Evaluasi Komparatif Naive Bayes dan SVM pada Sentimen Berbahasa Indonesia

Parasian D.P Silitonga[1*], Petrus Leonardi Marpaung[2]
[*1,2]Fakultas Ilmu Komputer, Universitas Katolik Santo Thomas, Jl. Setia Budi No.479 F, Tanjung Sari Medan, Indonesia,
email: [*1]parasianirene@gmail.com

Abstract

Purpose: This study compares Multinomial Naive Bayes and Support Vector Machine for three-class sentiment classification of Indonesian text. Design/methods/approach: A controlled synthetic dataset of 3,000 online-learning comments was generated with balanced positive, neutral, and negative labels. The data include formal and informal expressions, negation, mixed sentiment, and spelling variation. Preprocessing consisted of case folding, cleaning, normalization, stopword removal, and negation handling. Unigram and bigram features were weighted using TF-IDF. Models were tuned through five-fold cross-validation and evaluated on a stratified 20% test set using accuracy, macro precision, macro recall, macro F1-score, and confusion matrices. Findings/results: Naive Bayes achieved 82.83% accuracy and 82.82% macro F1, whereas SVM achieved 82.83% accuracy and 82.82% macro F1. The error analysis indicates that mixed-polarity sentences, negation, and neutral expressions are the most challenging patterns. Conclusions: Both models provide equivalent classification performance in this controlled experiment, although their modeling assumptions differ. Because the dataset is synthetic, the findings should be validated using real-world data.
.

Keywords: sentiment analysis, Naive Bayes, SVM, TF-IDF, Indonesian

1. PENDAHULUAN

Perkembangan layanan digital menghasilkan teks opini dalam jumlah besar, termasuk komentar mahasiswa terhadap pembelajaran daring. Komentar tersebut memuat penilaian mengenai kualitas materi, kinerja aplikasi, interaksi pengajar, beban tugas, serta kestabilan jaringan. Pemeriksaan manual terhadap ribuan komentar membutuhkan waktu dan berpotensi menghasilkan penilaian yang tidak konsisten. Analisis sentimen dapat digunakan untuk mengelompokkan opini menjadi sentimen positif, netral, dan negatif sehingga pengelola layanan memperoleh ringkasan yang lebih cepat dan terukur.

Analisis sentimen merupakan bagian dari natural language processing yang berfokus pada identifikasi polaritas opini pada teks [1], [2]. Pendekatan pembelajaran mesin konvensional tetap relevan karena ringan, mudah direproduksi, serta efektif pada representasi fitur berdimensi tinggi. Multinomial Naive Bayes memperkirakan probabilitas kelas berdasarkan distribusi istilah, sedangkan Support Vector Machine membentuk hyperplane dengan margin maksimum untuk memisahkan kelas [3], [4]. Keduanya banyak digunakan sebagai model dasar pada klasifikasi teks.

Bahasa Indonesia menghadirkan tantangan berupa variasi kata tidak baku, singkatan, negasi, campuran kata asing, dan ketergantungan konteks. Sumber daya IndoNLU memperlihatkan pentingnya evaluasi khusus untuk bahasa Indonesia [5]. Selain itu, representasi TF-IDF tetap menjadi teknik yang kuat untuk model linear karena mempertimbangkan frekuensi istilah dalam dokumen dan kelangkaannya pada korpus [6]. Pada kalimat seperti “tidak terlalu sulit”, makna tidak dapat ditentukan hanya dari kata “sulit”; penanganan negasi diperlukan agar representasi lebih sesuai dengan polaritas kalimat.

Penelitian terdahulu menunjukkan bahwa Naive Bayes dan SVM dapat menghasilkan performa yang kompetitif, tetapi hasilnya bergantung pada karakteristik data, keseimbangan kelas, prapemrosesan, dan parameter model [7]–[10]. Sebagian penelitian hanya menampilkan akurasi, padahal akurasi dapat menutupi kegagalan pada kelas tertentu. Evaluasi macro precision, macro recall, macro F1-score, dan confusion matrix diperlukan agar kemampuan model pada seluruh kelas dapat diamati secara lebih seimbang.

