Management of Information System Journal
Vol 4, No 3, Juli 2026, Hal. 151 - 159
ISSN: 2964-9455 (media online)
DOI: 10.47065/mis.v2i3.2783
https://journal.fkpt.org/index.php/MIS

Analisis Sentimen Ulasan Aplikasi Mobile Legends Berbasis Pelabelan IndoBERT dan SMOTE dengan Komparasi Algoritma Klasifikasi Naive Bayes dan SVM

Muhammad Fauzan Aditiya Mufid[1], Erna Daniati[2], Arie Nugroho[3]
[123] Fakultas Teknik dan Ilmu Komputer, Program Studi Sistem Informasi, Universitas Nusantara PGRI Kediri, Kota Kediri, Indonesia
Email: [1]faizaditiya1915@email.com, [2]ernadaniati@unpkediri.ac.id, [3]arienugroho@unpkediri.ac.id
Email Penulis Korespondensi: ernadaniati@unpkediri.ac.id

Abstrak− Studi ini dilakukan guna mengevaluasi kecenderungan sentimen pengguna terhadap aplikasi permainan Mobile Legends: Bang Bang melalui ulasan yang dipublikasikan pada platform Google Play Store. Penelitian ini menerapkan pelabelan otomatis berbasis model IndoBERT untuk menghasilkan label sentimen yang lebih merepresentasikan makna tekstual ulasan. Ketidakseimbangan distribusi kelas yaang muncul akibat proses pelabelan tersebut ditangani menggunakan teknik Synthetic Minority Oversampling Technique (SMOTE) pada tahap pelatihan model. Dua algoritma klasifikasi, yaitu Naive Bayes Classifier dan Support Vector Machine (SVM), dibandingkan performanya melalui proses hyperparameter tuning menggunakan GridSearchCV, dengan metrik F1-Macro. Hasil penelitian menunjukkan bahwa model Naive Bayes Classifier memberikan performa terbaik dengan akurasi sebesar 87.55%, F1-Score sebesar 85.78%, dan F1-Score sebesar 56.70%. Meskipun demikian, model tetap menunjukkan keterbatasan signifikan dalam mengenali kelas sentimen netral (F1-Score 0.21) yang dianalisis lebih lanjut disebabkan oleh proporsi kelas netral yang sangat kecil (2.3% dari total data) serta karakteristik ulasan netral yang cenderung berupa permohonan atau saran kepada pengembang, bukan pernyataan yang eksplisit. Penelitian ini berkontribusi pada pengembangan metodologi analisis sentimen yang lebih representatif melalui kombinasi pelabelan berbasis transformer, penanganan ketidakseimbangan data dan komparasi algoritma klasifikasi.

Kata Kunci: Analisis Sentimen, Naive Bayes Classifier, Support Vector Machine, IndoBERT, Mobile Legends

Abstract– This study was conducted to evaluate user sentiment trends toward the Mobile Legends: Bang Bang game app through reviews published on the Google Play Store platform. This research applied IndoBERT-based automatic labeling to generate sentiment labels that better represent the textual meaning of the reviews. The class distribution imbalance resulting from the labeling process was addressed using the Synthetic Minority Oversampling Technique (SMOTE) during the model training phase. The performance of two classification algorithms—the Naive Bayes Classifier and the Support Vector Machine (SVM)—was compared through hyperparameter tuning using GridSearchCV, with the F1-Macro metric. The results show that the Naive Bayes Classifier model delivers the best performance with an accuracy of 87.55%, an F1-Score of 85.78%, and an F1-Score of 56.70%. However, the model still exhibits significant limitations in recognizing the neutral sentiment class (F1-Score 0.21), which was further analyzed and found to be caused by the very small proportion of the neutral class (2.3% of the total data) as well as the characteristic of neutral reviews, which tend to be requests or suggestions to developers rather than explicit statements. This study contributes to the development of a more representative sentiment analysis methodology through a combination of transformer-based labeling, data imbalance handling, and classification algorithm comparison.

Keywords: Sentiment Analysis, Naive Bayes Classifier, Support Vector Machine, IndoBERT, Mobile Legends

1. PENDAHULUAN

Industri *mobile game* berbasis Multiplayer Online Battle Arena (MOBA) di Indonesia tumbuh pesat, didorong penetrasi *smartphone* dan jaringan internet yang kian luas. Mobile Legends: Bang Bang menjadi salah satu yang paling dominan, menarik jutaan pengguna aktif harian lewat gameplay kompetitif, ekosistem *esports*, dan pembaruan fitur berkala. Ulasan yang ditinggalkan pengguna di Google Play Store menyimpan opini dan kritik bernilai tinggi terkait stabilitas server, keadilan *matchmaking*, kualitas grafis, hingga pengalaman bermain secara keseluruhan. Namun volume ulasan yang terus bertambah membuat analisis manual tidak lagi efisien, terlebih bahasanya didominasi istilah tidak baku, singkatan, *code-mixing*, dan slang komunitas *gamer* seperti *lag*, *buff*, *nerf*, *afk*, atau *noob*.

Analisis sentimen, sebagai bagian dari *text mining*, menjadi pendekatan komputasional yang efektif untuk mendeteksi dan mengelompokkan opini pengguna menjadi sentimen positif, negatif, atau netral, sekaligus mengungkap pola tersembunyi dari tumpukan data ulasan [5]. Salah satu algoritma klasifikasi yang paling sering diandalkan adalah *Naive Bayes Classifier* (NBC), bekerja berdasarkan *Teorema Bayes* dengan asumsi independensi antar fitur. Algoritma ini dikenal karena kompleksitas komputasi yang rendah dan konsisten memberikan akurasi baik pada domain ulasan aplikasi digital [6].

