

Analisis Komparatif Algoritma *Machine learning* Untuk Klasifikasi Sentimen Multi-Kelas Berita Online Multi-Domain Menggunakan TF-IDF

Daniel Christian Gunawan^{1*}, Albertus Dwiyo Widianoro¹

¹Program Studi Sistem Informasi, Fakultas Ilmu Komputer, Universitas Katolik Soegijapranata,
Semarang, Jawa Tengah, Indonesia

e-mail: 23.n1.0001@student.unika.ac.id, yoga@unika.ac.id

Diterima
04-08-2026

Direvisi
19-08-2026

Disetujui
20-08-2026

Abstract: The rapid expansion of online news platforms requires automated sentiment analysis across diverse domains. This study evaluates and compares four supervised machine learning classifiers Logistic Regression, Naïve Bayes, Random Forest, and Support Vector Machine (SVM) for multiclass sentiment classification on 3,500 English news articles across seven categories. The raw textual data underwent preprocessing (case folding, URL/HTML removal, punctuation/numeral removal, stopword filtering, and stemming), followed by TF-IDF vectorization and variance-based feature selection. The dataset was partitioned using an 80:20 stratified split, and models were evaluated using macro-averaged Accuracy, Precision, Recall, and F1-score to handle class imbalance. Among the evaluated algorithms, Random Forest achieved the highest classification performance with an overall Accuracy of 75.10% and a macro F1-score of 63.39%. These outcomes provide practical insights into selecting effective machine learning algorithms for automated multi-domain news sentiment classification.

Keywords: Sentiment Analysis, Machine Learning, Text Mining, TF-IDF, Random Forest

Abstrak: Pesatnya pertumbuhan platform berita online memerlukan analisis sentimen otomatis di berbagai domain. Penelitian ini mengevaluasi dan membandingkan empat algoritma supervised machine learning Logistic Regression, Naïve Bayes, Random Forest, dan Support Vector Machine (SVM) untuk klasifikasi sentimen multi-kelas pada 3.500 artikel berita dalam tujuh kategori. Data teks mentah melalui tahap preprocessing (case folding, pembersihan URL/HTML, penghapusan tanda baca/angka, filtering stopword, dan stemming), dilanjutkan dengan vektorisasi TF-IDF serta seleksi fitur berbasis varians. Dataset dibagi menggunakan stratified split 80:20, dan model dievaluasi menggunakan Akurasi, Presisi, Recall, dan F1-score rata-rata makro untuk menangani ketidakseimbangan kelas. Di antara algoritma yang diuji, Random Forest meraih kinerja klasifikasi tertinggi dengan akurasi keseluruhan sebesar 75,10% dan macro F1-score sebesar 63,39%. Hasil ini memberikan wawasan praktis dalam pemilihan algoritma machine learning yang efektif untuk klasifikasi sentimen berita online multi-domain secara otomatis.

Kata kunci: Analisis Sentimen, Machine Learning, Text Mining, TF-IDF, Random Forest

I. PENDAHULUAN

Kemajuan teknologi informasi secara berkelanjutan telah mengubah cara berita diproduksi, didistribusikan, dan dikonsumsi. Platform berita digital dan agregator telah mengukuhkan diri sebagai perantara utama bagi jutaan pembaca untuk mengakses informasi setiap hari, yang mencakup berbagai domain seperti bisnis, kesehatan, sains, olahraga, hiburan, teknologi, dan isu umum. Nada serta pembingkai yang tertuang dalam pemberitaan membawa pengaruh signifikan terhadap persepsi publik, sentimen pasar, dan wacana sosial, sehingga pemahaman sistematis mengenai sentimen dalam konten berita menjadi topik yang sangat penting baik secara praktis maupun akademis (Israt et al., 2024)(Harmandini, 2024).

Peningkatan volume konten berita digital dalam skala besar menimbulkan kesulitan nyata jika interpretasi sentimen dilakukan secara manual. Akumulasi data teks yang dinamis menyulitkan organisasi, jurnal, dan analis untuk mengekstraksi pola sentimen secara konsisten dan tepat waktu. Sebagai respons, analisis sentimen hadir sebagai salah satu teknik yang paling luas diterapkan dalam Natural Language Processing (NLP), menyediakan sarana otomatis untuk

mengeksktraksi informasi subjektif dari korpus teks. Tujuan utama analisis sentimen adalah menentukan apakah suatu teks menyampaikan nada positif, netral, atau negatif, sehingga mendukung pemantauan tren sentimen media secara terstruktur (Prastyo et al., 2024)(Paudel & Yogi, 2020).