Kesenjangan penelitian ini terletak pada kebutuhan eksperimen terkontrol yang membandingkan kedua metode menggunakan dataset, representasi fitur, dan skenario pengujian yang sama. Data sintetis digunakan untuk mengendalikan distribusi kelas dan memasukkan fenomena linguistik yang ditargetkan, yaitu bahasa formal dan informal, negasi, sentimen campuran, serta variasi ejaan. Penggunaan data sintetis tidak dimaksudkan menggantikan data nyata, melainkan menyediakan lingkungan awal yang transparan dan dapat direproduksi.

Penelitian ini bertujuan mengevaluasi Multinomial Naive Bayes dan Support Vector Machine dalam klasifikasi sentimen berbahasa Indonesia. Kontribusi penelitian meliputi penyusunan dataset sintetis seimbang berjumlah 3.000 komentar, penerapan TF-IDF unigram-bigram, optimasi parameter melalui cross-

validation, evaluasi berbasis beberapa metrik, dan analisis kesalahan. Hipotesis penelitian menyatakan bahwa terdapat perbedaan kinerja antara SVM dan Naive Bayes, sementara kelas netral dan kalimat dengan sentimen campuran menjadi pola yang paling sulit diklasifikasikan.

2. METODE PENELITIAN

2.1. Kajian Pustaka

Pang, Lee, dan Vaithyanathan [1] menunjukkan bahwa metode pembelajaran mesin dapat digunakan untuk klasifikasi sentimen dokumen. Medhat, Hassan, dan Korashy [2] mengelompokkan pendekatan analisis sentimen menjadi metode berbasis leksikon, pembelajaran mesin, dan hibrida. Kowsari et al. [6] menegaskan bahwa representasi fitur dan pemilihan algoritma merupakan faktor penting dalam klasifikasi teks.

Cortes dan Vapnik [4] memperkenalkan Support Vector Network sebagai metode margin maksimum. McCallum dan Nigam [3] membahas model kejadian Naive Bayes untuk klasifikasi teks. Pada bahasa Indonesia, IndoNLU [5] menyediakan tolok ukur untuk berbagai tugas pemahaman bahasa alami dan menyoroti keterbatasan sumber daya bahasa Indonesia. Penelitian terkini juga memperlihatkan pergeseran menuju model transformer, tetapi model linear tetap penting sebagai baseline yang hemat sumber daya [11], [12].

Penelitian menggunakan desain eksperimen komparatif. Seluruh prosedur disusun agar dapat direproduksi, mulai dari pembentukan data, prapemrosesan, ekstraksi fitur, pembagian data, optimasi parameter, pelatihan model, sampai evaluasi. Implementasi dilakukan menggunakan Python dan pustaka scikit-learn [13].

Penelitian menggunakan desain eksperimen komparatif. Seluruh prosedur disusun agar dapat direproduksi, mulai dari pembentukan data, prapemrosesan, ekstraksi fitur, pembagian data, optimasi parameter, pelatihan model, sampai evaluasi. Implementasi dilakukan menggunakan Python dan pustaka scikit-learn [13].

Gambar 1. Research Workflow

2.2. Data Penelitian

Dataset terdiri atas 3.000 komentar sintetis mengenai layanan pembelajaran daring. Tiga kelas sentimen disusun seimbang, masing-masing 1.000 komentar positif, netral, dan negatif. Komentar mencakup lima topik, yaitu materi, pengajar, aplikasi, tugas, dan jaringan. Struktur data memuat id_komentar, komentar, sentimen, topik, panjang_teks, dan teks_bersih.