Algoritma lain yang juga populer adalah *Support Vector Machine* (SVM), unggul dalam menemukan *hyperplane* pemisah optimal pada data berdimensi tinggi seperti representasi TF-IDF [4]. Performa kedua algoritma sangat bergantung pada konfigurasi *hyperparameter*, sehingga *tuning* melalui *GridSearchCV* menjadi langkah penting untuk memperoleh model optimal sekaligus memungkinkan perbandingan yang objektif [3].

Efektivitas *Naive Bayes* pada ranah opini pengguna telah dibuktikan di berbagai domain, misalnya pada aplikasi perbankan digital MyBCA dengan akurasi memuaskan [7], maupun pada ulasan e-commerce yang dikombinasikan TF-IDF untuk meningkatkan kualitas ekstraksi fitur [8]. Kualitas data turut menentukan hasil klasifikasi, sehingga *preprocessing* berupa *cleansing*, *case folding*, *stopword removal*, dan *stemming* menjadi tahap wajib untuk membuang derau sebelum pelatihan model [9].

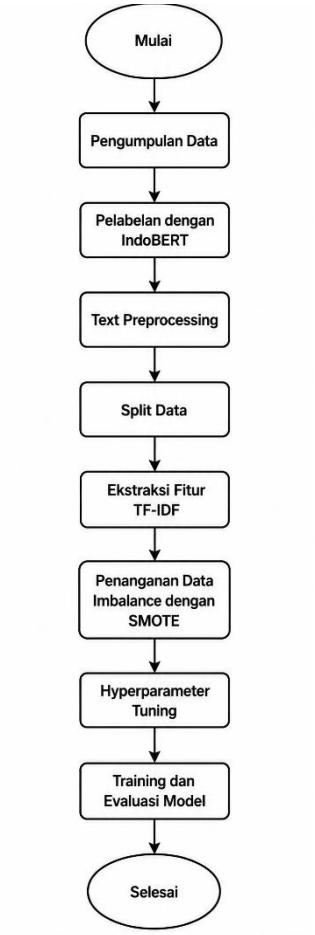
Management of Information System Journal
Vol 4, No 3, Juli 2026, Hal. 151 - 159
ISSN: 2964-9455 (media online)
DOI: 10.47065/mis.v2i3.2783
https://journal.fkpt.org/index.php/MIS

Meski demikian, penelitian sentimen masih terpusat pada layanan umum, media sosial, dan *marketplace*, sementara ulasan *mobile game* kompetitif dengan jargon teknis dan bahasa informal yang terus berubah masih belum banyak dieksplorasi. Mayoritas penelitian juga hanya mengandalkan satu algoritma tanpa *hyperparameter tuning* maupun pembanding, serta jarang menangani ketidakseimbangan kelas sentimen yang lazim terjadi karena pengguna cenderung menulis ulasan saat mengalami masalah. Pelabelan pun umumnya masih memakai skor *rating* sebagai proksi, padahal *rating* tidak selalu selaras dengan muatan sentimen tekstual ulasan. Penelitian ini mengisi celah tersebut melalui pelabelan otomatis berbasis IndoBERT, penanganan ketidakseimbangan data dengan SMOTE, serta komparasi dan *tuning* dua algoritma klasifikasi pada domain ulasan *mobile game*.

2. METODOLOGI PENELITIAN

2.1 Diagram Alur

Tahapan penelitian yang dilakukan seperti yang digambarkan pada , dimulai dari proses pengumpulan data, pelabelan sentimen otomatis, *text preprocessing*, ekstraksi fitur, penanganan ketidakseimbangan data, hingga proses pelatihan dan evaluasi model.



Gambar 1. Diagram Alur

2.2 Pengumpulan Dataset

Objek pada studi yang dilaksanakan adalah ulasan pengguna terhadap aplikasi *Mobile Legends*: *Bang Bang* pada *platform Google Play Store*. Proses pengambilan data dilakukan melalui metode *scraping* didukung oleh pustaka *crawler* otomatis dengan kriteria pengambilan yang dibatasi pada ulasan berbahasa Indonesia. Total awal data yang berhasil ditarik adalah berjumlah 10.000 ulasan. Setelah melalui proses tahap pembersihan dan eliminasi ulasan kosong (*null values*), diperoleh sebanyak 9.997 data ulasan yang lolos tahap validasi siap dimanfaatkan sebagai sumber data penelitian.

2.3 Pelabelan dengan IndoBERT

Penelitian ini menerapkan pelabelan otomatis menggunakan model IndoBERT yang telah di-fine-tune untuk tugas klasifikasi sentimen Bahasa Indonesia, yaitu mdhugol/indonesia-bert-sentiment-classification [1]. Pendekatan ini dipilih karena skor rating tidak selalu selaras dengan sentimen yang terkandung dalam teks ulasan. Seorang pengguna dapat memberikan rating rendah namun menuliskan ulasan yang sebagian bersifat netral atau bahkan positif pada aspek tertentu

Management of Information System Journal
Vol 4, No 3, Juli 2026, Hal. 151 - 159
ISSN: 2964-9455 (media online)
DOI: 10.47065/mis.v2i3.2783
https://journal.fkpt.org/index.php/MIS

