dan Komput.*, vol. 6, no. 1, pp. 75–85, 2024.
- [2] M. M. Kananga and V. Ayumi, "Perancangan Sistem Untuk Monitoring Operasional Kontainer dengan Implementasi Metode Agile dan Decision Tree," *JUSIBI (Jurnal Sist. Inf. dan E-Bisnis)*, vol. 8, no. 1, pp. 31–37, 2026.
- [3] B. Y. Geni, D. Ramayanti, and A. Ratnasari, "Implementasi Sistem Poin of Sale Terintegrasi Berbasis Python," *JATI (Jurnal Mhs. Tek. Inform.)*, vol. 8, no. 4, pp. 4387–4393, 2024.
- [4] M. Purba, J. Junaidi, L. Iryani, N. Umilizah, H. Noprisson, and N. Ani, "An Ensemble Boosting Approach with Boruta Feature Selection for Predicting E-Payment Adoption," *Int. J. Adv. Comput. Sci. Appl.*, vol. 17, no. 4, p. 403, 2026.
- [5] N. Luhyana and V. Ayumi, "Model Sistem Manajemen Proyek Sosial Media di Perusahaan Periklanan Menggunakan Metode Waterfall," *JUKOMIKA (Jurnal Ilmu Komput. dan Inform.)*, vol. 8, no. 2, pp. 136–141, 2025.

- [6] C.-P. Chan and J.-H. Yang, "Instagram text sentiment analysis combining machine learning and NLP," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, 2023.
- [7] A. Suhendi and V. Ayumi, "Development of Web-Based Cash Management System (CMS) Based on Cash Processing Center (CPC) at PT. XYZ," *JSAI (Journal Sci. Appl. Informatics)*, vol. 8, no. 2, pp. 503–508, 2025.
- [8] S. Mailana and V. Ayumi, "Aplikasi ConTrack Untuk Manajemen dan Monitoring Data Stok Barang di Perusahaan Jasa Konstruksi," *JSAI (Journal Sci. Appl. Informatics)*, vol. 8, no. 2, pp. 520–525, 2025.
- [9] I. P. D. W. Darmawan, G. A. Pradnyana, and I. B. N. Pascima, "Optimasi Parameter Support Vector Machine Dengan Algoritma Genetika Untuk Analisis Sentimen Pada Media Sosial Instagram," *SINTECH (Science Inf. Technol. J.)*, vol. 6, no. 1, pp. 58–67, 2023.
- [10] Y. Asri, W. N. Suliyanti, D. Kuswardani, and M. Fajri, "Pelabelan Otomatis Lexicon Vader dan Klasifikasi Naive Bayes dalam menganalisis sentimen data ulasan PLN Mobile," *PETIR J. Pengkaj. dan Penerapan Tek. Inf.*, vol. 15, no. 2, pp. 264–275, 2022.
- [11] S. A. Nugraha, "Penerapan Lexicon Based Untuk Analisis Sentimen Masyarakat Indonesia Terhadap Danantara," *JATI (Jurnal Mhs. Tek. Inform.)*, vol. 9, no. 3, pp. 4949–4957, 2025.
- [12] S. Yahya, Z. Sitorus, M. Iqbal, D. Nasution, and R. F. Wijaya, "Analisis Sentimen Penerapan Deep Learning dan Analisis Sentimen terhadap Gap Kompetensi Lulusan Lembaga Pendidikan dan Pelatihan Vokasi terhadap Dunia Kerja dengan Metode Long Short-Term Memory (LSTM)," *Bull. Inf. Technol.*, vol. 6, no. 2, pp. 161–172, 2025.
- [13] V. H. Leonardi, A. Ibrahim, R. D. Kurnia, and M. Afrina, "Analisis sentimen ulasan aplikasi pembelajaran bahasa menggunakan metode VADER," *J. Algoritm.*, vol. 22, no. 1, pp. 767–776, 2025.
- [14] M. Sari, S. Syahrullah, N. T. Lapatta, and R. Ardiansyah, "Twitter (X) Sentiment Analysis of Kampus Merdeka Program Using Support Vector Machine Algorithm and Selection Feature Chi-Square," *J. Tek. Inform.*, vol. 5, no. 5, pp. 1249–1256, 2024.
- [15] S. A. Fitri, S. H. Wijoyo, and K. Rahman, "Analisis Sentimen Kampus Mengajar Menggunakan Algoritma Naïve Bayes dari Komentar pada X," *J. Sist. Informasi, Teknol. Informasi, dan Edukasi Sist. Inf.*, vol. 6, no. 2, pp. 88–101, 2025.
- [16] R. Sapitri and L. Costaner, "Analisis Sentimen Terhadap Program Kampus Merdeka Menggunakan Algoritma Naïve Bayes Classifier Dan Support Vector Machine Di Twitter," *J. Karya Ilm. Multidisiplin*, vol. 5, no. 2, pp. 112–124, 2025.