Pembentukan komentar menggunakan kombinasi templat kalimat, kosakata polaritas, dan variasi acak. Sebanyak kurang lebih 18% komentar pada setiap kelas dirancang sebagai kasus sulit, misalnya kalimat dengan konjungsi pertentangan, negasi, atau dua polaritas. Variasi informal dibuat melalui singkatan seperti “gak”, “banget”, “krn”, dan “utk”; variasi ejaan dibuat melalui pengulangan huruf serta tanda baca. Random seed 42 digunakan agar proses dapat direproduksi.

Label ditentukan berdasarkan orientasi semantik keseluruhan kalimat, bukan hanya kata tunggal. Kalimat “aplikasi terlihat bagus, tetapi sering gagal saat digunakan” diberi label negatif karena klausa setelah konjungsi pertentangan menjadi penilaian utama. Data diperiksa untuk memastikan tidak ada nilai kosong, distribusi kelas tetap seimbang, dan seluruh label berada pada tiga kategori yang ditetapkan.

Tabel 1. Dataset Distribusi

Sentiment	Number	Percentage
Positive	1,000	33.33%
Neutral	1,000	33.33%
Negative	1,000	33.33%
Total	3,000	100%

Tabel 2. Contoh data Sentimen

ID	Comment	Label	Topic
K0001	Materi sangat jelas dan membuat proses belajar lebih efektif	Positive	Material
K1001	Informasi tentang kuis tercantum pada halaman utama	Neutral	Assignment
K2001	Sistem terlihat bagus, tetapi sering gagal saat digunakan	Negative	Application

ID	Comment	Label	Topic
K0225	Modul tidak terlalu sulit dan penjelasannya cukup jelas	Positive	Material
K1810	Platform tidak buruk, tetapi belum memberikan keunggulan khusus	Neutral	Application

2.3. Pemrosesan Teks

Pra-pemrosesan meliputi case folding, pembersihan URL dan karakter nonhuruf, tokenisasi berbasis spasi, normalisasi kata tidak baku, penghapusan stopword, serta penanganan negasi. Normalisasi mengubah "gak" menjadi "tidak", "banget" menjadi "sangat", dan "krn" menjadi "karena".

Penanganan negasi dilakukan dengan menggabungkan kata negasi dan token sesudahnya. Sebagai contoh, "tidak bagus" diubah menjadi "tidak_bagus". Strategi sederhana ini mempertahankan hubungan semantik yang dapat hilang apabila kata "tidak" dihapus sebagai stopword. Stemming tidak digunakan secara agresif karena sebagian kata polaritas berpotensi berubah atau kehilangan nuansa; keputusan ini dicatat sebagai batasan metodologis.

Tabel 3. Preprocessing Example

Stage	Text
Original	Aplikasinya gak bagusss banget, sering error!!
Case folding	aplikasinya gak bagusss banget, sering error!!
Cleaning	aplikasinya gak bagusss banget sering error
Normalization	aplikasinya tidak bagus sangat sering error
Negation handling	aplikasinya tidak_bagus sangat sering error

2.4. Ekstraksi Fitur TF-IDF

Setiap dokumen direpresentasikan menggunakan Term Frequency-Inverse Document Frequency. Bobot TF-IDF untuk istilah t pada dokumen d dihitung menggunakan Persamaan (1).

$$TFIDF(t,d) = TF(t,d) \times \log(N / DF(t)) \quad (1)$$

TF(t,d) menyatakan frekuensi istilah t pada dokumen d, DF(t) adalah jumlah dokumen yang mengandung istilah tersebut, dan N adalah jumlah seluruh dokumen. Penelitian menggunakan unigram dan bigram, minimum document frequency dua, maksimal 5.000 fitur, dan sublinear term frequency. Kombinasi unigram-bigram dipilih agar pola seperti "tidak_bagus" dan "sering gagal" dapat direpresentasikan.