[1]. Proses pelabelan dilakukan pada teks ulasan asli (belum melalui tahap preprocessing) untuk menjaga konteks kalimat yang utuh, mengingat model berbasis *transformer* seperti IndoBERT memanfaatkan *tokenizer* serta representasi kontekstual yang dirancang untuk teks berbahasa alami. Hasil pelabelan menghasilkan distribusi kelas seperti pada Tabel:

Tabel 1 Distribusi Label Sentimen Hasil IndoBERT

Kelas Sentimen	Jumlah Data	Presentase
Negatif	8.267	82.7%
Positif	1.505	15.1%
Netral	225	2.3%

2.4 Tahap Text Preprocessing

Mengingat ulasan pada *game mobile* memiliki tingkat ambiguitas tinggi serta dipenuhi oleh istilah *slang* komunitas seperti *lag, buff, nerf,* dan *afk,* maka diperlukan tahapan *text preprocessing* yang mendalam. Langkah pembersihan ini krusial karena kualitas data teks yang kotor dapat menurunkan akurasi model klasifikasi secara drastis [18]. Proses penyiapan teks ini terdiri dari lima tahapan utama yang dilakukan secara berurutan. Tahap pertama adalah *cleansing* yang berfungsi menghilangkan komponen teks kurang berkaitan, seperti angka, simbol baca, simbol khusus, emoji, dan tautan *URL.* Tahap kedua adalah *case folding* yang mengganti seluruh karakter huruf kapital mengenai dokumen diubah ke bentuk huruf kecil guna menyeragamkan format teks. Tahap ketiga adalah *normalizing* atau normalisasi untuk mengubah kata kata tidak baku, singkatan informal, dan bahasa gaul komunitas *game* diubah menjadi bentuk kata baku bahasa Indonesia sesuai dengan kamus normalisasi yang telah dibuat. Tahap keempat adalah *stopword removal* yang menghilangkan kata-kata umum yang sering muncul namun tidak memiliki nilai informasi atau kontribusi terhadap polaritas sentimen seperti kata "yang", "dan", "di", maupun "dari". Tahap kelima adalah *stemming* yang mengubah kata kata berimbun menjadi kata dasarnya dengan bantuan algortima *stemmer* khusus bahasa Indonesia untuk menyederhanakan variasi fitur kata. Tahapan *text preprocessing* dilakukan untuk nanti dilakukan transformasi menggunakan TF-IDF.

2.5 Split Data

Pembagian dataset menjadi data latih dan data uji merupakan tahapan penting dalam pengembangan model machine learning untuk menjamin kemampuan generalisasi model terhadap data yang belum pernah dilihat sebelumnya. Penelitian ini menerapkan skema train_test_split dengan rasio 80:20, proporsi yang umum digunakan dalam penelitian klasifikasi teks karena memberikan keseimbangan antara ketersediaan data latih yang memadai dan data uji yang representatif untuk evaluasi [1]. Proses pembagian data dilakukan sebelum tahap ekstraksi fitur, bukan sesudahnya, untuk mencegah *data leakage*, dimana kondisi di mana informasi statistik dari data uji secara tidak sengaja memengaruhi proses pembelajaran model, misalnya melalui kosakata atau bobot *Inverse Document Frequency* (IDF) yang ikut terbentuk dari data uji apabila ekstraksi fitur dilakukan pada keseluruhan dataset sebelum pembagian. Selain itu, parameter *stratify* diterapkan pada proses *splitting* untuk mempertahankan proporsi tiap kelas sentimen secara konsisten pada data latih maupun data uji, mengingat distribusi kelas pada penelitian ini tidak seimbang sejak tahap pelabelan [2].

2.6 Ekstraksi Fitur TF-IDF

Setelah proses data kalimat diubah menjadi rapi serta terstruktur, tahapan berikutnya yaitu melakukan ekstraksi fitur untuk memproses data tekstual tersebut dikonversi menjadi data numerik berupa vector yang mampu di proses oleh algoritma *machine learning.* Penelitian tersebut menerapkan teknik pembobotan *term Frequency Inverse Document Frequency* (TF-IDF). Metode TF-IDF terbukti optimal saat membedakan istilah yang memiliki informasi signifikan dengan kata yang memiliki frekuensi kemunculan tinggi pada seluruh dokumen [20]. Nilai bobot dari sebuah kata atau term (t) pada suatu dokumen (d) dapat dihitung menggunakan persamaan (1).

$$w_{d,t} = tf_{d,t} \times \log\left(\frac{N}{df_t}\right) \quad (1)$$

Berdasarkan Persamaan (1) tersebut, $w_{d,t}$ menunjukkan bobot akhir dari kata, sedangkan $tf_{d,t}$ merepresentasikan intensitas kemunculan kata t di bagian dokumen d. Komponen N menyatakan akumulasi jumlah keseluruhan dokumen yang ada di dalam Kumpulan data teks, dan df_t menunjukkan banyak nya dokumen yang berisi kata t tersebut. Melalui proses pembobotan ini, fitur kata kunci yang representatif akan mendapatkan nilai bobot yang lebih tinggi. Hasil akhir dari ekstraksi fitur ini dibatasi hingga menghasilkan 5.000 fitur term utama yang paling relevan dalam mencapai keseluruhan.