Selama dekade terakhir, analisis sentimen telah menarik perhatian akademis yang luas karena penerapannya di berbagai bidang seperti pemantauan media sosial, *e-commerce*, wacana politik, dan studi media massa. Banyak penelitian terdahulu menunjukkan bahwa algoritma *machine learning* mencapai performa yang kompetitif pada tugas klasifikasi teks. Namun demikian, kinerja prediktif dari masing-masing algoritma bervariasi tergantung pada karakteristik dataset, prosedur pra-pemrosesan yang diterapkan, serta strategi representasi fitur yang dipilih (Das et al., 2023).

Aspek mendasar dari klasifikasi sentimen terletak pada proses ekstraksi fitur, yang bertanggung jawab mengubah dokumen teks mentah menjadi representasi numerik. Di antara berbagai metodologi ekstraksi fitur, skema pembobotan *Term Frequency Inverse Document Frequency* (TF-IDF) tetap menjadi salah satu pendekatan yang paling populer karena kesederhanaan komputasi, efisiensi proses, dan kemampuannya menonjolkan kata-kata yang informatif secara semantik (Harmandini, 2024)(Zhao et al., 2026). Selain itu, prosedur pra-pemrosesan teks seperti *case folding*, pembersihan tanda baca, penghapusan *stopword*, dan stemming merupakan tahapan krusial untuk meningkatkan kualitas data teks sebelum ekstraksi fitur (Aufar et al., 2023).

Sejumlah algoritma *supervised learning* telah dipelajari secara luas dalam riset klasifikasi sentimen. *Logistic Regression* diakui karena formulasi matematika yang efisien pada masalah klasifikasi linear (E. Safitri et al., 2025). *Naïve Bayes* banyak digunakan karena kerangka inferensi probabilitasnya yang sangat cepat untuk dataset berdimensi tinggi (Sitorus & Zufria, 2024). *Random Forest* secara konsisten menunjukkan performa unggul melalui mekanisme ensemble learning yang menggabungkan prediksi banyak *decision tree*, sehingga meningkatkan akurasi dan ketahanan terhadap ketidakseimbangan kelas (Jalal et al., 2022). Sementara itu, *Support Vector Machine* (SVM) terbukti efektif dalam klasifikasi teks dengan mengonstruksi hyperplane pemisah optimal dalam ruang fitur berdimensi tinggi (Widodo et al., 2025).

Meskipun telah banyak penelitian analisis sentimen pada media sosial dan ulasan produk, evaluasi komparatif yang melibatkan beberapa algoritma *machine learning* klasik pada artikel berita *online* multi-domain di bawah kondisi eksperimen yang terkontrol secara ketat masih terbatas (Zhang, 2025). Berdasarkan pertimbangan tersebut, penelitian ini melakukan analisis komparatif sistematis terhadap empat algoritma *supervised machine learning Logistic Regression, Naïve Bayes, Random Forest*, dan *Support Vector Machine* dalam klasifikasi sentimen multi-kelas artikel berita *online* multi-domain dengan skema ekstraksi fitur TF-IDF dan seleksi fitur berbasis varians (Bhalerao, 2021).

II. TUJUAN DAN KONTRIBUSI PENELITIAN

1. Tujuan Penelitian

Pelaksanaan penelitian ini bertujuan untuk membandingkan performa empat algoritma *machine learning* yaitu *Logistic Regression, Naïve Bayes, Random Forest*, dan *Support Vector Machine* dalam mengklasifikasikan sentimen berita ke dalam kategori positif, netral, dan negatif. Dari perbandingan tersebut, penelitian ini berupaya mengidentifikasi algoritma mana yang paling andal untuk diterapkan pada analisis sentimen berita multi-domain. Selain itu, penelitian ini juga dilakukan untuk mengevaluasi sejauh mana pengaruh penerapan ekstraksi fitur TF-IDF serta tahap pra-pemrosesan teks terhadap hasil akhir klasifikasi sentimen yang diperoleh.