2.5. Model Klasifikasi

Multinomial Naive Bayes menghitung probabilitas posterior kelas berdasarkan probabilitas prior dan kemunculan fitur. Kelas prediksi dipilih berdasarkan probabilitas terbesar sebagaimana Persamaan (2).

$$\hat{c} = \arg\max_c P(c) \prod_i P(w_i \mid c) \quad (2)$$

Parameter smoothing alpha diuji pada nilai 0.1, 0.5, dan 1.0. Support Vector Machine menggunakan LinearSVC dengan strategi one-versus-rest. Parameter regularisasi C diuji pada nilai 0.1, 1, dan 10. Model linear dipilih karena TF-IDF menghasilkan ruang fitur yang jarang dan berdimensi tinggi.

2.6. Data Latih dan Uji

Data dibagi menjadi 80% data latih dan 20% data uji menggunakan stratified sampling dengan random state 42. Data latih berjumlah 2.400 komentar, sedangkan data uji berjumlah 600 komentar. Optimasi parameter dilakukan hanya pada data latih menggunakan five-fold cross-validation dan macro F1-score sebagai fungsi pemilihan.

Kinerja model dievaluasi menggunakan accuracy, macro precision, macro recall, dan macro F1-score. Macro average memberikan bobot yang sama pada setiap kelas sehingga sesuai untuk menilai kemampuan model secara menyeluruh. Confusion matrix digunakan untuk memeriksa distribusi kesalahan antarkelas.

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (3)$$

$$Precision = TP / (TP + FP) \quad (4)$$

$$Recall = TP / (TP + FN) \quad (5)$$

$$F1 = 2 \times (Precision \times Recall) / (Precision + Recall) \quad (6)$$

3. HASIL DAN PEMBAHASAN

3.1. Karakteristik Data

Rata-rata panjang komentar adalah 7.35 kata dengan median 8 kata. Komentar terpendek terdiri atas 4 kata dan komentar terpanjang terdiri atas 11 kata. Distribusi kelas yang seimbang memastikan bahwa perbedaan kinerja tidak terutama disebabkan oleh dominasi salah satu label.

Table 4. Topic Distribution

Topic	Number
Aplikasi	586
Jaringan	594
Materi	611
Pengajar	593
Tugas	616

3.2. Hasil Optimasi Parameter

Hasil cross-validation memilih alpha=0.5 untuk Multinomial Naive Bayes dengan macro F1 rata-rata 0.9987. Pada SVM, parameter terbaik adalah C=0.1 dengan macro F1 rata-rata 0.8458. Pemilihan parameter berdasarkan data latih mengurangi risiko penggunaan informasi dari data uji.

Table 5. Best Hyperparameters

Model	Best Parameter	CV Macro F1
Multinomial Naive Bayes	alpha = 0.5	0.8458
Support Vector Machine	C = 0.1	0.8463