2.7 Penanganan Data Imbalance dengan SMOTE

Distribusi label hasil pelabelan IndoBERT menunjukkan ketidakseimbangan kelas yang signifikan,dengan kelas negatif mendominasi (82,7%) sementara kelas netral hanya berjumlah 2,3% dari keseluruhan data. Ketidakseimbangan ini berpotensi menyebabkan model bias dalam memprediksi kelas mayoritas dan mengabaikan kelas minoritas. Untuk mengatasi permasalahan tersebut, penelitian ini menerapkan *Synthetic Minority Oversampling Technique* (SMOTE) [2], yaitu teknik oversampling yang menghasilkan data sintetis pada kelas minoritas melalui interpolasi antar sampel yang berdekatan dalam ruang fitur. SMOTE diterapkan secara eksklusif pada data latih (train set) di dalam skema *cross-validation*, sehingga data uji (test set) tetap mempertahankan distribusi aslinya guna menjamin validitas evaluasi model.

Management of Information System Journal
Vol 4, No 3, Juli 2026, Hal. 151 - 159
ISSN: 2964-9455 (media online)
DOI: 10.47065/mis.v2i3.2783
https://journal.fkpt.org/index.php/MIS

Sebelum penerapan SMOTE, distribusi data latih tercatat sebanyak 6.613 data negatif, 180 data netral, dan 1.204 data positif. Strategi *resampling* ditetapkan secara proporsional, yaitu menaikkan jumlah data kelas netral menjadi 1.500 namun tidak menyamakan seluruh kelas pada rasio 1:1, guna menghindari *oversampling* ekstrem yang berisiko menghasilkan data sintetis dengan kualitas rendah akibat basis data asli yang terlalu sedikit.

2.8 Hyperparameter Tuning

Dua algoritma klasifikasi dibandingkan pada penelitian ini, yaitu *Naive Bayes Classifier* dan *Support Vector Machine* (SVM) kernel linear. Kedua algoritma dituning menggunakan *GridSearchCV* dengan skema *5-fold cross-validation* dan metrik F1-macro sebagai acuan *scoring*, dipilih karena memberi bobot setara pada setiap kelas sentimen sehingga lebih representatif pada kondisi data tidak seimbang [21]. Proses tuning dilakukan pada data latih pasca-SMOTE, dengan pencarian parameter alpha (*Naive Bayes*) pada rentang {0,1; 0,5; 1,0; 2,0} dan parameter C (SVM) pada rentang {0,1; 0,5; 1,0; 2,0; 5,0}. Hasil tuning diperoleh parameter optimal alpha=0,1 untuk Naive Bayes (CV F1-macro 57,21%) dan C=0,5 untuk SVM (CV F1-macro 59,37%).

2.9 Training dan Evaluasi Model

Parameter optimal hasil tuning selanjutnya digunakan untuk melatih ulang kedua model pada keseluruhan data latih pasca-SMOTE. Kedua model yang telah dilatih kemudian diuji pada 2.000 data uji yang tetap mempertahankan distribusi kelas asli, guna menjamin validitas hasil evaluasi. Perbandingan performa kedua model dilakukan menggunakan pendekatan *confusion matrix* sebagai instrumen evaluasi standar dalam klasifikasi [21], yang menghasilkan empat metrik utama yaitu akurasi (persentase prediksi benar dari seluruh data uji), presisi (ketepatan prediksi positif suatu kelas terhadap keseluruhan prediksi kelas tersebut), recall (kemampuan model mengenali seluruh data aktual suatu kelas), dan F1-score (rata-rata harmonis presisi dan recall). Selain F1-score per kelas, F1-macro turut dilaporkan sebagai metrik pembanding utama antara kedua model, karena tidak bias terhadap kelas mayoritas. Struktur matriks evaluasi yang digunakan ditampilkan pada Tabel 2.

Tabel 2 Struktur Matriks Evaluasi

Prediksi	Prediksi Positif	Prediksi Negatif
Positif	Positif Benar	Negatif Salah
Negatif	Positif Salah	Negatif Benar

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

Dataset diperoleh melalui scraping ulasan Mobile Legends: Bang Bang pada Google Play Store, terdiri dari 10.000 data dengan 11 atribut (reviewId, userName, content, score, dan lainnya). Penelitian ini hanya memanfaatkan kolom content sebagai teks ulasan. Ulasan didominasi keluhan seputar lag, jaringan, matchmaking, dan bug, meski sebagian mengapresiasi gameplay dan pembaruan fitur. Rincian dataset tersaji pada Tabel 3.1.

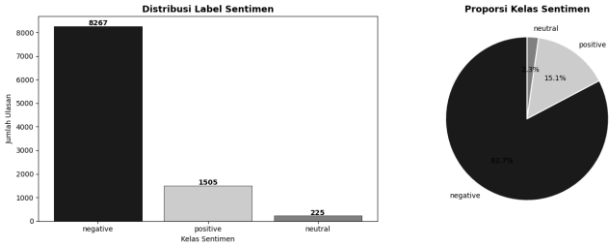
Tabel 3. Informasi Dataset

Keterangan	Hasil
Jumlah Data	10.000
Jumlah Kolom	11
Sumber Data	Google Play Store
Objek Penelitian	Ulasan Mobile Legends
Bahasa	Bahasa Indonesia

3.2 Pelabelan Sentimen dengan IndoBERT

Pelabelan dilakukan menggunakan IndoBERT pada teks ulasan. Hasil pelabelan menghasilkan 8.267 data negatif (82,7%), 1.505 data positif (15,1%), dan 225 data netral (2,3%), sebagaimana Gambar 3.2. Proporsi kelas netral yang kecil ini mencerminkan karakteristik ulasan aplikasi permainan, di mana pengguna cenderung menuliskan keluhan atau pujian secara eksplisit ketimbang pernyataan bernada netral, sehingga sebagian besar ulasan condong terklasifikasi ke salah satu kutub sentimen.