2. Kontribusi Penelitian

Sementara itu, kontribusi utama dari penelitian ini terletak pada pengembangan alur analisis sentimen *end-to-end* yang disesuaikan secara khusus untuk data artikel berita *online* multi-domain menggunakan standar pemrosesan teks yang terstruktur. Penelitian ini juga menyediakan evaluasi komparatif empiris dari empat algoritma *supervised learning* pada kondisi

eksperimen yang seragam demi menjamin validitas hasil. Melalui pendekatan tersebut, penelitian ini mampu menentukan algoritma terbaik untuk klasifikasi sentimen berita berdasarkan berbagai metrik evaluasi kuantitatif.

III. METODE PENELITIAN

Penelitian ini menggunakan pendekatan eksperimental kuantitatif untuk mengevaluasi dan membandingkan kinerja prediktif dari empat algoritma *supervised machine learning*. Alur penelitian dirancang secara runtut yang mencakup pengumpulan dataset, pembersihan data, pra-pemrosesan teks, ekstraksi fitur TF-IDF, seleksi fitur, pembagian data, pelatihan model, dan evaluasi performa.

1. Kerangka Penelitian

Kerangka kerja penelitian ini disusun dalam 9 tahap utama secara sistematis:

- a. Pengumpulan Dataset Berita *Online (Dataset Collection)* Proses pengumpulan data primer atau sekunder berupa artikel berita digital dari berbagai domain (multi-domain) melalui teknik *web scraping* atau penggunaan *API news aggregator*. Tahap ini bertujuan untuk membentuk repositori data mentah yang merepresentasikan dinamika ragam opini dan sentimen publik.
- b. Pembersihan Data dari Duplikat dan Data Kosong (*Data Cleaning*) Tahap validasi kualitas data yang berfokus pada penghapusan baris data yang terduplikasi (*duplicate removal*), penanganan nilai yang hilang (*missing values/NaN*), serta penghapusan data yang tidak relevan. Langkah ini krusial untuk mencegah bias pelatihan model dan menjaga integritas hasil analisis.
- c. Pra-pemrosesan Teks (*Text Preprocessing*) Serangkaian operasi standar untuk mentransformasi teks mentah yang tidak terstruktur menjadi bentuk yang bersih dan siap diolah. Tahap ini umumnya mencakup *case folding* (penyeragaman huruf kecil), filtering (penghapusan tanda baca, angka, dan karakter khusus), *stopword removal* (penghapusan kata umum yang kurang bermakna), serta *stemming* atau *lemmatization* (pengembalian kata ke bentuk dasarnya).
- d. Ekstraksi Fitur TF-IDF (*TF-IDF Feature Extraction*) Metode pembobotan statistik *Term Frequency-Inverse Document Frequency* yang mengubah himpunan teks menjadi vektor numerik (*vectorization*). Teknik ini mengukur seberapa penting suatu kata dalam sebuah dokumen relatif terhadap seluruh korpus berita, sehingga mampu menangkap keunikan kata dalam mencerminkan sentimen.
- e. Seleksi Fitur Berbasis Varians (*Variance-Based Feature Selection*) Teknik pengurangan dimensi (*dimensionality reduction*) dengan cara mengeliminasi kata/fitur yang memiliki nilai varians di bawah ambang batas (*threshold*) tertentu (misalnya *VarianceThreshold*). Fitur dengan varians yang sangat rendah (seperti kata yang muncul pada hampir semua dokumen atau sangat jarang) dianggap tidak memiliki daya pembeda (*discriminative power*) yang signifikan untuk klasifikasi.
- f. Pembagian Data *Train-Test Split* 80:20 (*Stratified Split*) Prosedur pembagian dataset menjadi 80% data latih (*training set*) untuk melatih algoritma dan 20% data uji (*test set*) untuk menguji performa model. Penggunaan teknik *stratified* memastikan bahwa proporsi distribusi label kelas (positif, netral, negatif) pada data latih dan data uji tetap seimbang selaras dengan distribusi data aslinya.
- g. Pelatihan Model Klasifikasi (*Model Training: LR, NB, RF, SVM*) Proses pembentukan model prediktif dengan melatih empat algoritma *supervised learning* (*Logistic Regression, Naïve Bayes, Random Forest, dan Support Vector Machine*) menggunakan *training set*. Setiap algoritma mempelajari pola hubungan antara vektor fitur TF-IDF dengan label sentimen yang sesuai.
- h. Evaluasi Performa Model (*Performance Evaluation*) Tahap pengujian dan pengukuran kemampuan generalisasi model terhadap data uji (*test set*) yang belum pernah dilihat sebelumnya. Evaluasi dilakukan dengan menghitung metrik kuantitatif berbasis *Confusion Matrix*, seperti *Accuracy, Precision, Recall, dan F1-score*.