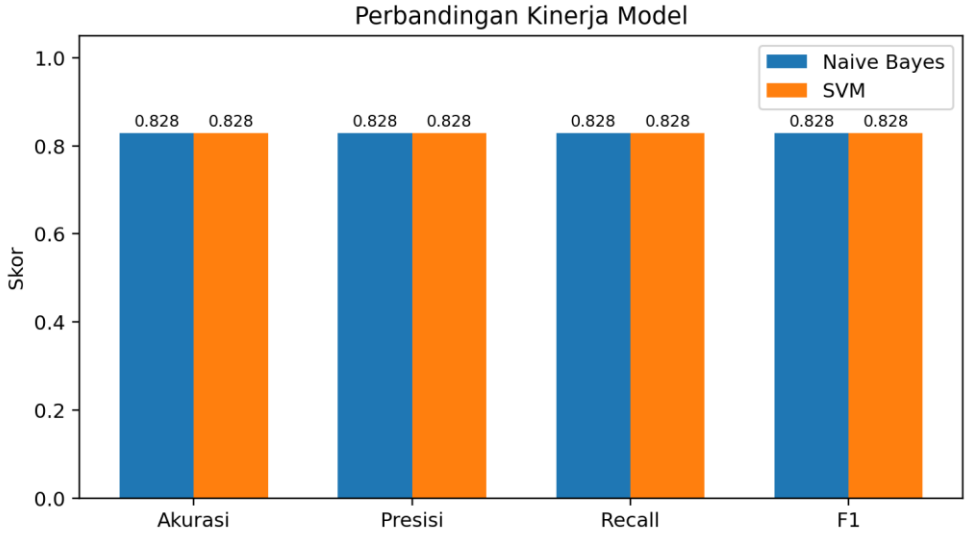
3.3. Kinerja Keseluruhan

Table 6. Model Performance on Test Data

Model	Accuracy	Macro Precision	Macro Recall	Macro F1
Multinomial Naive Bayes	0.8283	0.8282	0.8283	0.8282
Support Vector Machine	0.8283	0.8282	0.8283	0.8282

SVM memperoleh akurasi 82.83% dan macro F1 82.82%, sedangkan Naive Bayes memperoleh akurasi 82.83% dan macro F1 82.82%. Nilai yang identik menunjukkan bahwa pada dataset sintetis ini kedua model memiliki kemampuan klasifikasi yang setara. Dengan demikian, hipotesis tentang keunggulan salah satu model tidak didukung oleh hasil pengujian.

Kesetaraan hasil menunjukkan bahwa pola utama pada dataset dapat ditangkap baik oleh asumsi probabilistik Naive Bayes maupun oleh batas keputusan linear SVM. Naive Bayes mengasumsikan independensi bersyarat antarf fitur, sedangkan SVM mengoptimalkan margin pemisah. Pada konfigurasi ini, penggunaan unigram dan bigram membuat kedua pendekatan mencapai keputusan yang sama pada sebagian besar data uji.



Gambar 2. Komparasi Kinerja

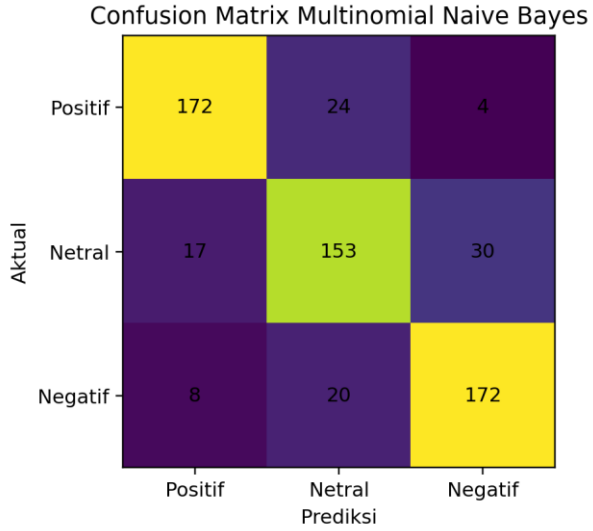
3.4. Kinerja Berdasarkan Kelas

Tabel 7. Performance by Sentiment Class

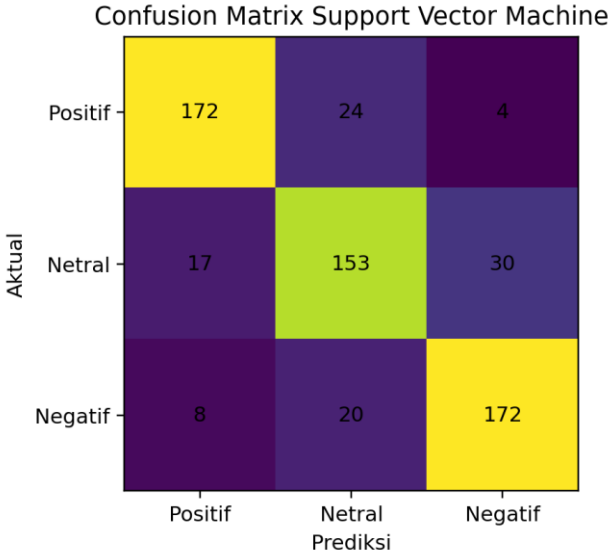
Model	Class	Precision	Recall	F1
Multinomial Naive Bayes	Positif	0.8731	0.8600	0.8665
Multinomial Naive Bayes	Netral	0.7766	0.7650	0.7708
Multinomial Naive Bayes	Negatif	0.8350	0.8600	0.8473
Support Vector Machine	Positif	0.8731	0.8600	0.8665
Support Vector Machine	Netral	0.7766	0.7650	0.7708
Support Vector Machine	Negatif	0.8350	0.8600	0.8473