Management of Information System Journal
Vol 4, No 3, Juli 2026, Hal. 151 - 159
ISSN: 2964-9455 (media online)
DOI: 10.47065/mis.v2i3.2783
https://journal.fkpt.org/index.php/MIS



Gambar 2 Distribusi Sentimen IndoBERT

Setelah pelabelan dilakukan, maka perlu untuk membersihkan noise yang ada pada teks ulasan yang nantinya akan dilakukan pelatihan model, sehingga selanjutnya adalah melakukan *text preprocessing*.

3.3 Text Preprocessing

Implementasi *preprocessing* menggunakan kamus normalisasi sebanyak 93 kosa kata, 828 *stopword* Bahasa Indonesia ditambah *custom stopword*, serta *library* Sastrawi untuk *stemming*. Kata negasi ("tidak", "bukan", "jangan", "kurang", "tanpa", "belum") sengaja dipertahankan pada tahap *stopword removal* karena membalikkan polaritas makna kalimat, didukung fitur *bigram* pada tahap TF-IDF. Proses ini menghasilkan 9.997 data valid siap diekstraksi fiturnya. Contoh transformasi teks ditampilkan pada Tabel 4 dan 5

Tabel 4 Hasil Preprocessing

Tahapan	Contoh Hasil
Original	Aku turunkan rating bintang nya
Cleansing	Aku turunkan rating bintang nya
Case Folding	aku turunkan rating bintang nya
Stopword Removal	Turunkan rating Bintang nya
Stemming	Turun rating bintang

Tabel 5 Perbandingan Sebelum dan Sesudah Preprocessing

Teks Asli	Hasil Preprocessing	Label
Game gak guna	Game guna	Negatif
Ayolah ini kadang game nya ngelag	Ayo game kadang ngelag	Negatif
Ini game seru banged	Game seru banged	Positif

Tahapan *preprocessing* divisualisasikan secara rinci melalui hasil output program dan digunakan guna meningkatkan mutu data sebelum proses ekstraksi fitur dilakukan.

3.4 Split Data

Dataset dibagi kedalam data latih dan data uji menggunakan metode train_test_split dengan rasio 80:20. Sebanyak 7.997 data dimanfaatkan sebagai data pelatihan dan 2.000 data digunakan sebagai data uji. Penguraian data dilakukan dengan bantuan parameter strafity agar distribusi kelas tetap seimbang sesuai dataset asli sebagaimana ditampilkan pada Tabel 6

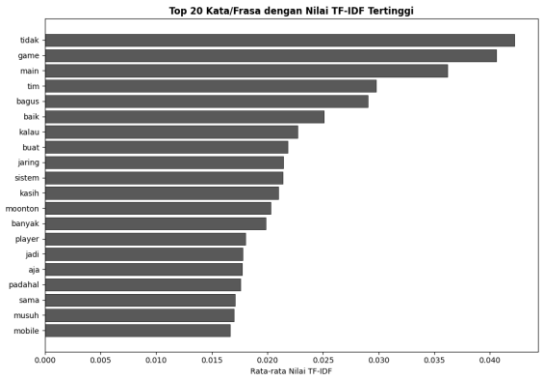
Tabel 6. Distribusi Kelas Data Latih dan Uji

Sentimen	Data Latih	Data Uji
Negatif	6.613	1.654
Positif	1.204	301
Netral	180	45

Management of Information System Journal
Vol 4, No 3, Juli 2026, Hal. 151 - 159
ISSN: 2964-9455 (media online)
DOI: 10.47065/mis.v2i3.2783
https://journal.fkpt.org/index.php/MIS

3.5 Ekstraksi Fitur TF-IDF

Langkah pengambilan fitur diterapkan menggunakan metode TF-IDF (Term Frequency Inverse Document Frequency). Pada penelitian ini memanfaatkan parameter *max_features* sebanyak 5.000 fitur dengan kombinasi *unigram* dan *bigram*. Hasil ekstraksi menunjukkan bahwa kata "tidak", "game", dan "main" merupakan kata yang paling dominan muncul pada dataset ulasan pengguna. Kata kata tersebut memiliki bobot TF-IDF yang tinggi karena sering muncul pada berbagai ulasan pengguna Mobile Legends. Hasil ekstraksi fitur divisualisasikan pada Gambar 3



Gambar 3. Top 20 TF-IDF

3.6 Penanganan Data Imbalance dengan SMOTE

Strategi oversampling diterapkan secara proporsional dengan menaikkan jumlah data netral menjadi 1.500, tanpa menyamakan seluruh kelas pada rasio 1:1, guna menghindari data sintetis berkualitas rendah akibat basis data asli yang terlalu sedikit. Tabel 7. menyajikan perbandingan distribusi kelas sebelum dan sesudah SMOTE.

Tabel 7. Distribusi Kelas Sebelum dan Sesudah SMOTE

Kelas	Sebelum SMOTE	Sesudah SMOTE
Negatif	6.613	6.613
Positif	1.204	1.204
Netral	180	1.500

3.7 Hyperparameter Tuning

Setelah data latih diseimbangkan melalui SMOTE, tahap selanjutnya adalah menentukan konfigurasi parameter terbaik bagi kedua algoritma. Proses pencarian dilakukan menggunakan GridSearchCV dengan skema 5-fold cross-validation dan F1-macro sebagai metrik acuan, pada rentang alpha {0,1; 0,5; 1,0; 2,0} untuk Naive Bayes dan C {0,1; 0,5; 1,0; 2,0; 5,0} untuk SVM. Konfigurasi dengan skor validasi tertinggi untuk masing-masing algoritma ditampilkan pada Tabel 8.