- i. Analisis Komparatif (*Comparative Analysis*) Proses perbandingan secara empiris dan mendalam terhadap hasil pengukuran performa dari keempat algoritma (Surya et al., 2026). Analisis ini bertujuan untuk mengidentifikasi algoritma terbaik (*best-performing model*), menganalisis kelebihan/kekurangan masing-masing metode pada teks berita multi-domain, serta menyimpulkan efektivitas alur pemrosesan yang diterapkan.

2. Dataset

Dataset yang digunakan mencakup 3.500 artikel berita berbahasa Inggris yang dikumpulkan dari berbagai media online dalam 7 kategori domain: Bisnis, Hiburan, Umum, Kesehatan, Sains, Olahraga, dan Teknologi. Untuk kebutuhan analisis sentimen, judul artikel dan deskripsi digabungkan menjadi dokumen teks utama. Setelah menghapus entri kosong dan 1.082 data duplikat eksak, sebanyak 2.418 artikel berita unik dipertahankan untuk tahap eksperimen.

3. Pra Pemrosesan Teks

Serangkaian teknik pra-pemrosesan diterapkan secara berurutan guna menghilangkan *noise* dan menstandarkan data teks:

- a. *Case Folding*: Mengubah seluruh huruf menjadi huruf kecil (*lowercase*).
- b. *URL & HTML Removal*: Menghapus link web serta entitas HTML.
- c. *Mention & Punctuation Removal*: Menghapus simbol mention, tanda baca, serta karakter angka.
- d. *Tokenization & Stopword Removal*: Memotong kalimat menjadi token kata dan memfilter kata-kata umum yang kurang informatif.
- e. *Stemming*: Mengubah kata berimbuhan menjadi kata dasar menggunakan algoritma *Porter Stemmer*.

4. Ekstraksi dan Seleksi Fitur

Setelah pra-pemrosesan, dokumen dikonversi menjadi matriks numerik menggunakan TF-IDF. Korpus awal menghasilkan 15.930 fitur unik. *Trimming* frekuensi dokumen diterapkan untuk membuang istilah yang muncul kurang dari 5 kali, sehingga mengurangi dimensi menjadi 3.711 fitur. Selanjutnya, seleksi fitur berbasis *varians* memilih 1.500 fitur dengan *varians* tertinggi untuk digunakan dalam pelatihan model.

5. Pembagian Dataset dan Pelatihan Model

Dataset dibagi dengan teknik *Stratified Hold-Out Split* 80:20 (1.936 data latih dan 482 data uji) dengan *seed* acak bernilai 123. Empat model *machine learning* diimplementasikan sebagai berikut:

- a. *Logistic Regression*: Menggunakan fungsi multinomial untuk mendukung multikelas sentimen.
- b. *Naïve Bayes*: Menggunakan estimasi probabilitas posterior *Bayes* dengan *Laplace smoothing*.
- c. *Random Forest*: Membangun ensemble dari 200 *decision trees* dengan mayoritas suara (*majority voting*).
- d. *Support Vector Machine*: Menggunakan konfigurasi kernel linear untuk mencari *hyperplane* pemisah optimal.

6. Matriks Evaluasi

Kinerja model dievaluasi berdasarkan *Confusion Matrix* dengan menghitung Akurasi, Presisi, *Recall*, dan *F1-score* secara *macro-averaging* untuk memperhitungkan ketidakseimbangan kelas pada dataset.

IV. HASIL DAN PEMBAHASAN

Proses pembersihan data menghasilkan 2.418 dokumen artikel berita valid. Distribusi kelas sentimen ditunjukkan pada Tabel 1.