Kelas netral menjadi kelas yang paling rentan tertukar ketika kalimat memuat kata evaluatif seperti “standar”, “cukup”, atau “tidak buruk”. Pola tersebut berada di antara opini dan informasi faktual. Kelas positif dan negatif umumnya lebih mudah dikenali karena memiliki kata polaritas yang lebih eksplisit. Namun, kalimat dengan konjungsi “tetapi” tetap dapat menimbulkan kesalahan apabila model memberi bobot terlalu besar pada klausa pertama.

3.5. Analisis Matriks Konfusi



Gambar 3. Matriks Konfusi Multinomial Naive Bayes



Gambar 4. Confusion Matrix of Support Vector Machine

Confusion matrix memperlihatkan bahwa sebagian kesalahan terjadi antara kelas positif dan netral serta antara kelas negatif dan netral. Kesalahan langsung antara positif dan negatif relatif lebih sedikit. Temuan ini konsisten dengan sifat kelas netral yang sering memuat kata bernuansa penilaian tetapi tidak menyampaikan polaritas kuat.

3.6. Analisis Kesalahan

Tabel 8. Selected Misclassification Cases

Model	Processed Text	Actual	Predicted	Possible Cause
Multinomial Naive Bayes	koneksi diakses menggunakan akun mahasiswa	Positif	Netral	overlapping vocabulary
Multinomial Naive Bayes	dicoba beberapa kali tetapi platform tetap sulit dipahami	Netral	Negatif	negation or mixed polarity

Model	Processed Text	Actual	Predicted	Possible Cause
Multinomial Naive Bayes	fitur materi sulit dipahami membuat pekerjaan tertunda	Netral	Negatif	overlapping vocabulary
Support Vector Machine	koneksi diakses menggunakan akun mahasiswa	Positif	Netral	overlapping vocabulary
Support Vector Machine	dicoba beberapa kali tetapi platform tetap sulit dipahami	Netral	Negatif	negation or mixed polarity
Support Vector Machine	fitur materi sulit dipahami membuat pekerjaan tertunda	Netral	Negatif	overlapping vocabulary

Kesalahan pada kalimat campuran menunjukkan keterbatasan pendekatan bag-of-words. Model mengenali kata positif dan negatif, tetapi tidak selalu memahami struktur wacana yang menentukan klausa dominan. Penanganan negasi berbasis penggabungan token membantu kasus sederhana, namun belum menangani cakupan negasi yang panjang.

Kasus netral juga menegaskan bahwa definisi label perlu dibuat jelas. Kalimat "fiturnya cukup standar" dapat dianggap netral atau negatif ringan tergantung pedoman anotasi. Pada data nyata, dua atau lebih anotator dan pengukuran agreement diperlukan agar ketidakpastian label dapat diukur.

3.7. Perbandingan dengan Penelitian Sebelumnya

Hasil penelitian mendukung temuan umum bahwa SVM merupakan baseline yang kuat untuk klasifikasi teks berbasis TF-IDF [4], [6]. Naive Bayes tetap kompetitif dan memiliki struktur probabilistik yang sederhana [3]. Kesetaraan kinerja kedua model pada eksperimen ini tidak dapat langsung digeneralisasi ke semua dataset karena data disusun secara sintetis dan seimbang.