Tabel 8. Hasil Hyperparameter Tuning

Model	Parameter Terbaik	CV F1-Macro
Naïve Bayes	Alpha = 0.1	57.21%
SVM (Linear)	C = 0.5	59.37%

Berdasarkan Tabel 8, pada tahap validasi ini SVM memperoleh skor F1-macro yang sedikit lebih tinggi dibandingkan Naive Bayes. Perlu dicatat bahwa skor tersebut merupakan rata-rata performa lintas *fold* pada data latih, sehingga masih bersifat sementara dan belum tentu mencerminkan performa kedua model ketika dihadapkan pada data uji yang sepenuhnya terpisah dari proses pelatihan. Pengujian pada data uji dilakukan pada tahap berikutnya untuk memperoleh gambaran performa yang sesungguhnya.

Management of Information System Journal
Vol 4, No 3, Juli 2026, Hal. 151 - 159
ISSN: 2964-9455 (media online)
DOI: 10.47065/mis.v2i3.2783
https://journal.fkpt.org/index.php/MIS

3.8 Training dan Evaluasi Model

Parameter optimal hasil tuning selanjutnya digunakan untuk melatih ulang kedua model pada keseluruhan data latih pasca-SMOTE. Model yang telah dilatih kemudian diuji pada 2.000 data uji yang tetap mempertahankan distribusi kelas asli, guna menjamin validitas hasil evaluasi. Ringkasan performa akhir kedua model disajikan pada Tabel 9

Tabel 9 Hasil Evaluasi *Naive Bayes* dan SVM

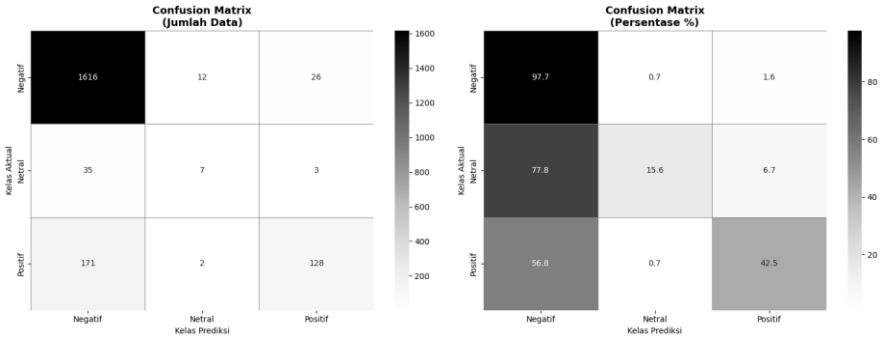
Model	Accuracy	Precision	Recall	F1-Score	F1-Macro
Naïve Bayes	87.55%	86.37%	87.55%	85.78%	56.70%
SVM (Linear)	87.15%	85.98%	87.15%	86.06%	56.53%

Berbeda dengan hasil validasi pada tahap *tuning*, performa pada data uji menunjukkan *Naive Bayes* justru unggul tipis dibandingkan SVM, baik pada akurasi (87,55% berbanding 87,15%) maupun F1-macro (56,70% berbanding 56,53%). SVM hanya sedikit lebih tinggi pada F1-score berbasis *weighted average* (86,06% berbanding 85,78%). Perbedaan arah keunggulan ini terjadi karena *F1-weighted* memberi bobot lebih besar pada kelas mayoritas, sedangkan *F1-macro* memperlakukan seluruh kelas secara setara tanpa memandang jumlah datanya. Karena kelas netral menjadi perhatian utama penelitian ini, F1-macro dipilih sebagai kriteria penentu, sehingga Naive Bayes ditetapkan sebagai model terbaik. Untuk mengevaluasi performa model terbaik (*Naive Bayes*) secara lebih rinci, Tabel 10 menyajikan *classification report* per kelas sentimen dari model *Naive Bayes*.

Tabel 10 *Classification Report* Model *Naive Bayes*

Kelas	Precision	Recall	F1-Score	Support
Negatif	0.89	0.98	0.93	1.654
Netral	0.33	0.16	0.21	45
Positif	0.82	0.43	0.56	301

Tabel 10 menunjukkan model bekerja sangat baik pada kelas negatif (F1-score 0,93), sejalan dengan dominasinya sebagai kelas mayoritas. Kelas positif berada pada posisi menengah dengan presisi cukup tinggi (0,82) namun recall rendah (0,43), sedangkan kelas netral memperoleh performa terendah (F1-score 0,21). Pola kesalahan klasifikasi tersebut divisualisasikan melalui confusion matrix pada Gambar 4.



Gambar 4 *Heatmap Confusion Matrix Naive Bayes*

Dengan detail hasil confusion matrix seperti pada Tabel 11.