Table 1. Distribusi Label Sentimen

Kelas Sentimen	Jumlah Artikel
Positif	1507
Netral	514
Negatif	397
Total	2418

Hasil Pra-pemrosesan Teks dan Reduksi Fitur Contoh hasil pra-pemrosesan teks disajikan pada Tabel 2, sedangkan tahapan reduksi dimensi fitur disajikan pada Tabel 3.

Table 2. Contoh Hasil Pra-pemrosesan Teks

Teks Teks Asli (Original Text)	Teks Hasil <i>Preprocessing</i>
Watch: Travis busk outside the Barrowlands on bus tour of Glasgow. Band return home to ...	watch travis busk outsid barrowland bu tour glasgow band return home pr...
Life Sciences Firms Adopt Tools to Modernize Clinical Trials. Life Sciences Firms Adopt...	life scienc firm adopt tool modern clinic trial life scienc firm adopt...
Falcons receiver ready to breakout in new offense according to CBS Sports. Falcons recei...	falcon receiv readi breakout new offens accord cb sport falcon recei...

Table 3. Proses Reduksi Dimensi Fitur

Tahapan Proses	Jumlah Fitur (<i>Term</i>)
Kosakata Awal (Initial <i>Vocabulary</i>)	15930
Setelah <i>Trimming</i> Kata Jarang ($freq < 5$)	3711
Fitur Terpilih Akhir (Varians Tertinggi)	1500

Hasil Evaluasi Performa Model dan Analisis Komparatif merupakan hasil pengujian keempat model *machine learning* pada dataset uji (482 artikel) menggunakan empat metrik evaluasi berskala rata-rata makro (*macro-average*) dirangkum pada Tabel 4.

Table 4. Perbandingan Performa Model *Machine Learning*

Model Algoritma	<i>Accuracy</i>	<i>Precision</i>	<i>Recall</i>	<i>F1-score</i>
Logistic Regression	50,00%	44,61%	46,97%	44,65%
<i>Naïve Bayes</i>	26,76%	26,05%	46,15%	37,74%
<i>Random Forest</i>	75,10%	77,55%	60,98%	63,39%
<i>Support Vector Machine</i>	64,52%	55,95%	56,40%	56,17%

Berdasarkan hasil pengujian pada Tabel 4, *Random Forest* menunjukkan performa terbaik secara menyeluruh dengan akurasi 75,10%, presisi 77,55%, recall 60,98%, dan macro *F1-score* sebesar 63,39%. Keunggulan *Random Forest* berasal dari mekanisme *ensemble learning* yang mampu meminimalkan varians serta menangani *imbalance* data secara lebih adaptif dibandingkan model tunggal (E. D. Safitri et al., 2026).

Support Vector Machine berada di posisi kedua dengan akurasi 64,52% dan *F1-score* 56,17%, mengonfirmasi keandalannya dalam mengolah ruang fitur berdimensi tinggi yang bersifat *sparse*. *Logistic Regression* mencatatkan akurasi 50,00% dan *F1-score* 44,65%, cenderung memprediksi kelas mayoritas (positif) dengan baik tetapi kesulitan membedakan kelas netral dan negatif akibat ambiguitas kata. Sementara itu, *Naïve Bayes* menghasilkan performa terendah dengan akurasi 26,76% dan *F1-score* 37,74%, yang disebabkan oleh asumsi independensi bersyarat yang kurang sesuai untuk hubungan antar-kata kompleks dalam artikel berita (Sitorus et al., 2024).

V.CONCLUSION

Penelitian ini telah berhasil menganalisis dan membandingkan performa empat algoritma supervised *machine learning* yaitu *Logistic Regression*, *Naïve Bayes*, *Random Forest*, dan *Support Vector Machine* dalam mengklasifikasikan sentimen multi-kelas pada 2.418 artikel berita *online* multi-domain. Hasil eksperimen menunjukkan bahwa pemilihan algoritma memberikan pengaruh yang signifikan terhadap akurasi dan keandalan prediksi. Dalam hal ini, algoritma *Random Forest* terbukti menjadi model terbaik dengan capaian Akurasi sebesar 75,10% dan

Macro *F1-score* sebesar 63,39%. Untuk pengembangan penelitian selanjutnya, disarankan agar menggunakan teknik representasi kata berbasis konteks seperti *Word2Vec*, *FastText*, atau model *pretrained* transformer (seperti *BERT* dan *RoBERTa*). Selain itu, penerapan teknik penanganan ketidakseimbangan kelas seperti *SMOTE* atau *cost-sensitive learning* sangat direkomendasikan guna meningkatkan nilai *Recall* pada kelas minoritas, khususnya untuk kategori sentimen negatif dan netral.