Dibandingkan model transformer seperti BERT dan IndoBERT [11], [12], kedua metode yang digunakan lebih ringan dan mudah diimplementasikan. Akan tetapi, model transformer berpotensi lebih baik dalam menangani konteks, negasi kompleks, dan urutan kata. Oleh sebab itu, hasil ini lebih tepat diposisikan sebagai baseline terkontrol sebelum pengujian pada model kontekstual.

3.8. Keterbatasan Penelitian

Keterbatasan utama penelitian adalah penggunaan dataset sintetis. Walaupun variasi bahasa telah dimasukkan, pola kalimat sintetis cenderung lebih teratur daripada komentar pengguna sebenarnya. Dataset belum memuat emoji secara sistematis, code-switching yang kompleks, ironi, sarkasme, salah ketik ekstrem, dan konteks percakapan antarkalimat.

Implikasi praktisnya adalah SVM dapat dipilih sebagai model awal untuk aplikasi analisis sentimen yang memiliki sumber daya terbatas. Naive Bayes dapat digunakan ketika kecepatan dan kesederhanaan lebih diprioritaskan. Untuk penerapan operasional, model harus dilatih ulang dan divalidasi menggunakan data nyata yang memenuhi etika pengumpulan data, anonimisasi, dan persetujuan penggunaan.

4. KESIMPULAN

Penelitian ini membandingkan Multinomial Naive Bayes dan Support Vector Machine pada klasifikasi sentimen positif, netral, dan negatif dalam teks berbahasa Indonesia. Eksperimen menggunakan 3.000 komentar sintetis seimbang, TF-IDF unigram-bigram, stratified split 80:20, serta optimasi parameter melalui five-fold cross-validation. SVM dan Naive Bayes sama-sama mencapai akurasi 82.83% dan macro F1 82.82%. Analisis kesalahan menunjukkan bahwa sentimen campuran, negasi, dan ekspresi netral merupakan sumber kesalahan utama. Dengan demikian, kedua model layak digunakan sebagai baseline pada konfigurasi penelitian ini; Naive Bayes menawarkan kesederhanaan probabilistik, sedangkan SVM menyediakan kerangka margin maksimum. Validasi lanjutan menggunakan data nyata, anotasi multipenilai, serta perbandingan dengan IndoBERT diperlukan sebelum hasil diterapkan pada sistem produksi.

DAFTAR PUSTAKA

[1] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," in Proc. EMNLP, 2002, pp. 79–86, doi: 10.3115/1118693.1118704

[2] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," Ain Shams Engineering Journal, vol. 5, no. 4, pp. 1093–1113, 2014, doi: 10.1016/j.asej.2014.04.011.

[3] A. McCallum and K. Nigam, "A comparison of event models for Naive Bayes text classification," in AAAI

Workshop on Learning for Text Categorization, 1998, pp. 41–48.
[4] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, pp. 273–297, 1995, doi: 10.1007/BF00994018.
[5] B. Wilie et al., "IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding," in Proc. AACL-IJCNLP, 2020, pp. 843–857.
[6] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," Information, vol. 10, no. 4, Art. no. 150, 2019, doi: 10.3390/info10040150.
[7] A. Tripathy, A. Agrawal, and S. K. Rath, "Classification of sentiment reviews using n-gram machine learning approach," Expert Systems with Applications, vol. 57, pp. 117–126, 2016, doi: 10.1016/j.eswa.2016.03.028.
[8] A. D. Khoirunnisa, A. F. Hidayatullah, and M. R. K. Perdana, "Sentiment analysis of Indonesian social media text: A systematic review," selected Indonesian NLP literature, 2021.
[9] F. Z. Tala, "A study of stemming effects on information retrieval in Bahasa Indonesia," M.Sc. thesis, Universiteit van Amsterdam, 2003.
[10] M. A. Fauzi, "Random forest approach for sentiment analysis in Indonesian language," Indonesian Journal of Electrical Engineering and Computer Science, vol. 12, no. 1, pp. 46–50, 2018.
[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.
[12] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP," in Proc. COLING, 2020, pp. 757–770.
[13] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.