Tabel 11 *Confusion Matrix* per Kelas

Aktual \ Prediksi	Negatif	Netral	Positif
Negatif	1616	12	26
Netral	35	7	3
Positif	171	2	128

Rendahnya performa pada kelas netral tidak sepenuhnya disebabkan keterbatasan algoritma, melainkan turut dipengaruhi karakteristik data. Pertama, dari sisi kuantitas, kelas netral hanya bertumpu pada 225 data (2,3%), jauh lebih sedikit dibanding dua kelas lain sehingga basis pembelajaran model sangat terbatas bahkan setelah diperbesar melalui SMOTE. Kedua, dari sisi kualitas label, distribusi *confidence score* IndoBERT terhadap prediksi netral menunjukkan median 0,647

Management of Information System Journal
Vol 4, No 3, Juli 2026, Hal. 151 - 159
ISSN: 2964-9455 (media online)
DOI: 10.47065/mis.v2i3.2783
https://journal.fkpt.org/index.php/MIS

dengan 41,8% di antaranya di bawah 0,60, mengindikasikan hampir separuh label netral dihasilkan dari prediksi yang ragu-ragu. Ketiga, dan paling substansial, pemeriksaan terhadap sampel data netral menunjukkan 77,8% di antaranya bukan pernyataan sentimen eksplisit, melainkan permohonan, saran, atau keluhan teknis kepada pengembang (seperti permintaan perbaikan fitur atau pelaporan bug) yang secara linguistik tidak memiliki penanda polaritas jelas sehingga sulit dibedakan dari kelas negatif. Temuan ini menunjukkan bahwa tantangan pada kelas netral di domain ulasan aplikasi permainan bersifat inheren pada karakteristik data, bukan semata kelemahan model.

Tahap terakhir adalah pengujian model menggunakan beberapa contoh teks baru. Pengujian dilakukan guna menentukan kemampuan model dalam memprediksi penilaian terhadap data terpisah dari data pelatihan. Hasil evaluasi membuktikan kemampuan model dalam mengenali persepsi positif dan negatif dengan cukup relevan. Namun, model masih mengalami kesalahan prediksi pada kalimat ambigu dan sentimen netral. Kondisi ini menegaskan bahwa sistem *Naive Bayes* belum terlepas dari kendala saat memahami hubungan kalimat tertentu. Sebagai sampel hasil uji model manual terdapat pada Tabel 12

Tabel 12. Hasil Uji Coba Manual

Kalimat	Prediksi
Game bagus banged	Positif
Lag mulu gak bisa main	Negatif
Biasa aja si game nya	Negatif
Update bikin panas hp	Negatif
Hero baru ini strong banged	Negatif

4. KESIMPULAN

Berdasarkan seluruh rangkaian penelitian analisis sentimen pada ulasan pengguna aplikasi Mobile Legends di Google Play Store, dapat disimpulkan bahwa kombinasi pelabelan otomatis berbasis IndoBERT, penanganan ketidakseimbangan data menggunakan SMOTE, serta komparasi antara algoritma Naive Bayes dan Support Vector Machine mampu menghasilkan model klasifikasi sentimen dengan performa yang cukup baik, di mana tahap text preprocessing yang meliputi cleansing, case folding, normalizing, stopword removal, dan stemming terbukti krusial dalam menstransformasikan ulasan yang bersifat informal menjadi bentuk terstruktur sebelum diekstraksi menggunakan TF-IDF; dari kedua algoritma yang dibandingkan, Naive Bayes (alpha=0,1) terpilih sebagai model terbaik dengan akurasi 87,55% dan F1-macro 56,70%, meskipun model masih menunjukkan keterbatasan dalam mengenali kelas sentimen netral (F1-score 0,21) yang disebabkan oleh proporsi datanya yang sangat kecil serta karakteristik kalimatnya yang cenderung berupa permohonan teknis kepada pengembang tanpa penanda polaritas yang jelas, sementara dominasi sentimen negatif sebesar 82,7% mengindikasikan bahwa pengalaman pengguna masih diwarnai berbagai kendala teknis seperti ketidakstabilan koneksi jaringan, sistem matchmaking yang kurang seimbang, dan isu performa perangkat, sehingga temuan ini memberikan wawasan strategis bagi pengembang untuk memprioritaskan perbaikan pada aspek-aspek tersebut, dan untuk penelitian selanjutnya disarankan mengeksplorasi model klasifikasi berbasis deep learning atau transformer yang mampu menangkap konteks kalimat secara lebih mendalam, serta strategi penanganan kelas minoritas yang lebih lanjut seperti anotasi manual terbatas guna meningkatkan performa pada kelas netral.

UCAPAN TERIMAKASIH

Terima kasih disampaikan kepada pihak-pihak yang telah mendukung terlaksananya penelitian ini.

REFERENCES

D. I. Putri, A. N. Alfian, M. Y. Putra, and P. D. Mulyo, "IndoBERT Model Analysis: Twitter Sentiments on Indonesia's 2024 Presidential Election," J. Appl. Informatics Comput. (JAIC), vol. 8, no. 1, pp. 45-52, Jul. 2024, doi: 10.30871/jaic.v8i1.7440.

F. Y. A'la, "Optimasi Klasifikasi Sentimen Ulasan Game Berbahasa Indonesia: IndoBERT dan SMOTE untuk Menangani Ketidakseimbangan Kelas," J. Pendidikan Inform., vol. 9, no. 1, pp. 256-265, Jan. 2025.

H. Ahmadian et al., "Pretrained Models Against Traditional Machine Learning for Detecting Fake Hadith," Electronics, vol. 14, no. 17, p. 3484, Sep. 2025, doi: 10.3390/electronics14173484.

A. S. Safitri, I. Wijayanto, and S. Hadiyoso, "Improving Classification Accuracy with Preprocessing Techniques for Sentiment Analysis," in Proc. Int. Conf. Data Sci. Its Appl. (ICoDSA), Kuta, Indonesia, 2024, pp. 487-490, doi: 10.1109/ICoDSA61314.2024.10625412.