REFERENSI

- Aufar, A. F., Rosid, M. A., Eviyanti, A., & Astutik, I. R. I. (2023). Optimizing Text Preprocessing for Accurate Sentiment Analysis on E-Wallet Reviews: Mengoptimalkan Preprocessing Teks untuk Analisis Sentimen yang Akurat pada Ulasan E-Wallet. *JICTE (Journal of Information and Computer Technology Education)*, 7(2), 42–50.
- Bhalerao, K. V. (2021). *Customer Reviews Sentiment Analysis: A hybrid technique of Lexicon and Machine Learning based Classification model (SVM, NB, Logistic Regression)*. Dublin, National College of Ireland.
- Das, R. K., Islam, M., Hasan, M. M., Razia, S., Hassan, M., & Khushbu, S. A. (2023). Sentiment analysis in multilingual context: Comparative analysis of machine learning and hybrid deep learning models. *Heliyon*, 9(9).
- Harmandini, K. P. (2024). Analysis of TF-IDF and TF-RF feature extraction on product review sentiment. *Sinkron: Jurnal Dan Penelitian Teknik Informatika*, 8(2), 929–937.
- Israt, J., Islam, M. N., Hasan, M. M., & Siddiky, M. R. (2024). Comparative analysis of machine learning algorithms for sentiment classification in social media text. *World Journal of Advanced Research and Reviews*, 23(3), 2842–2852.
- Jalal, N., Mehmood, A., Choi, G. S., & Ashraf, I. (2022). A novel improved random forest for text classification using feature ranking and optimal number of trees. *Journal of King Saud University-Computer and Information Sciences*, 34(6), 2733–2742.
- Paudel, N., & Yogi, T. N. (2020). Comparative study of machine learning algorithms for rainfall prediction-a case study in Nepal. *International Journal of Advanced Research in Engineering and Technology (IJARET)*, 11(10), 1582–1591.
- Prastyo, D., Irawan, D., & Mursyidin, I. H. (2024). Klasifikasi sentimen komentar YouTube dengan NLP pada debat Pilkada Banten 2024. *Bit-Tech*, 7(2), 413–421.
- Safitri, E. D., Ichisasmita, L. H., Agustin, C., Muthoharoh, L., Satria, A., & Manullang, M. C. T. (2026). A Comparative Analysis of Classical Machine Learning and Deep Learning Approaches for Sentiment Classification on IMDb Movie Reviews. *ArXiv Preprint ArXiv:2605.07811*.
- Safitri, E., Syukrilla, W. A., & Fitriana, I. N. L. (2025). Logistic regression for sentiment analysis of insecurity phenomena on platform x. *J Statistika: Jurnal Ilmiah Teori Dan Aplikasi Statistika*, 18(1), 948–956.
- Sitorus, R. A., & Zufria, I. (2024). Application of the naïve Bayes algorithm in sentiment analysis of using the Shopee application on the Play Store. *Digital Zone: Jurnal Teknologi Informasi Dan Komunikasi*, 15(1), 53–66.
- Surya, L. F. Y., Valentina, L., Firmansyah, Z., Fathoni, F., & Ibrahim, A. (2026). Analisis Komparatif Algoritma Process Mining untuk Pemetaan Navigasi dan Deteksi Bottleneck E-Commerce. *Jutisi: Jurnal Ilmiah Teknik Informatika Dan Sistem Informasi*, 15(3), 861–874.
- Widodo, A., Herlambang, B. A., & Renaldy, R. (2025). Optimizing Support Vector Machine (SVM) for Sentiment Analysis of Blu by BCA Reviews with Chi-Square. *Journal of Applied Informatics and Computing*, 9(5), 2588–2597.

Zhang, L. (2025). Features extraction based on Naive Bayes algorithm and TF-IDF for news classification. *PLoS One*, 20(7), e0327347.

Zhao, C., Xi, X., Kang, L., & Bai, J. (2026). A weakly supervised SMOTE-XGBoost framework with sliding-window monitoring for temporal sentiment drift detection. *Knowledge and Information Systems*, 68(1), 186.