Management of Information System Journal
Vol 4, No 3, Juli 2026, Hal. 151 - 159
ISSN: 2964-9455 (media online)
DOI: 10.47065/mis.v2i3.2783
https://journal.fkpt.org/index.php/MIS

Purnajaya, A. R., & Pernando, Y. (2023). Analisa Sentimen Informasi Hoaks Pasca Pandemi Covid-19 dengan Text Mining. *Journal of Computer System and Informatics (JoSYC)*, *4*(3), 460–469. https://doi.org/10.47065/josyc.v4i3.3358

Daniati, E., & Utama, H. (2023). ANALISIS SENTIMEN DENGAN PENDEKATAN ENSEMBLE LEARNING DAN WORD EMBEDDING PADA TWITTER. *Journal of Information System Management (JOISM)*, *4*(2), 125–131. https://doi.org/10.24076/joism.2023v4i2.973

Dhamara, G. Z., Alamsyah, D. N., Saputro, P. W., Daniati, E., & Ristyawan, A. (2024). Analisis Sentimen Aplikasi Mybca Melalui Review Pengguna Di Google Play Store Menggunakan Algoritma Naive Bayes. *Inotek*, *8*.

Ferdiansyah, R., Abadi, K. R., & Daniati, E. (2025). Analisis Sentimen Dalam Tokopedia Terhadap Ulasan Pelanggan Menggunakan Metode Naïve Bayes. *Prosiding SEMNAS INOTEK (Seminar Nasional Inovasi Teknologi)*, *9*(1), 024–029. https://proceeding.unpkediri.ac.id/index.php/inotek/article/view/7201

Hakim, B. (2021). Analisa Sentimen Data Text Preprocessing Pada Data Mining Dengan Menggunakan Machine Learning. *JBASE - Journal of Business and Audit Information Systems*, *4*(2), 16–22. https://doi.org/10.30813/jbase.v4i2.3000

S. R. Sahoo and B. B. M. Kumar, "Customer Sentiment Analysis on E-Commerce Platforms Using Naive Bayes and Support Vector Machine Architectures," Springer: Journal of Consumer Behaviour and Analytics, vol. 12, no. 2, pp. 145-159, Jun. 2024, doi: 10.1007/s40537-024-00891-2.

K. H. Ahmed and T. O. Shareef, "A Hybrid Approach for Sentiment Analysis Using VADER Lexicon and Naive Bayes Classifier on App Store Reviews," Elsevier: Computer Speech & Language, vol. 89, p. 101712, Jan. 2025, doi: 10.1016/j.csl.2024.101712.

Hajar, S., Purwandira, A., Febiyanti, I., Daniati, E., & Ristyawan, A. (2024). Klasifikasi Sentimen Pengguna Aplikasi Livin By Mandiri Pada Playstore Menggunakan Algoritma Naive Bayes. *Agustus*, *8*, 2549–7952.

Hanum, A. R., Zetha, I. A., Putri, S. C., Wulandari, R. A., Andina, S. P., Fajrina, J. N., & Yudistira, N. (2024). Analisis Kinerja Algoritma Klasifikasi Teks Bert dalam Mendeteksi Berita Hoaks. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *11*(3), 537–546. https://doi.org/10.25126/jtiik.938093

T. Dinh, "Mining Insights from Esports Game Reviews with an Aspect-Based Sentiment Analysis Framework," IEEE Access, vol. 11, pp. 112045-112058, 2023, doi: 10.1109/ACCESS.2023.3323041.

Khoirunnisaa, N., Nabila Nastiti Kesuma, K., Setiawan, S., & Yunizar Pratama Yusuf, A. (2024). KLASIFIKASI TEKS ULASAN APLIKASI NETFLIX PADA GOOGLE PLAY STORE MENGGUNAKAN ALGORITMA NAÏVE BAYES DAN SVM. *SKANIKA: Sistem Komputer Dan Teknik Informatika*, *7*(1), 64–73. https://doi.org/10.36080/skanika.v7i1.3138

Latifah, U., Zuhriya, T. K., & Daniati, E. (2025). Analisis Sentimen Komentar Twitter (X) tentang Kesehatan Mental menggunakan Naïve Bayes untuk Mengukur Opini Publik. *Prosiding SEMNAS INOTEK (Seminar Nasional Inovasi Teknologi)*, *9*(X), 2549–7952. https://doi.org/https://doi.org/10.29407/pcj3sb07

H. Ahmadian et al., "Pretrained Models Against Traditional Machine Learning for Detecting Fake Hadith," Electronics, vol. 14, no. 17, p. 3484, 2025, doi: 10.3390/electronics14173484.

Pane, S. F., & Ramdan, J. (2022). Pemodelan Machine Learning : Analisis Sentimen Masyarakat Terhadap Kebijakan PPKM Menggunakan Data Twitter. *Jurnal Sistem Cerdas*, *5*(1), 12–20. https://doi.org/10.37396/jsc.v5i1.191

G. Kaur and V. Sharma, "A deep learning-based model using hybrid feature extraction approach for consumer sentiment analysis," Springer: Journal of Big Data, vol. 10, no. 1, p. 5, 2023, doi: 10.1186/s40537-022-00680-x.

Dianda Rifaldi, Tri Stiyo Famuji, Bella Okta Sari Miranda, Fauzan Purma Ramadhan, Iriene Putri Mulyadi, Vanji Saputra6, & Fanani, G. P. I. (2025). Evaluasi Sentimen Pengguna ChatGPT Menggunakan Naive Bayes: Tinjauan dari Confusion Matrix dan Classification Report. *Jurnal Riset Sistem Dan Teknologi Informasi*, *3*(2), 81–89. https://doi.org/10.30787/restia.v3i2.1990